1 Mechanistic and generative models of network structure

There are many models of network structure, and these largely can be divided into two classes: *mechanistic* models and *generative* or probabilistic models. The boundaries between these classes, however, are not sharp.

A mechanistic model, generally speaking, codifies or formalizes a notion of causality via a set of rules (often mathematical) that produces certain kinds of networks. Identifying the mechanism for some empirically observed pattern allows us to better understand and predict networks — if we see that pattern again, we can immediately generate hypotheses about what might have produced it. In network models, the mechanisms are often very simple (particularly for mechanisms proposed by physicists), and these produce specific kinds of topological patterns in networks. We will explore examples of such mechanisms later in the semester, including the *preferential attachment* mechanism, which is known to underlie the evolution of scientific citation networks and the World Wide Web.

Generative models, on the other hand, often represent weaker notions of causality and generate structure via a set of free parameters that may or may not have specific meanings. The most basic form of probabilistic network model is called the *random graph* (sometimes also the Erdös-Rényi random graph, after two of its most famous investigators, or the Poisson or Binomial random graph). In this and other generative models, edges exist probabilistically, where that probability may depend on other variables. The random graph model is the simplest such model, where every edge is an iid random variable from a fixed distribution. In this model, a single parameter determines everything about the network.

The attraction of simple generative models is that nearly every question about their structure, e.g., the network measures we have encountered so far, may be calculated analytically. This provides a useful baseline for deciding whether some empirically observed pattern is surprising. Let G denote a graph, and let $\Pr(G)$ be a probability distribution over all such graphs. The typical or expected value of some network measure is then given by

$$\langle x \rangle = \sum_G x(G) \times \Pr(G) \ ,$$

where $x(G)$ is the value of the measure x on a particular graph G. This equation has the usual form of an average, but is calculated by summing over the combinatoric space of graphs.[1] In this lecture, we will study the simply random graph and derive several of its most important properties.

[1] We may also be interested not only in the mean value, but in the full distribution of x, although this can be trickier to calculate.

2 The Erdös-Rényi random graph

This network model is the original random-graph model, and was studied extensively by the Hungarian mathematicians Paul Erdös (1913–1996)[2] and Alfréd Rényi (1921–1970)[3] although it was studied earlier, as well.

This model is typically denoted $G(n, p)$ for its two parameters: n the number of vertices and p the probability that an edge (i, j) exists, for all i, j,[4] and these parameter specify everything about the model. The utility of the random graph model lies mainly with its mathematical simplicity, not in its realism. Virtually none of its properties resemble those of real-world networks, but they provide a useful baseline for our expectations and provide a warmup for more complicated generative models.

To be precise $G(n, p)$ defines an *ensemble* or collection of networks, which is equivalent to the distribution over graphs $\Pr(G)$. When we calculate properties of this ensemble, we must be clear that we are not of individual instances of the ensemble, but rather making statements about the typical member.

2.1 Mean degree and degree distribution

In the $G(n, p)$ model, every edge exists independently and with the same probability. The total probability of drawing a graph with m edges from this ensemble is

$$\Pr(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m} \quad , \tag{1}$$

which is a binomial distribution choosing m edges out of the $\binom{n}{2}$ possible edges. (Note that this form implies that $G(n, p)$ is an undirected graph.) The mean value can be derived using the Binomial Theorem:

$$\begin{aligned} \langle m \rangle &= \sum_{m=0}^{\binom{n}{2}} m \Pr(m) \\ &= \binom{n}{2} p \quad . \end{aligned} \tag{2}$$

That is, the mean degree is the expected number of the $\binom{n}{2}$ possible ties that exist, given that each edge exists with probability p.

[2] `http://xkcd.com/599/`

[3] "A mathematician is a machine for turning coffee into theorems."

[4] Another version of this model is denoted $G(n, m)$ which places exactly m edges on n vertices. This version has the advantage that m is no longer a random variable.

For any network with m edges, the mean degree of a vertex is $\langle k \rangle = 2m/n$. Thus, the mean degree in $G(n,p)$ may be derived, using Eq. (2), as

$$\begin{aligned}\langle k \rangle &= \sum_{m=0}^{\binom{n}{2}} \frac{2m}{n} \Pr(m) \\ &= \frac{2}{n}\binom{n}{2} p \\ &= (n-1)p \ . \end{aligned} \tag{3}$$

In other words, each vertex has $n-1$ possible partners, and each of these exists with the same independent probability p. The product, by linearity of expectations, gives the mean degree, which is sometimes denoted c. (Sometimes, it is mathematically convenient to use the asymptoticly equivalent expression pn.)

