1 Vertex copying models

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. Indeed, 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 biological molecules like 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 *Networks* describes. The first instance of such model is considerably more recent than the first instance of preferential attachment, originating with Kleinberg et al. in 1999, which was a model of the web, and then again with Vazquez et al. in 2000 as an explicit model of biological networks.

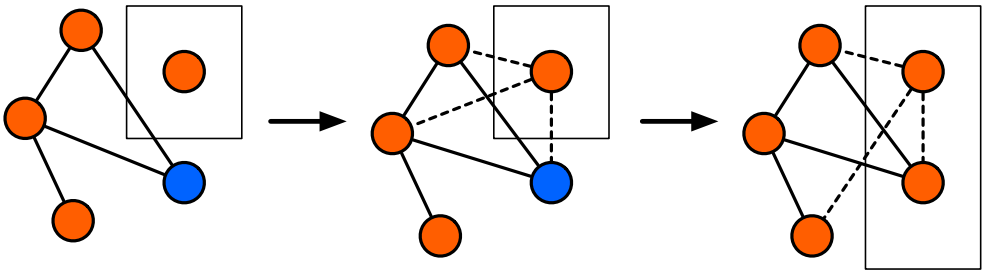
1.1 The basic mechanism

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. 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.

1.2 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

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

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. (1), 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

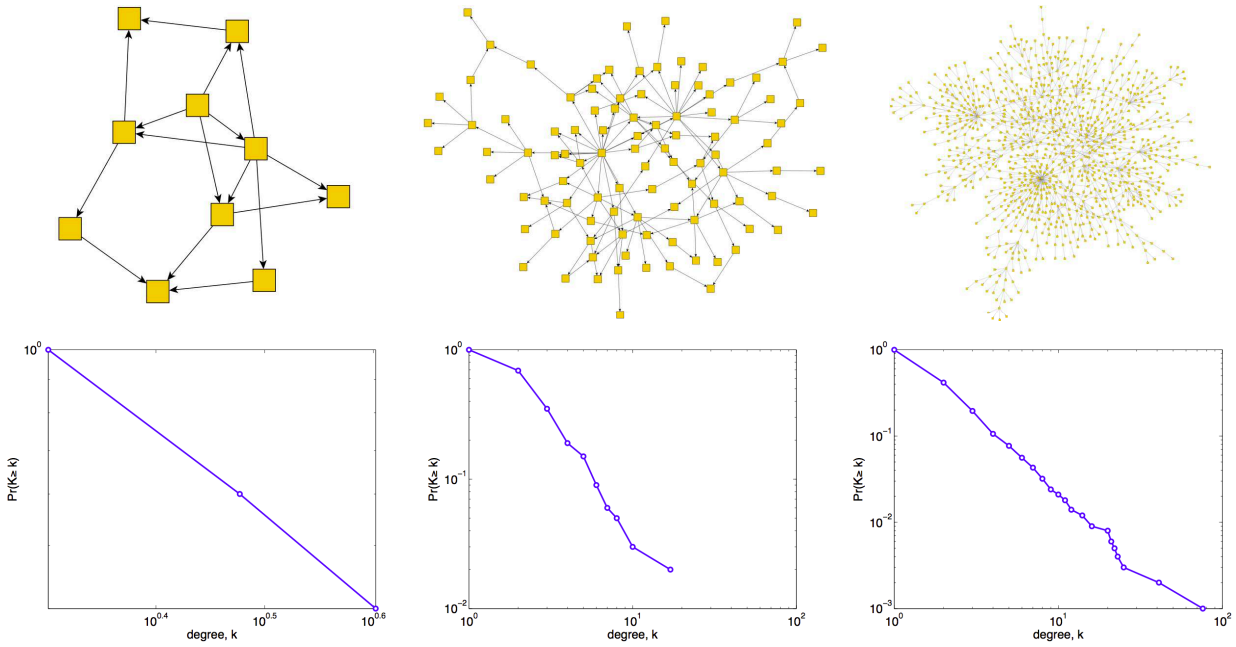
$$\alpha = 2 + \frac{a}{c} = 1 + \frac{1}{q} \ . \tag{2}$$

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. (2).

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.

1.3 Simulation of the model

Vertex copying is also easy to simulate (Matlab code is at the end of this file), and the following figures show network examples and degree distributions (ignoring the direction of edges) with $q = 1/2$ for $n = \{10, 100, 1000\}$. 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

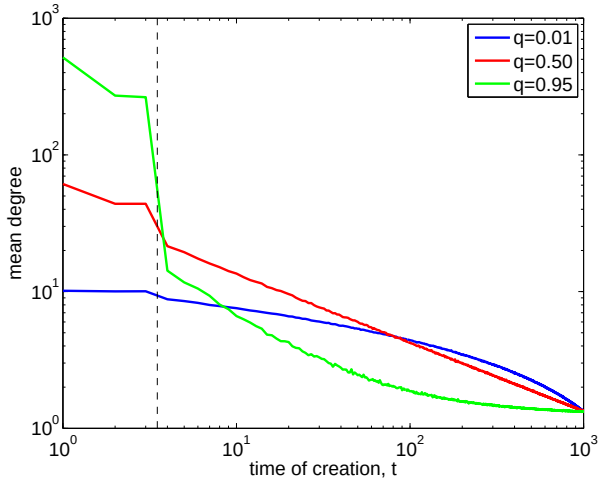


vertices within the network. But even here, the network visibly shows local density, reflecting the local nature by which it distributes edges.

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.)

2 At home

1. Read Chapters 14.5–14.6 (pages 534–548) in *Networks*
2. Next time: generative models of networks



3 Matlab code

```
% 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);
```