Appendix 1

Random graphs

A network is said to be random when the probability that an edge exists between two nodes is completely independent from the nodes' attributes. In other words, the only relevant function is the degree distribution $P(k)$. In the case that we are mainly interested in the node degree, by using the correlation function language used in Chapter 1 this implies that all degree correlation functions must be trivial. Even uncorrelated networks, however, must satisfy some basic constraints. The first one is the normalization relation

$$\sum_k P(k) = 1. \tag{A1.1}$$

A second constraint is imposed by the fact that all edges must point from one vertex to another, so that no edges with dangling ends exist in the network. This amounts to a degree detailed balance condition imposing that the total number of edges pointing from vertices of degree k to vertices of degree k' must be equal to the number of edges that point from vertices of degree k' to vertices of degree k. In order to state this condition mathematically let us denote by N_k the number of vertices with degree k. Since $\sum_k N_k = N$, we can define the degree distribution as

$$P(k) \equiv \lim_{N\to\infty} \frac{N_k}{N}. \tag{A1.2}$$

To define the network completely, apart from the relative number of vertices of a given degree, we need to specify how the vertices are connected through the symmetric matrix $E_{kk'}$ of the total number of edges between vertices of degree k and vertices of degree k' for $k \neq k'$. The diagonal values E_{kk} are equal to twice the number of connections between vertices in the same degree class, $k = k'$, yielding the following identities:

$$\sum_{k'} E_{kk'} = kN_k, \tag{A1.3}$$

$$\sum_{k,k'} E_{kk'} = \langle k\rangle N = 2E, \tag{A1.4}$$

where $\langle k\rangle$ is the average degree and E is the total number of edges in the network. In fact, the first relation states that the number of edges emanating from all the vertices of degree k is simply kN_k, while the second indicates that the sum of all the vertices' degrees is equal to twice the number of edges. The first identity allows us to write the conditional

https://doi.org/10.1017/CBO9780511791383.015 Published online by Cambridge University Press

probability $P(k' \mid k)$, defined as the probability that an edge from a k vertex points to a k' vertex, as

$$P(k' \mid k) = \frac{E_{kk'}}{kN_k}. \tag{A1.5}$$

The second identity allows the definition of the joint degree distribution as

$$P(k, k') = \frac{E_{kk'}}{\langle k \rangle N}, \tag{A1.6}$$

with the symmetric function $(2 - \delta_{k,k'})P(k, k')$ being the probability that a randomly chosen edge connects two vertices of degrees k and k'. This yields

$$P(k' \mid k) = \frac{E_{kk'}}{kN_k} \equiv \frac{\langle k \rangle P(k, k')}{kP(k)}, \tag{A1.7}$$

from where the detailed balance condition

$$kP(k' \mid k)P(k) = k'P(k \mid k')P(k'), \tag{A1.8}$$

follows immediately as a consequence of the symmetry of $P(k, k')$.

For uncorrelated networks, in which $P(k' \mid k)$ does not depend on k, application of the normalization condition $\sum_k P(k \mid k') = 1$ into Equation (A1.8) yields the form

$$P(k' \mid k) = \frac{1}{\langle k \rangle} k'P(k'), \tag{A1.9}$$

that has been used throughout this book. This recovers that for generalized random networks the conditional probability is just proportional to the relative proportion of edges departing from nodes k' and it is the same for all edges whatever the degree k of the emitting node. Equation (A1.9) allows the explicit calculation of some basic quantities of random networks with arbitrary degree distribution $P(k)$.

