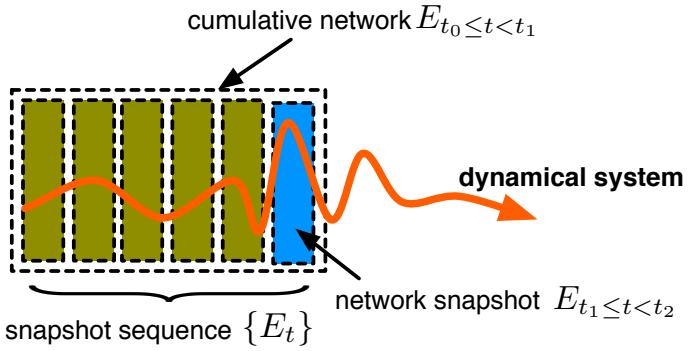
1 Modeling network growth

Most real-world networks are representations of an underlying dynamical systems, i.e., a system that changes over time, and the way we record or sample those changes over time can have a large impact on what structure we then observe.

In many cases, what appears to be a "static" network data set may in fact be a snapshot of the system's structure, taken at a particular moment in time t, recording the particular set of vertices and particular edges observable over some window of time $t_1 \leq t < t_2$. Or, we might have a sequence of such snapshots for some set of sampling times $\{t_0, t_1, \ldots, t_k\}$, or we might have a cumulative network, which includes all edges and vertices that ever occurred up to our observation time t. Even today, it remains relatively unusual to have a complete trace of the system's evolution, with timestamps for each change in the set of vertices or edges. At their most basic level, temporal (or evolving or dynamical) networks can change over time by either changing the number of vertices, or changing the number of edges they contain.



Growing networks are a subset of these temporal networks that represent those networks for which the number of vertices only grows over time. Two notable examples of growing networks relate to scientists and the papers the publish:[1]

- the network of *scientific citations*, where vertices are scientific papers and two papers are connected if one of them lists the other in its bibliography, and
- the network of *scientific collaborations*, where vertices are published scientists, and two scientists are connected if they have ever coauthored a paper together.

[1]Both of these are actually derived from the larger scientific literature network, in which vertices are either scientists or papers. Each scientist vertex connects to all the papers on which that scientist is listed as a coauthor, and each paper vertex links to all the papers that it cites in its bibliography. This is not actually a bipartite network, since there are links among the papers (but not among the scientists). The scientific collaboration network is the one-mode projection onto the scientists of the scientific literature network.

These networks are strictly growing because the set of nodes and edges are both cumulative over time. (One network is directed. The other is undirected. Do you see why?)

Similarly, we may define *shrinking networks* as those networks in which the number of vertices only shrinks over time. These two classes—always growing or always shrinking—define the extremes of a spectrum for temporal networks. The precise middle of this spectrum is the set of networks for which the number of vertices (and possibly also the number of edges, e.g., if the mean degree is constant) exists in a kind of dynamic equilibrium. These networks experience network "churn," in which new nodes join and new edges form at roughly the same rate that old nodes leave and old edges disappear. Across the literature of network science, the vast majority of work has been devoted to growing networks. In this lecture, we will investigate one of the better known network models for growing networks, called *preferential attachment*, which has been rediscovered many times, but was introduced first for networks by Derek De Solla Price in the 1960s.

A small number of studies have focused on shrinking or dynamically stable networks.[2] It is not clear why there is such a strong difference, but it may be due, in part to the fact that much data on networks comes from digital sources, and most of those sources represent expanding networks (like Facebook, Twitter, Instagram, TikTok, etc.) for fairly straightforward reasons.

Introducing time into our representation of networks presents several new opportunities for network analysis and modeling. It also introduces more directly the question of mechanism, that is, identifying the causes for the effects that we observe in the data.

1.1 Thinking about network mechanisms

Analyzing the structure of a network, e.g., using descriptive measures like degree distributions, degree assortativity, clustering coefficients, centrality measures, 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.

[2] Good studies of shrinking networks include Saavedra, Reed-Tsochas, and Uzzi, "Asymmetric disassembly and robustness in declining networks." *PNAS* **105**, 16466–16471 (2008) and Garcia, Mavrodiev, and Schweitzer, "Social Resilience in Online Communities: The Autopsy of Friendster." ACM Conference on Online Social Networks (COSN 2013) `http://arxiv.org/abs/1302.6109`. One study on dynamically stable networks is Merritt and Clauset, "Social Network Dynamics in a Massive Online Game." Workshop on Mining and Learning with Graphs (MLG 2013) `http://arxiv.org/abs/1306.4363`.

