1 Mechanisms for Network Structure

Analyzing the structure of a network, e.g., using centrality measures, degree distributions, relationships between network measures, comparison versus the configuration model, community structure, etc., allows us to identify a set of empirical patterns that characterize one or several networks. These patterns provide general insight into the network's large-scale organizational patterns, for instance, along what structural dimensions the network exhibits assortative or disassortative patterns, whether connectivity is concentrated among a minority of vertices, and whether those vertices are in the core or periphery of the network. Similarly, comparing the observed patterns against a good null model, like the configuration model, allows us to tell whether the pattern is surprising, given certain assumptions like fixing the degree distribution.

But, these techniques do not necessarily allow us to explain *why* we see these patterns and not others. For instance, why do social networks exhibit high clustering coefficients? Why do biological networks exhibit degree disassortativity? Why do citation networks exhibit power-law degree distributions? Why do online social networks also exhibit heavy-tailed degree distributions, while friendship networks exhibit much lighter tails?

Explaining the origin of a pattern requires identifying the underlying mechanism that generates it, i.e., the cause or causes that produce it as an effect.[1] Establishing causality for network patterns is a difficult task because typically we only have access to *observational data*, i.e., data that we observe passively rather than data we generate through a controlled experiment.[2] The central difficulty is that there are often multiple plausible mechanisms that can produce any particular empirical pattern, and the observational data alone may not provide the means to distinguish between them. That is, the data we want is rarely the data we can get. For this reason, caution should be employed in drawing any conclusions about causality.

Two things can make the mechanism inference task somewhat easier. First, temporal data, i.e., data over multiple points in time, allows us to eliminate mechanisms that do not match both the static and evolving empirical pattern. Second, requiring that a mechanism match multiple empirical patterns allows us to eliminate mechanisms that only match a subset of these patterns.

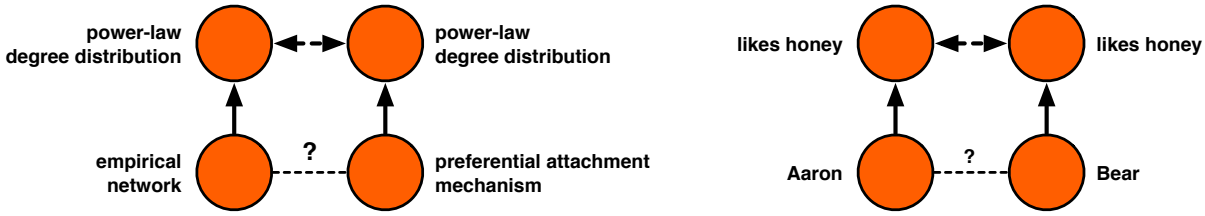
In this lecture, we will investigate these ideas in the context of the popular preferential attachment

[1] It is worth noting that there is a significant bias toward single-cause explanations in the natural sciences, and toward multiple-cause explanations in the social sciences. The biological sciences are more schizophrenic, with some fields taking after the natural sciences, and others after the social sciences.

[2] Social scientists have recently begun examining specific questions about networks and causality in controlled experiments. These efforts are generally exciting, although sometimes it can be unclear whether the results extend outside the laboratory setting. For good examples, see Salganik, Dodds, and Watts, "Experimental study of inequality and unpredictability in an artificial cultural market." *Science* **311**, 854–856 (2006), and Kearns, Suri, and Montfort, "Experimental Study of the Coloring Problem on Human Subject Networks." *Science* **313**, 824–827 (2006).

mechanism for network growth, whose signature output is a network with a power-law degree distribution. But, a power-law degree distribution can be produced by many network mechanisms. This implies that observing a power-law degree distribution in some network is a necessary but not a sufficient condition to conclude that the underlying network dynamics are governed by preferential attachment.