Because edges are iid random variables, the entire degree distribution has a simple form

$$\Pr(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \ , \tag{4}$$

which is a binomial distribution with parameter p for $n-1$ independent trials. What value of p should we choose? Commonly, we set $p = c/(n-1)$, where c is the target mean degree and is a finite value. (Verify using Eq. (3) that the expected value is indeed c under this choice for p.) That is, we choose the regime of $G(n,p)$ that produces *sparse networks*, where $c = O(1)$, which implies $p = O(1/n)$.

When p is very small, the binomial distribution may be simplified. When p is small, the last term in Eq. (4) may be approximated as

$$\begin{aligned}\ln\Big[(1-p)^{n-1-k}\Big] &= (n-1-k)\ln\left(1 - \frac{c}{n-1}\right) \\ &\simeq (n-1-k)\frac{-c}{n-1} \\ &\simeq -c \ , \end{aligned} \tag{5}$$

where we have used a first-order Taylor expansion of the logarithm[5] and taken the limit of large n. Taking the exponential of both sides yields the approximation $(1-p)^{n-1-k} \simeq \mathrm{e}^{-c}$, which is exact as $n \to \infty$. Thus, the expression for our degree distribution becomes

$$\Pr(k) \simeq \binom{n-1}{k} p^k \mathrm{e}^{-c} \ , \tag{6}$$

[5] A useful approximation: $\ln(1+x) \simeq x$, when x is small.

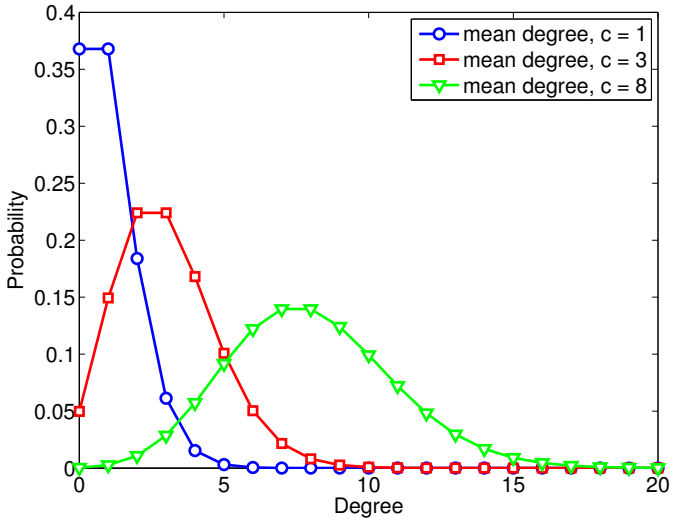
which may be simplified further still. The binomial coefficient is

$$\begin{aligned}\binom{n-1}{k} &= \frac{(n-1)!}{(n-1-k)!\,k!} \\ &\simeq \frac{(n-1)^k}{k!} \quad . \end{aligned} \tag{7}$$

Thus, the degree distribution is, in the limit of large n

$$\begin{aligned}\Pr(k) &\simeq \frac{(n-1)^k}{k!} p^k \mathrm{e}^{-c} \\ &= \frac{(n-1)^k}{k!} \left(\frac{c}{n-1}\right)^k \mathrm{e}^{-c} \\ &= \frac{c^k}{k!} \mathrm{e}^{-c} \quad , \end{aligned} \tag{8}$$

which is called the Poisson distribution. This distribution has mean and variance c, and is slightly asymmetric. The figure below shows examples of several Poisson distributions, all with $c \geq 1$. Recall, however, that most real-world networks exhibit heavy-tailed distributions. The degree distribution of the random graph model decays rapidly for $k > c$ and is thus highly unrealistic.