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.[3] 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.[4] 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.

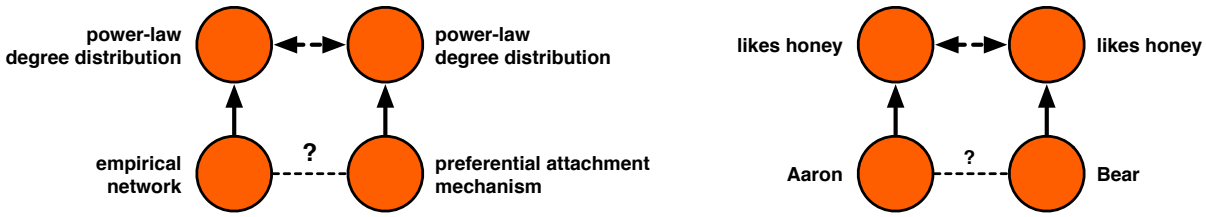
In this lecture, we will investigate these ideas in the context of the popular preferential attachment mechanism for network growth, whose signature output is a network with a power-law degree distribution. But, a power-law degree distribution can be produced by many network mechanisms. This implies 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, which describe "The Honey Bear Test" for complex systems. The left shows the usual situation with trying to determine whether some empirical network is governed by the preferential attachment mechanism.[5] The right shows

[3] 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.

[4] Social scientists have even examined 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 & 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).

[5] 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 like that described in Clauset, Shalizi and Newman, "Power-law distributions in empirical data." *SIAM Review* **51**(4), 661–703 (2009).



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 Aaron am a Bear, examine whether Aaron 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

Perhaps the best known of all mechanisms for network growth is *preferential attachment*. 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 nothing 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.

2.1 A brief history of many rediscoveries

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 scientomet-

rics), 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.[6]

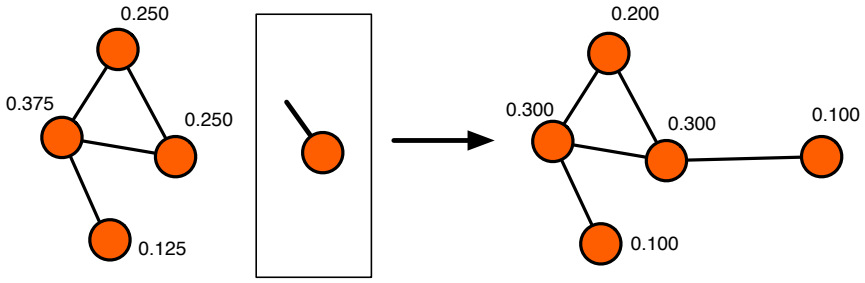
The central question for determining the evolution of the network structure is, How does a new paper choose which previous papers to cite? Price's assumption was those papers are chosen at random with probability proportional to the number of citations those previous papers already have.[7] Thus, highly 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

[6]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.

[7]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.

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.[8] Finally, we must also specify an initial network to which new vertices attach, e.g., two vertices joined by a single edge.

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 $\alpha = 3$.

[8] 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.

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 $G(n, p)$. (That being said, randomly grown networks are subtly different from Erdős-Rényi random graphs; for instance, if the seed graph is connected, then the grown network, no matter how large, is also connected.)

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

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

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

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

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[9] that a ratio of Beta functions is approximately a shifted power-law distribution,[10]

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

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[11] in 1955.

2.5 Simulation of the model

Price's model is easy to simulate (Matlab code is at the end of this file). Figure 1 (on the next page) 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.

[9] 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 $x \gg 1$.

[10] 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.

[11] 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.