To make this point clear, consider the following pair of logical diagrams. The left shows the usual situation with trying to determine whether some empirical network is governed by the preferential attachment mechanism.[3] The right shows an exactly analogous, albeit silly situation. The question



is whether we can conclude that some network is governed by the preferential attachment mechanism (or is "scale-free" in the manner implied by that mechanism) on the basis of it having a power-law degree distribution. Concluding in favor is a logical fallacy of the following sort: Aaron likes honey; Bears like honey; therefore, Aaron is a Bear. Clearly, this is wrong, but its wrongness highlights the utility of examining either (or both) temporal observations and multiple patterns as a way to constrain the space of plausible hypotheses. To determine if I am a Bear, examine whether I share other features in common with bears, e.g., possess claws, a fuzzy tail and fur, have a habit of eating moths for extra protein, amble around mostly on all fours, etc.

2 The preferential attachment mechanism

The preferential attachment mechanism is perhaps the best known mechanism for network growth. At its core, it describes how a new vertex joining a network will distribute any edges it brings with it. It is purely a model of network growth, and says little about how vertices or edges are removed from the network. The mechanism itself goes by many other names and has been reinvented (and renamed) several times over the past 100 years. Here is a brief summary of its storied history.

[3] Actually, there is a missing step on the left, which is determining that the network degree distribution does indeed follow a power law, which requires a statistical test of the kind we saw earlier in the semester.

2.1 A brief history

Mathematically, preferential attachment is equivalent to the *Yule process* for modeling the distribution of the sizes of biological taxa (for instance, how many species are in a genus), first studied by the statistician Udny Yule (1871–1951) in 1925. The Yule process is a kind of variation on the *Polya's urn* model, due to the mathematician George Pólya (1887–1985). The Yule process was named and generalized by the economist Herbert Simon (1916–2001; Turing Award in 1975 and Nobel Prize in Economics in 1978) to study the distribution of wealth. Simon showed mathematically that the mechanism produces power-law distributions. The "rich get richer" mechanism of preferential attachment was also recognized qualitatively by the sociologist Robert Merton (1910–2003), who called it the Matthew effect, after a passage in Biblical Gospel of Matthew. In the 1970s, the physicist Derek de Solla Price (1922–1983; sometimes called the "father" of scientometrics), inspired by Simon's work, adapted the Yule process to the study of the evolution of citation networks and renamed the mechanism *cumulative advantage*.

More recently, the physicists Albert-Laszlo Barabási and Reka Albert reinvented Price's network growth mechanism in a 1999 paper, giving it the name *preferential attachment*. Work did not stop there, of course. The *vertex-copying* models proposed in the past decade for the structure of gene networks, such as those proposed by the physicist Ricard Solé and colleagues (2002) and by the mathematician Alexei Vázquez and colleagues (2003), the fitness-based generalization of preferential attachment, proposed as a model of the WWW by physicists Ginestra Bianconi and Albert-Laszlo Barabási in 2001, the *forest fire* model for densification, proposed by the computer scientist Jure Leskovec and colleagues (2005), and the local-competition mechanism proposed by the physicist Raissa D'Souza and colleagues (2007), can all be framed as variations, explanations or generalizations of Price's model. Given its popularity and its mathematical simplicity, much is known about the behavior of this mechanism. We will cover a few highlights, returning to the question of mechanism inference at the end.

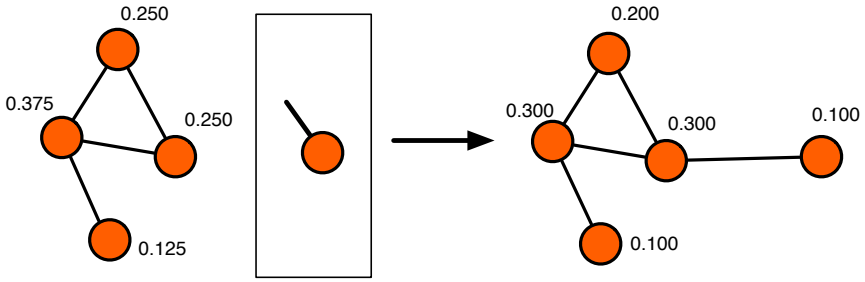
2.2 The basic mechanism

Price's model of a citation network is simple. Assume that papers are published continuously and that new papers only cite papers that have appeared previously. Because new papers are always being published and old papers are never destroyed, the network grows monotonically with time. For simplicity, let each new paper have a bibliography containing an average of c citations. That is, c is the average out-degree of a vertex joining the network.[4]