Following Newman (2003b), let us consider a generalized random graph with arbitrary degree distribution $P(k)$. Since edges are assigned at random between pairs of vertices, Equation (A1.9) implies that the probability that any edge points to a vertex of degree k is given by $kP(k)/\langle k \rangle$. Consider now a vertex i; following the edges emanating from it, we can arrive at k_i other vertices. The probability distribution that any of the neighboring vertices has k edges pointing to other vertices different from i (plus the edge from which we arrived) is therefore given by the function

$$q(k) = \frac{(k+1)P(k+1)}{\langle k \rangle}. \tag{A1.10}$$

In other words, $q(k)$ gives the probability distribution of the *number of second nearest neighbors* that can be reached following a given edge in a vertex. The average number of these second nearest neighbors is then given by

$$\sum_k kq(k) = \frac{1}{\langle k \rangle} \sum_k k(k+1)P(k+1) = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}. \tag{A1.11}$$

The absence of correlations also yields that any vertex j, neighbor of the vertex i, is connected to another vertex l, which is at the same time a neighbor of i, with probability

https://doi.org/10.1017/CBO9780511791383.015 Published online by Cambridge University Press

$k_j k_l/\langle k\rangle N$. Thus, the clustering coefficient is simply defined as the average of this quantity over the distribution of all possible neighbors of i, i.e.

$$\langle C\rangle = \frac{1}{N\langle k\rangle}\sum_{k_j}\sum_{k_l} k_j k_l q(k_j) q(k_l) = \frac{1}{N}\frac{(\langle k^2\rangle - \langle k\rangle)^2}{\langle k\rangle^3}. \tag{A1.12}$$

In order to provide an approximate expression for the scaling of the diameter of random networks[1] we can compute iteratively the average number of neighbors z_n at a distance n away from a given vertex as

$$z_n = \frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} z_{n-1}. \tag{A1.13}$$

Finally, by considering that $z_1 = \langle k\rangle$, it is possible to obtain the explicit expression

$$z_n = \left(\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right)^{n-1} \langle k\rangle. \tag{A1.14}$$

If the average shortest path length is $\langle \ell\rangle$, then the number of neighbors at this distance must be approximately equal to the size of the graph N, thus obtaining $z_{\langle \ell\rangle} = N$. Using the same argument as in the Erdős–Rényi model, we readily obtain

$$\langle \ell\rangle \approx 1 + \frac{\log[N/\langle k\rangle]}{\log[(\langle k^2\rangle - \langle k\rangle)/\langle k\rangle]}. \tag{A1.15}$$

The small-world properties are thus evident also for generalized random graphs and it is easy to check that for the case of a Poisson distribution, with second moment $\langle k^2\rangle = \langle k\rangle + \langle k\rangle^2$, one recovers the results for the Erdős–Rényi model. However, we must keep in mind that the previous expression is a rather crude approximation, which might fail for more complex degree distributions (especially in the presence of heavy tails).

The absence of degree correlations allows also the calculation of the dependence of the degree cut-off $k_c(N)$ induced by the finite size N of the network in scale-free networks.[2] The presence of the degree cut-off translates in the degree distribution into an explicit dependence on the network size (or time) that we can write in the scaling form

$$P(k, N) = k^{-\gamma} f\left[\frac{k}{k_c(N)}\right], \tag{A1.16}$$

where $f(x)$ is constant for $x \ll 1$ and decreases very quickly for $x \gg 1$. It is possible to obtain an upper bound for the functional dependence on N of the degree cut-off for generalized uncorrelated random graphs using an extremal theory argument. In the continuous k approximation, consider a random graph with normalized degree distribution in the infinite size limit $P(k) = (\gamma - 1)m^{\gamma-1}k^{-\gamma}$, with $\gamma > 2$ and $k \in [m, \infty]$, where m is the minimum degree of the graph. Consider now that we generate a graph of size N by sorting N independent random variables according to the distribution $P(k)$, obtaining

[1] This calculation is analogous to that used for the Erdős–Rényi model in Section 3.1.