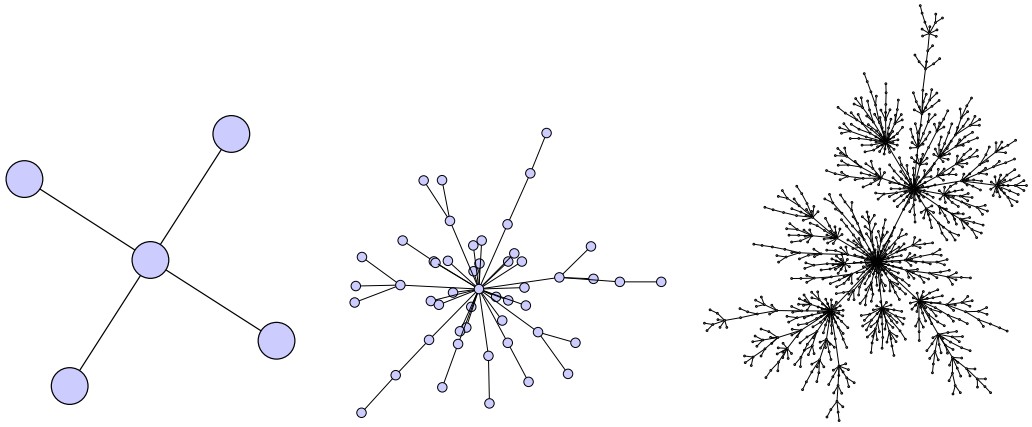


Figure 1: Price's model for preferential attachment with $r = c = 1$ for $n = \{5, 50, 1000\}$ vertices. Because $c = 1$, each grown network must be a tree. In the middle- and right-hand figures, one can see the emergence of very high-degree vertices. The right-most figure also illustrates the fractal structure of these networks.

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.[12]

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

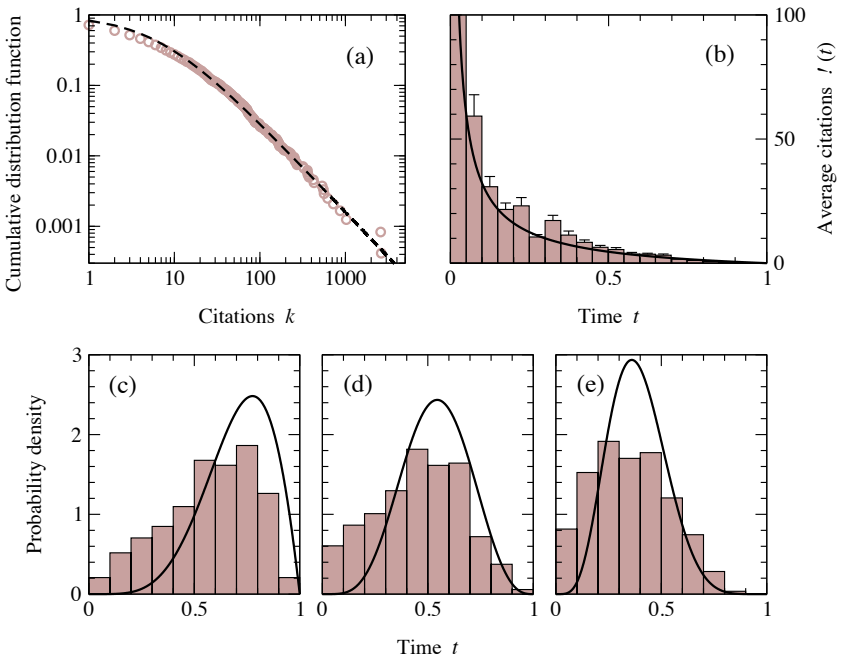


Figure 2: Empirical measurements in brown; theoretical predictions in black. (a) The ccdf of the degree distribution. The best fit is achieved for $\alpha = 2.28$ and $r = 6.38$. (b) The mean number of citations received by papers as a function of time from beginning ($t = 0$) to end ($t = 1$) of the 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 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