[4]For the mathematics to work, the distribution of bibliography lengths $\Pr(c)$ simply must have a finite variance. This rules out bibliography lengths distributed according to a power law with $\alpha < 3$. Fortunately, empirical work supports this assumption, although it also shows that bibliography lengths vary by field, and have been getting longer in recent years. Interestingly, the average number of authors on a paper has also been slowly increasing.

The central question for determining the evolution of the network structure is, How does a new paper choose which previous papers to cite? Price's assumption was those papers are chosen at random with probability proportional to the number of citations those previous papers already have.[5] Thus, highly cited papers are likely to gain additional citations and the "rich get richer."

For instance, consider an existing network of four vertices to which we will add a single new vertex. For simplicity, we let this vertex have a single edge to distribute. In the figure, each vertex is annotated with its fraction of the total degree ($2m$). To make the new edge, we flip a coin and connect the new vertex to the lucky existing vertex. We may recalculate the relative share of edge "wealth" held by each vertex. Repeating this process for each new vertex grows the network.



Naturally, this model is highly simplified as it ignores contributions such as the quality or importance of a paper, the fame of the authors, the fame of the publishing journal, the influence of the peer review process, the paper's topic, etc. In fact, this model ignores *everything* about the papers themselves except for their degree. In reality, the probability that a vertex gains a citation *cannot* be precisely proportional to its degree because every paper is born with zero citations. Price circumvented this problem by letting the probability be proportional to the current number of citations k plus a constant r, which could be interpreted as a number of "free" citations every new paper receives or a certain fraction of all citations that are distributed uniformly at random.[6] Finally, we must also specify an initial network to which new vertices attach, e.g., two vertices joined by a single edge.

[5] This particular assumption is rather unrealistic, however, because it assumes that every scientist who writes a new paper knows the distribution of citations for all other papers every written. An equivalent and more realistic mechanism that has the consequences, however, is the following. To choose which paper to cite, a scientist chooses an arbitrary paper and cites a uniformly random paper listed in its bibliography. This is equivalent to choosing a random neighbor of a random vertex, which, in a random graph, leads to choosing a vertex with probability proportional to its degree.

[6] This constant r thus plays a role similar to the "teleportation" probability in the PageRank model of vertex centrality. It also implies that the way a paper accumulates citations varies depending on which of these two mechanisms is larger; for young papers with small degrees, the uniform attachment mechanism should dominate, but older papers, with larger degrees, will mainly gain new citations from the preferential attachment mechanism.

2.3 Structural patterns of preferential attachment

With these simple assumptions in place, we can mathematically analyze the entire dynamics of the model and what kinds of network structures it produces. We will sketch some of this analysis in the next section. In general, however, the main consequence of Price's assumptions is that the degree distribution of papers exhibits a power-law tail $\Pr(k) = L(x)\, x^{-\alpha}$ where $\alpha = 2 + r/c$, the variable r is the "uniform attachment" mechanism described above and c is the average degree of a vertex joining the network. In the special case studied by Barabási and Albert, they chose $r = c$ yielding a power-law distribution with scaling exponent exactly $\alpha = 3$.

A second consequence of this model is a strong correlation between the "age" of a vertex, i.e., how early in the network-growth process it joined the network, and its degree. Basically, the longer a vertex is in the network, the more chances it has to accumulate additional edges, and further, older vertices tend to have a larger share of citations and thus they gain additional edges faster. The result is that the oldest vertices tend to have the largest degrees. This effect is sometimes called the "first-mover advantage."