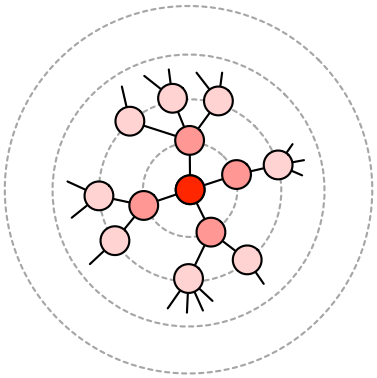
2.2 Clustering coefficient, triangles and other loops

The density of triangles in $G(n,p)$ is easy to calculate because very edge is iid. The clustering coefficient is

$$C = \frac{\text{(number of triangles)}}{\text{(number of connected triples)}} \propto \frac{\binom{n}{3}p^3}{\binom{n}{3}p^2} = p = \frac{c}{n-1} \ .$$

In the sparse case, this further implies that $C = O(1/n)$, i.e., the density of triangles in the network decays toward zero in the limit of large graph.

This calculation can be generalized to loops of longer length or cliques of larger size and produces the same result: the density of such structures decays to zero in the large-n limit. This implies that $G(n,p)$ graphs are locally *tree-like* (see figure below), meaning that if we build a tree outward from some vertex in the graph, we rarely encounter a "cross edge" that links between two branches of the tree.

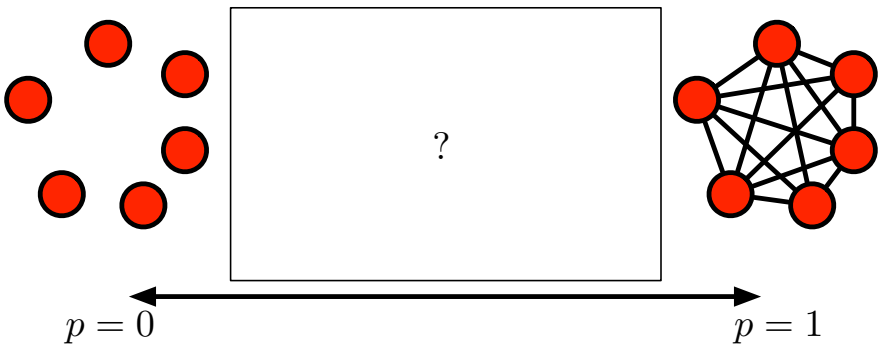


This property is another that differs sharply from real-world networks, particularly social networks, which we have seen exhibit fairly high clustering coefficients, even in very large graphs.

2.3 The giant component

This random graph model exhibits one extremely interesting property, which is the sudden appearance, as we vary the mean degree c, of a *giant component*, i.e., a component whose size is proportional to the size of the network n. This sudden appearance is called a *phase transition*.[6]

[6]The term "phase transition" comes from the study of critical phenomena in physics. Classic examples include the melting of ice, the evaporation of water, the magnetization of a metal, etc. Generally, a phase transition characterizes a sudden and qualitative shift in the bulk properties or global statistical behavior of a system. In this case, the transition is discontinuous and characterizes the transition between a mostly disconnected and a mostly connected networked.



Consider the two limiting cases for the parameter p. If $p = 0$ we have a fully empty network with n completely disconnected vertices. Every component in this network has the same size, and that size is a $O(1/n)$ fraction of the size of the network. In the jargon of physics, the size of the largest component here is an *intensive* property, meaning that it is independent of the size of the network.

On the other hand, if $p = 1$, then every edge exists and the network is an n-clique. This single component has a size that is a $O(1)$ fraction of the size of the network. In the jargon of physics, the size of the largest component here is an extensive property, meaning that it depends on the size of the network.[7] Thus, as we vary p, the size of the largest component transforms from an intensive property to an extensive one, and this is the hallmark of a phase transition. Of course, it could be that the size of the largest component becomes extensive only in the limit $p \to 1$, but in fact, something much more interesting happens. (When a graph is sparse, what other network measures are intensive? What measures are extensive?)

2.3.1 A phase transition

Let u denote the average fraction of vertices in $G(n, p)$ that do *not* belong to the giant component. Thus, if there is no giant component (e.g., $p = 0$), then $u = 1$, and if there is then $u < 1$. In other words, let u be the probability that a vertex chosen uniformly at random does not belong to the giant component.

For a vertex i not to belong the giant component, it must not be connected to any other vertex that belongs to the giant component. This means that for every other vertex j in the network, either (i) i is not connected to j by an edge or (ii) i is connected to j, but j does not belong to the giant component. Because edges are iid, the former happens with probability $1-p$, the latter with probability

[7] Other examples of extensive properties in physics include mass, volume and entropy. Other examples of *intensive* properties—those that are independent of the size of the system—include the density, temperature, melting point, and pressure.

pu, and the total probability that i does not belong to the giant component via vertex j is $1-p+pu$.