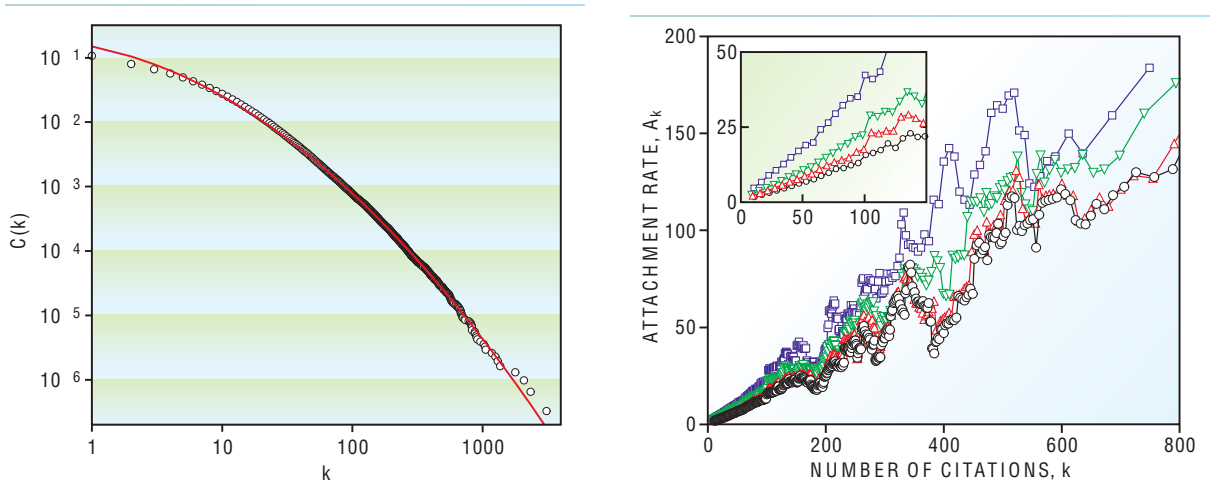


Figure 3: (a) The degree distribution of 353,268 papers (which generated 3,110,839 citations) published in the *Physical Review* journals from July 1893 – June 2003 (in this case, fitted with a left-truncated log-normal distribution). (b) An empirical estimate of the attachment rate q_k for this network (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).

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 The vertex copy mechanism

Although preferential attachment and its variations are perhaps the most widely known (and rediscovered) mechanisms for producing a power-law degree distribution in a growing network, it is not the only class of such mechanisms. In fact, there are now dozens of other models that produce this particular pattern and some are substantially more plausible explanations for other types of networks, e.g., networks among molecules, as in gene regulation or protein interaction networks.

These types of networks also exhibit heavy-tailed degree distributions, and the best current explanation for their structure is a class of "vertex copying" models, which are also called duplication-mutation or duplication-mutation-complementation models in biology. However, the mechanism is simple and can also be expressed as a form of document network model, which the textbook *Networks* describes. The first instance of such model is considerably more recent than the first instance of preferential attachment, originating with Kleinberg and colleagues in 1999, which was specifically a model of how the World Wide Web grows over time. A similar model was also introduced by Vazquez and colleagues in 2003 as an explicit model of protein-protein interaction networks.[13]

3.1 The basic mechanism

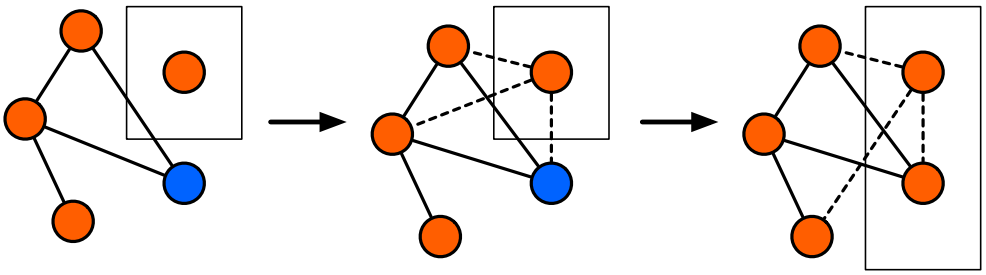
Recall from Price's model of a growing citation network that each new vertex added to the network includes c edge stubs. Each of these stubs is attached to an existing vertex i with probability proportional to its degree k_i. Alternatively, we can imagine flipping a coin: with probability q, we attach preferentially, and with probability $1-q$ we attach the edge uniformly at random.

The vertex copying models take a variation on this idea: instead of choosing a different vertex for each edge, we first choose a single vertex uniformly at random, and then copy all of its edges. That is, we copy or duplicate an existing vertex, and all of its connections, in order to grow the network.