Another consequence is that, like a random graph with heavy-tailed degree distribution, the highest-degree vertices tend also to have high closeness and betweenness centrality scores. This creates a kind of global *core-periphery* structure, in which the high-degree vertices cluster together in the center of the network, surrounded by a sea of lower degree vertices. This also induces degree dis-assortativity, with high-degree vertices linking to each other, but mainly to many very low degree vertices (which tend to be very young).

Finally, many variations of this model have been studied, including alterations to the attachment function. The traditional version, in which the probability of attachment is proportional to the degree is called *linear preferential attachment*. A simple generalization is to take a power of the attachment probability, like so

$$q_i = \left(\frac{r + k_i}{\sum_j (r + k_j)} \right)^{\gamma} \quad ,$$

where $\gamma = 1$ returns Price's linear attachment model. When $\gamma > 1$, the attachment behavior is super-linear. It can be shown that in this case a *condensation* or *winner-take-all* effect happens, and asymptotically one vertex (the highest degree one) will gain all new edges. When $\gamma < 1$, the attachment behavior is sub-linear, and the distribution of new connections is more equitable across the network. In the limit $\gamma \to 0$, we return to a kind of randomly grown network where the connection probabilities are iid variables, like in $G(n, p)$.

2.4 A mathematical sketch of the degree distribution

The full derivation of the degree distribution's form is given in *Networks*. Here, we will briefly sketch the mathematical form predicted by Price's model. The linear attachment function is given by

$$q_i = \frac{r + k_i}{\sum_j (r + k_j)} = \frac{r + k_i}{nr + n\langle k \rangle} = \frac{r + k_i}{n(r + c)} \ ,$$

where we use the definition of the average degree $\langle k \rangle = \frac{1}{n}\sum_j k_j$ and the fact that $\langle k \rangle = c$ by definition. When a new vertex joins the network, it distributes on average c new connections to the other vertices. We can model the fraction of vertices with degree k, denoted p_k, by using the master equation approach to write down a set of coupled equations that represent how those populations change over time and then letting the size of the network $n \to \infty$. This yields the expressions

$$\begin{aligned} p_k &= \left(\tfrac{r+(k-1)}{(r+1+r/c)+k}\right) p_{k-1} && \text{for } k \geq 1 \ , \\ p_0 &= \left(\tfrac{1+r/c}{r+(1+r/c)}\right) && \text{for } k = 0 \ . \end{aligned} \tag{1}$$

The second equation is necessary because we are only modeling the in-degree distribution of vertices. Iterating these recursive equations and recognizing some patterns in their functional form, we find that they can be expressed as

$$p_k = \frac{\mathrm{B}(k + r, 2 + r/c)}{\mathrm{B}(r, 1 + r/c)} \ ,$$

where $\mathrm{B}(x, y)$ is Euler's beta function. In this form, it is not obvious that p_k follows a power-law distribution. However, it can be shown[7] that a ratio of Beta functions is approximately a shifted power-law distribution,[8]

$$\begin{aligned} p_k &\propto (k + r)^{-\alpha} \\ &\approx k^{-\alpha} \qquad \text{for } k \gg r \ , \end{aligned} \tag{2}$$

where $\alpha = 2 + r/c$ and when $r = c$, we have $p_k \approx k^{-3}$.

The full distribution, given by the ratio of the two beta functions is called the Yule-Simon distribution, and was first derived by Herbert Simon[9] in 1955.

