Network Models

Vincenzo Bonifaci

April 3, 2017

1 Random graph models: $G(n, m)$ (Erdős-Rényi, variant 1)

Let Ω be the number of simple graphs on n nodes with m edges. If we associate a probability $\Pr(G) = 1/\Omega$ to each such graph, we obtain the $G(n, m)$ model.

The expected average degree of a graph in the $G(n, m)$ model, $c := \mathbb{E}[\frac{1}{n}\sum_{v \in V} \deg(v)]$ is exactly $2m/n$. However, the expected values of other quantities are quite difficult to compute.

2 Random graph models: $G(n, p)$ (Erdős-Rényi, variant 2)

Let $0 \leq p \leq 1$. If we associate a probability p to every pair of nodes among n nodes, and sample a graph where every edge is independently present with probability p, we obtain the $G(n, p)$ model. In this model, the probability associated to a *specific* graph with m edges is

$$p^m (1-p)^{\binom{n}{2} - m}.$$

The probability of getting *some* graph with exactly m edges is

$$P(m) := \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2} - m}.$$

2.1 Number of edges

Let X_{ij} be the indicator random variable representing whether i is linked to j or not. The expected "number of edges" between two specific nodes i and j is $\mathbb{E}[X_{ij}] = p \cdot 1 + (1-p) \cdot 0 = p$. Therefore the expected number of edges of a $G(n, p)$ random graph is $\mathbb{E}[\sum_{i,j:i<j} X_{ij}] = \sum_{i,j:i<j} \mathbb{E}[X_{ij}] = p \sum_{i,j:i<j} 1 = \binom{n}{2} p$. Note that this must also equal $\mathbb{E}[m] = \sum_{m=0}^{\binom{n}{2}} m P(m)$.

2.2 Average degree

What about expected average degree? Call the average degree c and recall that $c = 2m/n$. Therefore,

$$\mathbb{E}[c] = \mathbb{E}[2m/n] = (2/n) \sum_{m=0}^{\binom{n}{2}} m P(m) = \frac{2}{n} \binom{n}{2} p = (n-1)p.$$

2.3 Degree distribution

The probability that a specific node is linked to k specific others is $p^k(1-p)^{n-1-k}$. So the probability of being linked to exactly k any others is

$$p_k := \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

So $G(n, p)$ has a binomial degree distribution. In the limit of $n \to \infty$ (keeping $c = (n-1)p$ constant), we show that it approaches the Poisson distribution $p_k = \lambda^k e^{-\lambda}/k!$, with $\lambda = c$. Indeed,

$$\begin{aligned} \log((1-p)^{n-1-k}) &= (n-1-k)\log(1 - \frac{c}{n-1}) \\ &\simeq -(n-1-k)\frac{c}{n-1} \to -c. \end{aligned}$$

So $(1-p)^{n-1-k} \to e^{-c}$ as $n \to \infty$. Additionally,

$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!k!} \simeq \frac{(n-1)^k}{k!},$$

so

$$p_k \to \frac{(n-1)^k}{k!} p^k e^{-c} = \frac{(n-1)^k}{k!} \left(\frac{c}{n-1}\right)^k e^{-c} = e^{-c}\frac{c^k}{k!}.$$

Why should we care about this calculation? The point is that the Poisson distribution is highly concentrated, with a short tail: the tail probability decreases like $c^k/k!$. Instead, in most social networks we know from measurements that the tail probability of the degree distribution decreases much more slowly, like $1/k^\alpha$. Therefore, the Erdős-Rényi model is *not* a good model of social networks, at least from the point of view of the average degree.

2.4 Clustering coefficient

The clustering coefficient is another example of parameter that, for most social networks, is *not* predicted correctly by the random graph model $G(n, p)$.

Remember that the clustering coefficient is the probability that a random path of length 2 (u–v–z) is part of a triangle (u–v–z–u). In the $G(n, p)$ model, this probability is exactly p (why?). For fixed degree $c = p(n-1)$, as $n \to \infty$, this probability goes down like $c/(n-1) = O(1/n)$. In fact, even if the average degree is increasing with the size of the network (say, $c = O((\log n)^r)$, which seems a conservative assumption for most social networks observed in practice), the clustering coefficient of the $G(n, p)$ model tends to zero as $n \to \infty$. But in social networks observed in practice, the clustering coefficient has been found to be lower-bounded by a *non-negligible constant*, independently of the size of the network.

2.5 Giant components

When p gets larger, the probability that a connected component of size $\Theta(n)$ will form in the network (a *giant component*) also gets larger. We consider a heuristic argument to determine the ratio of the size of the giant component to n, as a function of p (in the limit of large n).

Let u be the expected fraction of nodes in the random graph that do not belong to the giant component. Note that u is also the probability that a randomly chosen node does not belong to a giant component.

The probability, for a specific node i, of not being connected to the giant component through another specific node j, is $1-p+pu$: with prob. $1-p$, i is not adjacent to j; with probability pu, i is adjacent to j, but j is not in the giant component.

The probability of not being connected to the giant component through *any* other node is therefore

$$u = (1-p+pu)^{n-1} = \left(1 - \frac{c}{n-1}(1-u)\right)^{n-1},$$

and taking logs,

$$\log u = (n-1)\log\left(1 - \frac{c}{n-1}(1-u)\right) \simeq -(n-1)\frac{c}{n-1}(1-u) = -c(1-u),$$

so u satisfies $u = e^{-c(1-u)}$, and $z := 1-u$ (the fraction of nodes in the giant component) satisfies

$$z = 1 - e^{-cz}.$$

This equation in z can be solved numerically, or graphically. There is only the solution $z = 0$ for small c, and there are two solutions for large c. The transition takes place when

$$\left.\frac{d}{dz}(1-e^{-cz})\right|_{z=0} = 1,$$

giving $ce^{-c\cdot 0} = 1$, i.e., $c = 1$. So we do not expect a giant component unless $c > 1$.

Do we really expect a giant component when $c > 1$? Consider a large enough set T of s nodes. If we expand this set by adding its immediate neighbors not in it, we expect to find $p(n-s)$ edges out of each node in T to nodes out of T. But

$$p(n-s) = c\frac{n-s}{n-1} \simeq c,$$

so when $c > 1$ the size of the periphery gets larger by the constant factor c.

Can there be *two* giant components? It can again be argued that for large n this happens only with vanishing probability. Indeed, if there are two giant components with fractions z_1 and z_2 of the total number of nodes, there are $z_1 n \cdot z_2 n$ ways to connect them, and none of this potential edges should be in the graph, which happens with probability

$$q := (1-p)^{z_1 z_2 n^2} = \left(1 - \frac{c}{n-1}\right)^{z_1 z_2 n^2}.$$

Taking logs and using the second order Taylor approximation $\ln(1-x) \simeq -x - x^2/2$ yields

$$\log q = z_1 z_2 \left(n^2 \log(1 - \frac{c}{n-1}) \right) \simeq z_1 z_2(-c(n+1) - c^2/2) = c z_1 z_2(-n - (c/2+1)).$$

But for large n, $\log q \to -\infty$, so $q \to 0$.