[13] See Kleinberg, Kumar, Raghavan, Rajagopalan, and Tomkins, "The Web as a graph: Measurements, models and methods." In *International Conference on Combinatorics and Computing* (1999), and Vazquez, Flammini, Maritan and Vespignani, "Modeling of protein interaction networks." *Complexus* **1**, 38–44 (2003) `arXiv:cond-mat/0108043`. For an excellent discussion of the accuracy of the vertex copy model for protein interaction networks, see Middendorf, Ziv, and Wiggins, "Inferring network mechanisms: The Drosophila melanogaster protein interaction network." *PNAS* **102**(9), 3192–3197 (2005).

This model presents a similar problem to a "pure" preferential attachment model, which is that only vertices that already have connections can gain new ones. A modification much like the one we made for preferential attachment solves this problem: instead of copying every connection, we flip a coin for each connection, so that with probability q we copy that connection and with probability $1-q$ we use the uniform attachment mechanism.

The following figure illustrates this process, in which the blue vertex is the one chosen to duplicate. As with preferential attachment, there are now many variations of the basic model, some of which assume that the new vertex always adds a connection to the copied vertex (as done in the figure), and others of which also modify the connections of the copied vertex. These variations are common in biological flavors of the model, as they are intended to capture the impact of specific biological processes on the structure of the network, e.g., the tendency for genes to occasionally be duplicated during the DNA replication process, the tendency for functionality to diverge among duplicate copies of a gene, and the tendency for proteins to stick to themselves (as in polymerization). Notably, all such models share the common assumption of network growth, and relatively few models consider the more general case of allowing vertices to be removed from the network (gene extinction). Here we will ignore these alternatives and focus on the basic model.



Much like the preferential attachment mechanism, the vertex copying mechanism will tend to produce a "rich get richer" effect, as vertices with many connections to them have a higher probability that one of their neighbors will be chosen for copying, and when such an event occurs, their degree will increase.

3.2 The degree distribution

There are two ways the degree of some vertex i could increase. Either one of i's neighbors is copied by the new vertex, in which case with probability q the connection to i will also be copied, or it is chosen directly for uniform attachment.

The probability that any particular vertex is chosen to be copied is $1/n$, where n is the current size

of the network. If vertex i has degree k_i already, then the probability that such a randomly chosen vertex connects to i is k_i/n. Because each connection from the copied vertex is preserved independently with probability q, the probability that i increases its degree as a result of the copying step is $k_i q/n$.

The probability that i receives a new connection as a result of the uniform attachment depends on the mean degree of the network. As in Price's model, we fix this value at c. Because connections of the copied vertex are copied independently with probability q, the number the copied vertex's connections that will be discarded is $(1-q)c$, each of which is replaced with uniformly random connection for the new vertex. Thus, the probability that i receives one of these connections is $(1-q)c/n$.

Combining these terms yields the total probability that vertex i will increase its degree:

$$\begin{aligned}\Pr(k_i \to k_i + 1) &= \frac{k_i q}{n} + \frac{(1-q)c}{n} \\ &= \frac{k_i q(1-q)c}{n} \ .\end{aligned}$$

For a network with n vertices, let $p_k(n)$ denote the fraction of these with degree k. Thus, the expected number of such vertices receiving a new connection is

$$\begin{aligned}n\, p_k(n) \times \frac{k_i q(1-q)c}{n} &= [k_i q(1-q)c]\, p_k(n) \\ &= \frac{c(k+a)}{c+a} p_k(n) \ , \end{aligned} \tag{3}$$

where we have employed a clever change of variables, letting $a = c(q^{-1} - 1)$, which implies $q = c/(c+a)$. The form of Eq. (3), which counts the number of vertices in the network that will increase their degree at any particular step of the growth process, is exactly the same as the same expression for the preferential attachment model. This symmetry implies that the degree distribution for the vertex copy model also follows a power law $p_k \propto k^{-\alpha}$, but with an exponent

$$\alpha = 2 + \frac{a}{c} = 1 + \frac{1}{q} \ . \tag{4}$$

The fidelity or accuracy of the copy mechanism thus tells us how heavy a tailed distribution the mechanism produces. Perfect or near-perfect copying yields exponents at or close to 2, while poor copying yields larger exponents. If a particular class of networks could be shown to follow this mechanism in general, then we could estimate the accuracy of the copying by estimating α from the empirical data and then applying Eq. (4).