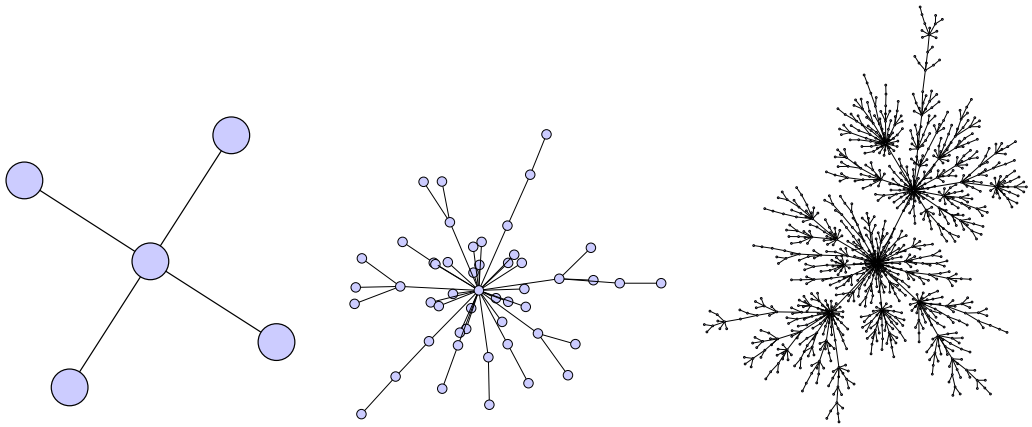
[7] First, recall that $\mathrm{B}(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$, where $\Gamma(x)$ is the gamma function, a continuous-variable generalization of the standard factorial $x! = x(x - 1)(x - 2)\ldots$. Stirling's approximation for the gamma function is $\Gamma(x) \simeq \sqrt{2\pi}\mathrm{e}^{-x}x^{x-1/2}$, which allows us to re-express $\mathrm{B}(x, y)$ in closed form. Then applying the approximation $(x + y)^z \simeq x^z\mathrm{e}^y$, yields $\mathrm{B}(x, y) \simeq x^{-y}\Gamma(y)$, which decays like a power law for $x \gg 1$.

[8] This mathematical structure, the ratio of two slightly offset Gamma functions, appears in the analysis of many simple models of network structure and always yields a power law form.

[9] Simon was a giant of an intellect, and was into "complex systems" several decades before the term was coined. His book *The Sciences of the Artificial* is highly recommended.

2.5 Simulation of the model

Price's model is easy to simulate (Matlab code is at the end of this file). The following figure shows the results of a single simulation (ignoring the direction of the edges) with $r = c = 1$ for $n = \{5, 50, 1000\}$. Because the average degree is $c = 1$, the network is always a tree. Note the large degree heterogeneity emerging, even at modest values of n.



2.6 Empirical tests of the model

The empirical support for Price's model largely comes from indirect comparisons of the model with data. That is, from comparing the predictions of the model on certain structural regularities with empirical tabulations of those patterns. The first of these was done by Price himself, looking at the degree distribution of real citation networks. This comparison continues to be the dominant test of the model. However, since many models of network growth are known to produce heavy-tailed or even power-law degree distributions, agreement here is not a very powerful test of the model.

Given full bibliographic information about a set of papers, that is, their publication date and the set of previous papers they cite, the there are no free parameters in the model. However, if we do not have the arrival times of the vertices, i.e., we only have a current snapshot of the structure of the network such that we can see which vertices connect to which other ones, Price's model can be cast in terms of likelihood functions and fitted directly to the network structure. This leads to estimates of its free parameters r, c and the arrival times of the vertices. The inference step is mainly to search through the permutations of the n vertices to find the one under which the observed topology is most likely.[10]

[10]This inference problem was recently formalized by Wiuf et al. in 2006; however, it's much harder than it sounds