[2] The result obtained here is valid only if the only origin of the degree cut-off resides in the finite number of vertices forming the network. In other situations, networks may exhibit a degree cut-off due to external constraints and finite connectivity resources (Amaral *et al.*, 2000; Mossa *et al.*, 2002). In this case k_c is not related to the network size and has to be considered as an external parameter.

https://doi.org/10.1017/CBO9780511791383.015 Published online by Cambridge University Press

the sample $\{k_1, \ldots, k_N\}$. Let us define K the maximum value of this particular sample, $K = \max\{k_1, \ldots, k_N\}$. When generating an ensemble of graphs, we will obtain in each case a different value of the maximum degree K. Thus, we can define the cut-off $k_c(N)$ as the average value of K, weighted by the distribution $P(k)$. It is easy to see that the probability of this maximum being less than or equal to K is equal to the probability of all the individual values k_i being in their turn less than or equal to K. This means that the distribution function of the maximum value K is just

$$\Pi(K, N) = \left[\Psi(K)\right]^N, \tag{A1.17}$$

where $\Psi(K)$ is the distribution function of the probability $P(k)$, i.e. $\Psi(K) = \int_m^K P(k)\ \mathrm{d}k = 1 - (K/m)^{-\gamma+1}$. By differentiating Equation (A1.17), we obtain the probability distribution of maximum values, namely

$$\pi(K, N) = \frac{\mathrm{d}\Pi(K, N)}{\mathrm{d}K} = \frac{N(\gamma - 1)}{m}\left(\frac{K}{m}\right)^{-\gamma}\left[1 - \left(\frac{K}{m}\right)^{-\gamma+1}\right]^{N-1}. \tag{A1.18}$$

If the degree cut-off is defined as the average value of the maximum K, then we have that

$$k_c(N) = \int_m^\infty K\pi(K, N)\ \mathrm{d}K = \frac{Nm}{\lambda}\frac{\Gamma(1+\lambda)\Gamma(N)}{\Gamma(N+\lambda)}, \tag{A1.19}$$

where $\Gamma(x)$ is the Gamma function and we have defined the constant $\lambda = (\gamma - 2)/(\gamma - 1)$. Using the asymptotic relation $\lim_{N\to\infty} \Gamma(N + a)/\Gamma(N + b) \simeq N^{a-b}$ (Abramowitz and Stegun, 1972), we obtain the leading behavior for large N

$$k_c(N) \simeq m\frac{\Gamma(1+\lambda)}{\lambda}N^{1-\lambda} \sim mN^{1/(\gamma-1)}. \tag{A1.20}$$

The previous equation is in fact an upper bound for $k_c(N)$, since we have only considered the possible values that the random variables k_i can take, according to the probability distribution $P(k)$. If those values must represent the degree sequence of an actual graph, some constraints would then apply, especially if we want to avoid the presence of loops or multiple edges.

The previous scaling relation with the size of the graph allows, finally, an important consideration on the scaling of $\langle C\rangle$ and $\langle \ell\rangle$. In generalized random graphs these quantities depend essentially on the first and second moments of $P(k)$. For a bounded degree distribution, in which the degree fluctuations $\langle k^2\rangle$ are finite, we observe that $\langle C\rangle \sim 1/N$ and $\langle \ell\rangle \sim \log N$, in agreement with the results found for the Erdős–Rényi model. On the other hand, for degree distributions with a fat tail, such as a power law, the second moment $\langle k^2\rangle$ can be very large, and even diverge with N. In this case, the prefactor in Equation (A1.12) can be noticeably large, and yield a clustering coefficient that might be higher than the one corresponding to an Erdős–Rényi graph with the same size and average degree. Indeed, if we consider a scale-free graph with degree distribution $P(k) \sim k^{-\gamma}$, we have that $\langle k\rangle$ is finite, while the degree fluctuations scale as $\langle k^2\rangle \simeq k_c^{3-\gamma}$, where k_c is the maximum degree present in the graph. By plugging the behavior of k_c given by Equation (A1.20) into Equation (A1.12), we obtain an average clustering coefficient depending on the network size as

$$\langle C\rangle_N \simeq N^{(7-3\gamma)/(\gamma-1)}. \tag{A1.21}$$

https://doi.org/10.1017/CBO9780511791383.015 Published online by Cambridge University Press

Since the clustering coefficient cannot be larger than 1, Equation (A1.12) must be restricted to values of the degree exponent $\gamma > 7/3$. The last consideration points out that correlations must arise naturally to allow the connectivity pattern observed in some real-world networks. The random approximation is, however, a basic model often used in calculations as it generally allows an explicit analytic expression to be obtained.