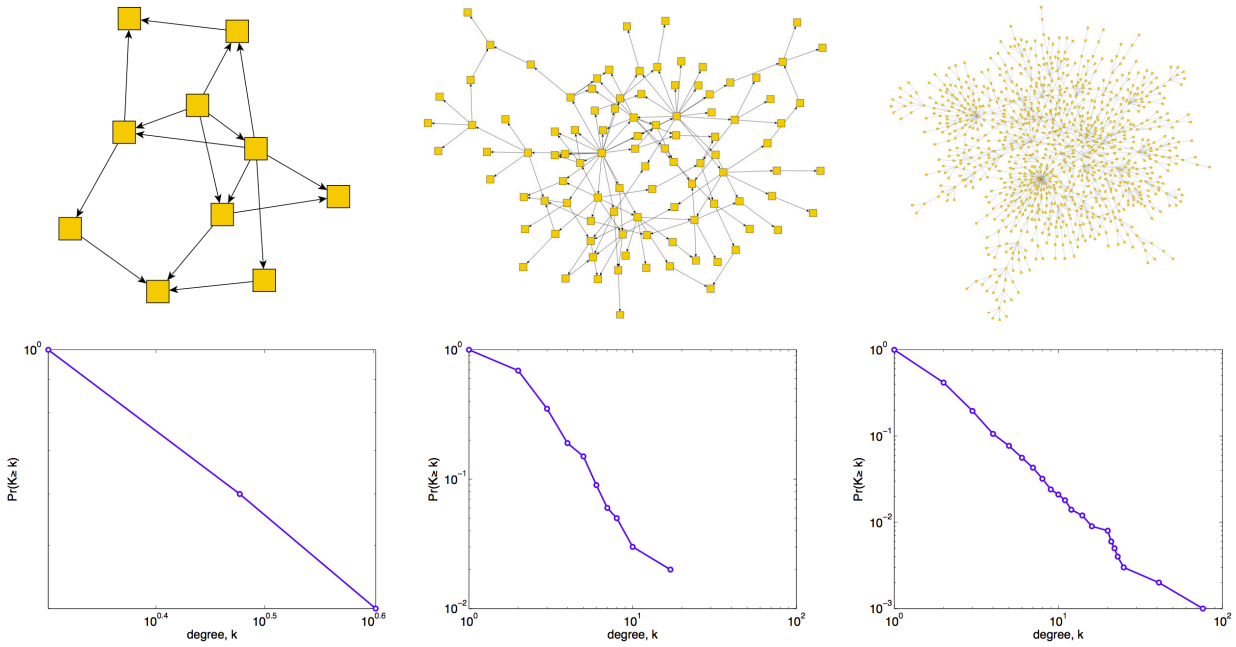


Figure 4: The vertex copy model with $q = 1/2$ for $n = \{10, 100, 1000\}$ vertices. The mean degree here is close to 1, which makes the large networks very tree-like. As with Price's model, the middle- and right-hand figures show the emergence of very high-degree vertices.

The symmetry with Price's model also implies that its other properties also carry over, including the tendency for the oldest vertices to have the largest degrees. Not all the properties carry over, however. For instance, vertex copying will tend to produce greater local clustering (triangles) as a result of copying many connections from a single vertex, while preferential attachment tends to distribute a new vertex's connections more broadly across the network.

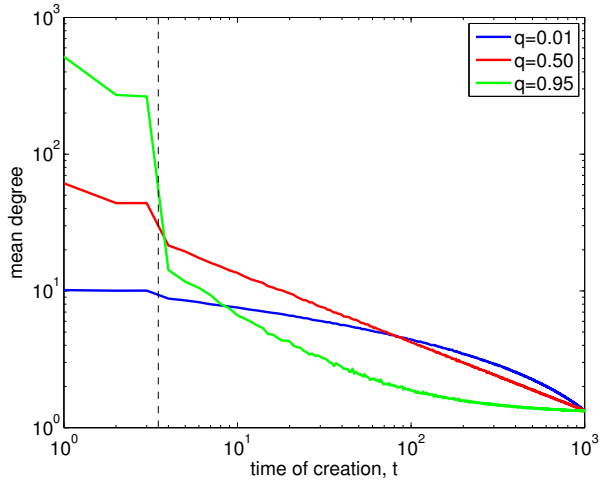
3.3 Simulation of the model

Vertex copying is also easy to simulate (Matlab code is at the end of this file), and Figure 4 show network examples and degree distributions (ignoring the direction of edges) with $q = 1/2$.