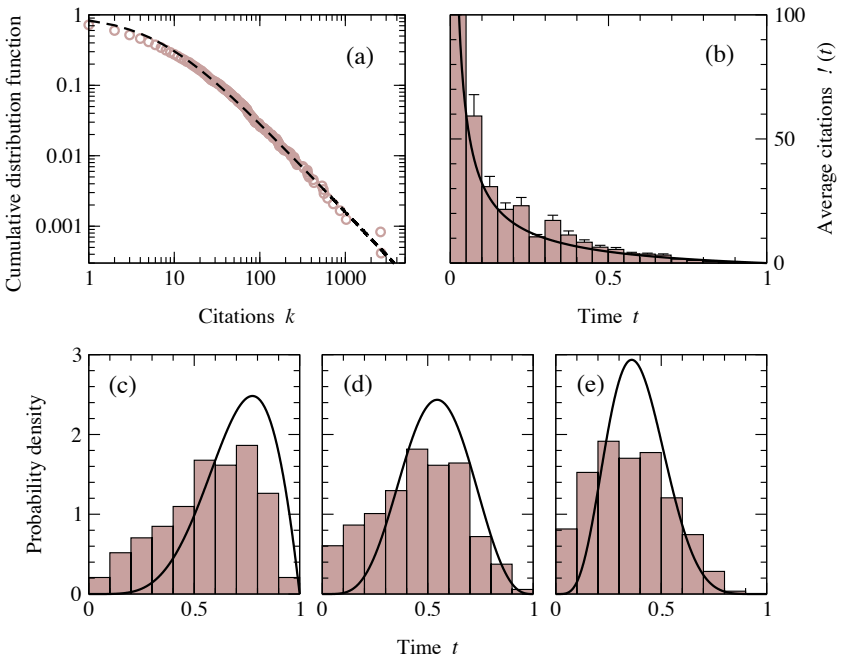


Figure 1: Empirical measurements in brown; theoretical predictions in black. (a) The ccdf of the degree distribution. The best fit is achieved for $\alpha = 2.28$ and $r = 6.38$. (b) The mean number of citations received by papers as a function of time from beginning ($t = 0$) to end ($t = 1$) of the period covered. (c), (d) and (e): Probability that a paper with a given number of citations is published at time t, for papers with (c) 1 or 2 citations, (d) 3 to 5 citations, and (e) 6 to 10 citations at time $t = 1$. Figure reproduced from M.E.J. Newman, *Eur. Phys. Lett.* **86**, 68001 (2009).

If the arrival times are known, we can estimate r and c directly from the empirical degree distribution by fitting the predicted form to the empirical data. We can then make effectively zero-parameter predictions about the local structure of the network. This was recently done by Mark Newman in a 2009 paper on the first-mover advantage in which he applied Price's model to the evolution of the citation network of papers on the theory of networks. Figure 2 shows some of his

because the structure of a network evolving under the preferential attachment mechanism exhibits a strong dependence on its past, which makes estimation of the likelihood of the data given a choice of arrival times of the nodes difficult. In general, all evolving network models exhibit the same problem and thus it becomes easier to work with indirect comparisons of the model with data.