For i to be disconnected from the giant component, this must be true for all $n-1$ choices of j, and the total probability u that some i is not in the giant component is

$$\begin{aligned} u &= (1-p+pu)^{n-1} \\ &= \left[1-\frac{c}{n-1}(1-u)\right]^{n-1} \qquad (9) \\ &= \mathrm{e}^{-c(1-u)} \qquad (10) \end{aligned}$$

where we use the identity $p = c/(n-1)$ in the first step, and the identity $\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n = \mathrm{e}^{-x}$ in the second.[8]

If u is the probability that i is not in the giant component, then let $S = 1-u$ be the probability that i belongs to the giant component. Plugging this expression into Eq. (10) and eliminating u in favor of S yields a single equation for the size of the giant component, expressed as a fraction of the total network size, as a function of the mean degree c:

$$S = 1-\mathrm{e}^{-cS} \quad . \qquad (11)$$

Note that this equation is transcendental and there is no simple closed form that isolates S from the other variables.[9]

We can visualize the shape of this function by first plotting the function $y = 1-\mathrm{e}^{-cS}$ for $S \in [0,1]$ and asking where it intersects the line $y = S$. The location of the intersection is the solution to Eq. (11) and gives the size of the giant component. Figure 1 (next page) shows this exercise graphically (and Section 4 below contains the Matlab code that generates these figures). In the "sub-critical" regime $c < 1$, the curves only intersect at $S = 0$, implying that no giant component exists. In the "super-critical" regime $c > 1$, the lines always intersect at a second point $S > 0$, implying the existing of a giant component. The transition between these two "phases" happens at $c = 1$, which is called the "critical point".

[8] We can sidestep using the second identity by taking the logarithms of both sides of Eq. (9):

$$\ln u = (n-1)\ln\left[1-\frac{c}{n-1}(1-u)\right] \simeq -(n-1)\frac{c}{n-1}(1-u) = -c(1-u)$$

where the approximate equality becomes exact in the limit of large n. Exponentiating both sides of our approximation then yields Eq. (10). This should look familiar.

[9] For numerical calculations, it may be useful to express it as $S = 1+(1/c)W(-c\mathrm{e}^{-c})$ where $W(.)$ is the *Lambert W-function* and is defined as the solution to the equation $W(z)\mathrm{e}^{W(z)} = z$.

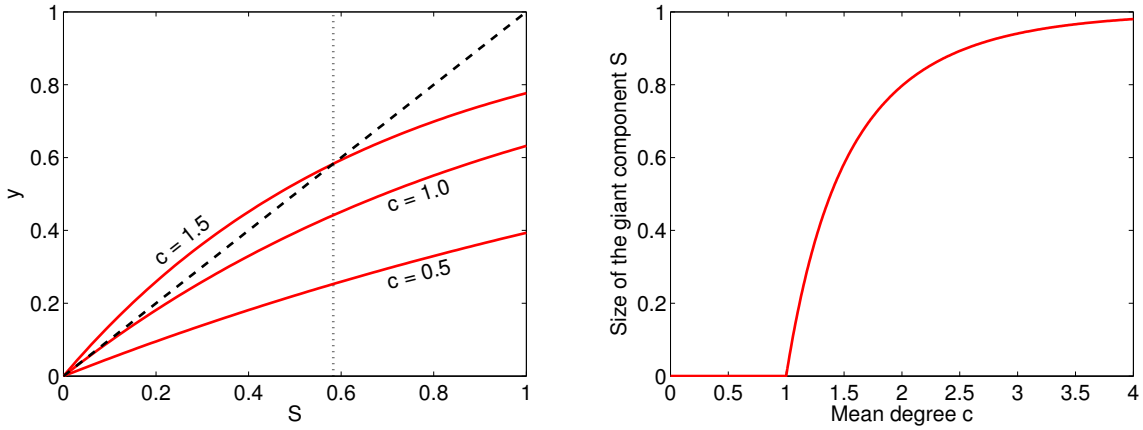


Figure 1: (a) Graphical solutions to Eq. (11), showing the curve $y = 1 - \mathrm{e}^{-cS}$ for three choices of c along with the curve $y = S$. The locations of their intersection gives the numerical solutions to Eq. (11). Any solution $S > 0$ implies a giant component. (b) The solution to Eq. (11) as a function of c, showing the discontinuous emergence of a giant component at the critical point $c = 1$.