What is noticeable about these networks, as compared to those grown by preferential attachment, is the prevalence of short loops, particularly for small n. As n increases, the heavy tail becomes more prominent, the variance increases, and we an easily spot high-degree vertices within the network. But even here, the network visibly shows local density, reflecting the local nature by which

it distributes edges.

The correlation between age and degree remains, however, which we can see via a simple simulation of many network instances. The figure below shows the mean degree as a function of the time-of-creation for each vertex, averaged across 10,000 networks, and for three choices of the copy probability q. As expected, as q approaches 0, connection copying occurs more rarely, and most vertices have similar degrees. In contrast, as q approaches 1, very few connections are rewired, leading to a greater concentration of edges among the oldest vertices. (The dashed black line shows the transition from the original seed network to the portion of time when the network is growing.)



Supplemental readings

1. Chapter 14.0–14.4 (pages 486–534) in *Networks* (preferential attachment)
2. Chapter 14.5–14.6 (pages 534–548) in *Networks* (vertex copying)
3. Overgoor, Benson, and Ugander, "Choosing to Grow a Graph: Modeling Network Formation as Discrete Choice." *Proc. The Web Conference* (WWW '19) (2019) `https://arxiv.org/abs/1811.05008`

4 Matlab code

```
% the preferential attachment mechanism
n = 10^6;    % size of the network
c = 3;       % k_out
r = 1;       % uniform attachment contribution
p = c/(c+r); % attachment probability
x = zeros(1,c*n); % stores the graph
x(1:12) = [2 3 4 1 3 4 1 2 4 1 2 3]; % a 4-clique seed graph
for t=5:n
     for j=1:c           % for each out-edge
          if rand(1)<p   % choose by preferential attachment
               d = x(randi(c*(t-1),1));
          else           % choose by uniform attachment
               d = randi(t-1,1);
          end
          x(c*(t-1)+j) = d; % record the attachment
     end
end
degs = hist(x,(1:n)); % in-degree sequence

% plot the degree distribution as a ccdf
hx = (0:max(degs))';
hc = hist(degs,hx)'./n;
hc = [[hx; hx(end)+1] 1-[0; cumsum(hc)]];
hc(hc(:,2)<10^-10,:) = [];
figure;
loglog(hc(2:end,1),hc(2:end,2),'r-','LineWidth',3);
set(gca,'FontSize',16);
```

```
% the vertex copy mechanism
n = 1000;       % size of network
q = 0.5;        % probability to copy a connection
x = zeros(n,2); % edge list
% initial condition: a reciprocally connected triple
x(1,:) = [1 2]; % node 1 points to node 2
x(2,:) = [2 1]; % node 2 points to node 1
x(3,:) = [2 3]; % node 2 points to node 3
x(4,:) = [3 1]; % node 3 points to node 2
m = 4;          % number of edges
% start copying
for t=4:n
     v  = ceil(rand(1)*(t-1));    % vertex to copy
     g  = x(x(:,1)==v,2);         % copied endpoints
     k  = length(g);              % its degree
     u  = ceil(rand(k,1).*(t-1)); % uniformly random endpoints
     s  = rand(size(g,1),1)<q;    % these endpoints are copied
     g(~s) = u(~s);               % these endpoints are not
     % make those edges
     x(m+1:m+k,:) = [t*ones(k,1) g];
     m = m+k;                     % increment edge count
end;

% make an undirected adjacency matrix
B = zeros(n,n);
for i=1:size(x,1)
     if x(i,2)~=x(i,1)
          B(x(i,1),x(i,2)) = 1;
          B(x(i,2),x(i,1)) = 1;
     end;
end;
degs = sum(B); % get degree sequence

% plot the degree distribution as a ccdf
pdf = hist(degs,unique(degs));
cdf = [[unique(degs)'; length(pdf)+1] 1-[0 cumsum(pdf./sum(pdf))]'];
cdf(cdf(:,2)<1/n,:) = [];
figure(1);
loglog(cdf(:,1),cdf(:,2),'bo-','LineWidth',2,'MarkerFaceColor',[1 1 1]);
set(gca,'FontSize',16);
xlabel('degree, k','FontSize',16);
ylabel('Pr(K\geq k)','FontSize',16);
```