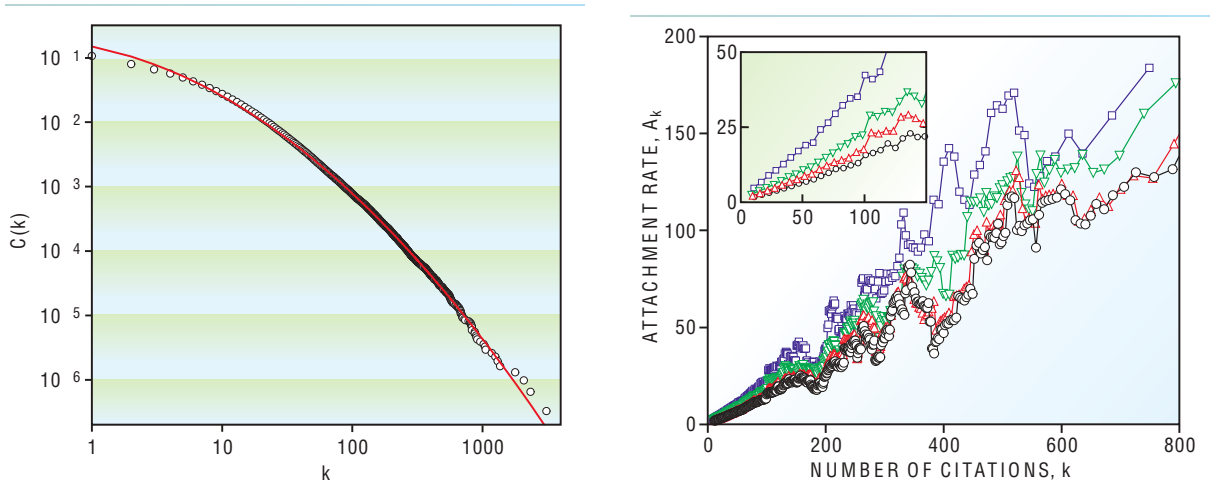


Figure 2: (a) The degree distribution of 353,268 papers (which generated 3,110,839 citations) published in the *Physical Review* journals from July 1893 – June 2003 (in this case, fitted with a left-truncated log-normal distribution). (b) An empirical estimate of the attachment rate q_k for this network (denoted A_k in the figure), showing a roughly linear shape, particularly for the first 100 or so citations (inset). The different colors denote different time periods for establishing k: 1990-99 (purple), 1980-99 (green), 1970-99 (red) and 1893–1999 (black). Figures reproduced from S. Redner, *Phys. Today* **58**, 49–54 (June 2005).

results. In the first step, the Yule-Simon distribution was fitted to the degree distribution via maximum likelihood to recover estimates of r and c. This then fully specifies the model and additional predictions, e.g., of the average citation count of a paper as a function of its age or of the probability that a paper with k citations was published at some time t, can be derived and compared with the empirical results. Perhaps unsurprisingly, Newman's results show a very strong preferential attachment mechanism among papers on network theory. He goes on, however, to use Price's model as a kind of null model and shows that some papers receive more citations than we would expect, given the age. The implication is that there is some contribution to the citation dynamics from the many aspects not represented by preferential attachment, e.g., the quality of the paper.

These tests mainly focus on testing the outcomes or predictions of the model, however, a crucial step to validating any hypothesis is to test the validity of its inputs or assumptions. The physicist Sid Redner conducted such a test in 2005 by analyzing 110 years of bibliographic data for the journals *Physical Review*, covering 353,268 papers and 3,110,839 citations. Most notably, he directly measured the form of the attachment function q_k and showed that it exhibits a plausibly linear

structure, particularly for the first 100 or so citations. Figure 3 illustrates his results. He also found, however, that citation networks exhibit a host of rich dynamics that have extremely low probability under Price's model. For instance, there are "sleeper" classics, which receive very few citations for a long period of time, but then suddenly begin accumulating large numbers of new connections, e.g., because an important paper was forgotten and then rediscovered.

Price's preferential attachment mechanism has been suggested as the underlying explanation of many other networks' structure, including the topology of the Internet at the level of Autonomous Systems, the evolution of the WWW, the structure of online social networks like Facebook and Twitter, and even some biological networks. However, in most cases, the tests of the model's accuracy have mainly compared the empirical and predicted degree distributions. Few have tested whether the attachment function exhibits linear behavior (as Redner did) or whether other predictions also line up (as Newman did).

3 At home

1. Peruse Chapter 14.0–14.4 (pages 486–534) in *Networks*
2. Next time: more network mechanisms

4 Matlab code

```
% the preferential attachment mechanism
n = 1000;       % size of network
p = 0.1;        % probability of uniform attachment
x = zeros(n,1);
d = zeros(n,1);
% initial condition: pair of reciprocal edges
x(1)   = 2;     % node 1 points to node 2
x(2)   = 1;     % node 2 points to node 1
d(1:2) = [1 2]; % each has degree 1
for t=3:n
    if rand(1)<p
        % uniform attachment
        g = ceil((t-1).*rand(1));
    else
        % preferential attachment
        g = d(ceil((t-1).*rand(1)));
    end;
    x(t) = g;
    d(t) = g;
end;

% plot the degree distribution as a ccdf
pdf = hist(x,unique(x));
cdf = [[unique(x); length(pdf)+1] 1-[0 cumsum(pdf./sum(pdf))]'];
cdf(cdf(:,2)<1/n,:) = [];
figure(1);
loglog(cdf(:,1),cdf(:,2),'ro');
set(gca,'FontSize',16);
```