2.3.2 Branching processes and percolation

An alternative analysis considers building each component, one vertex at a time, via a *branching process*. Here, the mean degree c plays the role of the expected number of additional vertices that are joined to a particular vertex i already in the component. The analysis can be made entirely analytical, but here is a simple sketch of the logic.

When $c < 1$, on average, this branching process will terminate after a finite number of steps, and the component will have a finite size. This is the "sub-critical" regime. In contrast, when $c > 1$, the average number of new vertices grows with each new vertex we add, and thus the branching process will never end. Of course, it must end at some point, and this point is when the component has grown to encompass the entire graph, i.e., it is a giant component. This is the "super-critical" regime. At the transition, when $c = 1$, the branching process could in principle go on forever, but instead, due to fluctuations in the number of actual new vertices found in the branching process, it does terminate. At $c = 1$, however, components of all sizes are found and their distribution can be shown to follow a power law.

2.4 A small world with $O(\log n)$ diameter

The branching-process argument for understanding the component structure in the sub- and super-critical regimes can also be used to argue that the diameter of a $G(n, p)$ graph should be small,

growing like $O(\log n)$ with the size of the graph n. Recall that the structure of the giant component is locally tree-like and that in the super-critical regime the average number of offspring in the branching process $c > 1$. Thus, the largest component is a little like a big tree, containing $O(n)$ nodes and thus, with high probability, has a depth $O(\log n)$, which will be the diameter of the network. This informal argument can be made mathematically rigorous, but we won't cover that here.

2.5 Drawing networks from $G(n, p)$

Generating instances of $G(n, p)$ is straight forward. There are at least two ways to do it: (i) loop over the upper triangle of the adjacency matrix, checking if a new uniform random deviate $r_{ij} < p$, which takes time $O(n^2)$; or (ii) generate a vector of length $n(n-1)/2$ of uniform random deviates, threshold them with respect to p, and then use a pair of nested loops to walk the length of the vector, which still takes time $O(n^2)$. A third way, which does not strictly generate an instance of $G(n, p)$, is to draw a degree sequence from the Poisson distribution to construct the network, which takes time $O(n + m \log m)$. In the sparse limit, the latter approach is essentially linear in the size of the network, and thus substantially faster for very large networks.

3 At home

1. Reread Chapter 12 (pages 397–425) in *Networks*
2. Next time: more random graphs

4 Matlab code

Matlab code for generating Figure 1a,b.

```
% Figure 1a
c = [0.5 1 1.5]; % three choices of mean degree
S = (0:0.01:1);  % a range of possible component sizes

figure(1);
plot(0.583.*[1 1],[0 1],'k:','LineWidth',2); hold on;
plot(S,1-exp(-c(1).*S),'r-','LineWidth',2); % c = 0.5 curve
plot(S,1-exp(-c(2).*S),'r-','LineWidth',2); % c = 1.0 curve
plot(S,1-exp(-c(3).*S),'r-','LineWidth',2); % c = 1.5 curve
plot(S,S,'k--','LineWidth',2); hold off     % y = S   curve
xlabel('S','FontSize',16);
ylabel('y','FontSize',16);
set(gca,'FontSize',16);
h1=text(0.7,0.26,'c = 0.5'); set(h1,'FontSize',16,'Rotation',14);
h1=text(0.7,0.47,'c = 1.0'); set(h1,'FontSize',16,'Rotation',18);
h1=text(0.2,0.32,'c = 1.5'); set(h1,'FontSize',16,'Rotation',38);

% Figure 1b
S  = (0:0.0001:1);  % a range of component sizes
c  = (0:0.01:4);    % a range of mean degree values
Ss = zeros(length(c),1);
for i=1:length(c)
    g     = find(S - (1-exp(-c(i).*S))>0, 1,'first'); % find the intersection point
    Ss(i) = S(g);                                     % store it
end;

figure(2);
plot(c,Ss,'r-','LineWidth',2);
xlabel('Mean degree c','FontSize',16);
ylabel('Size of the giant component S','FontSize',16);
set(gca,'FontSize',16,'XTick',(0:0.5:4));
```