9

Networks

9.1 Convergent Computation

9.2 Connectedness

9.2.1 *What is the minimum number of links needed to connect 100 nodes in one large component (a collection of nodes that are directly or indirectly connected to each other)?*

Answer Placing all nodes along a line, one needs $N-1 = 99$ links to connect the graph. Placing $N-1$ nodes connected by a single link to a central node similarly needs $N-1 = 99$ links. In fact, any architecture without a loop needs 99 links to be connected, as the network can be grown subsequently by adding one node with one link to any previous node. Starting with one non-connected node, one thus needs at least 99 links to connect any the network.

9.2.2 *What is the largest diameter one can have in a network with 100 nodes?*

Answer Placing all nodes along one line makes a diameter $D = N-1 = 99$.

9.2.3 *How does the diameter of a network scale with the number of nodes* N*, when these are organized on a square/cubic... grid in d dimensions. Consider, for example, 4096 nodes, organized in one-, two-, and three-dimensional lattices. Determine the number diameter of the 4096 node network when organized in an Erdos–Reynei network with an average connectivity of six (the same number of neighbors as a three- dimensional cubic lattice). Convince yourself that an Erdos–Reynei network has infinite dimensions.*

© K. Sneppen

Answer On a line, the number of nodes $n(L)$ within distance L is 2L: $n(L) \propto L$ and the dimension is 1. On a square lattice the number of nodes within a distance L is $\pi \cdot L^2$: $n(L) \propto L^2$ and the dimension is 2. In d dimensions the number of nodes within distance L is $n(L) \propto L^d$.

$N = 4096$ on a one-dimensional line is thus reached at a typical distance $L \sim N/2 \sim 2000$.

$N = 4096$ on a two-dimensional lattice is thus reached at a typical distance $L \sim \sqrt{N/\pi} \sim 35$.

$N = 4096$ on a three-dimensional lattice is reached within a distance $L \sim (3N/(4\pi))^{1/3} \sim 10$. For a Erdos-Reynei network we disregard loops and estimate the number of loops at distance L to be dominated by the contributions at an outer distance: $n(L) \sim \mathrm{d}n/\mathrm{d}L = k^L \sim n(L)$ and thus $n(L)k^L$. For $n(L) \sim 4000$ and $k = 6$: $6^L \sim 4000 \Rightarrow L = \ln(4000)/\ln(6) \sim 4.6$.

Erdos–Reynei networks have infinite dimensions because k^L grows faster than any power law L^d.

9.2.4 *Generate Erdos–Reynei networks with* $p = 3/(N-1)$ *(three neighbours per node on average, and not allowing self-interactions) for* $N = 10$, $N = 100$ *and* $N = 1000$*, and count how many triangles there are at various network sizes. How does the number of triangles change with* N *for a fixed average connectivity (fixed connectivity implies that* p *decreases with system size).*

Answer A random symmetric sparse matrix M is generated by assigning 0 to all entries, except a few, representing links. For each i, j, $i > j$ is assigned value 1 with probability p. Subsequently $M(j,i) = M(i,j)$, $(M(i,i) = 0)$. The results are shown in the left-hand panel of Fig. 9.1. Triangles are counted by calculating the matrix M^3, $n(\text{triangles}) = \text{trace}(M^3)/6$. The matrix shown has five triangles. A similarly generated matrix of size $N = 1000$ also has five triangles in total.

9.2.5 *Consider a non-directed network that only consists of one large component. Prove that a random walker, after an infinitely long time will visit each node with a probability that is proportional to its degree. Notice that random walks in networks are at the core of search engines such as Google, although the walkers also complete other moves to deal with the properties of directed networks. Hint: consider the steady-state flux between two connected nodes with different degrees.*

Answer The random walker can visit all nodes by chance. The average number of transitions from node i to a directly connected node j must after

© K. Sneppen

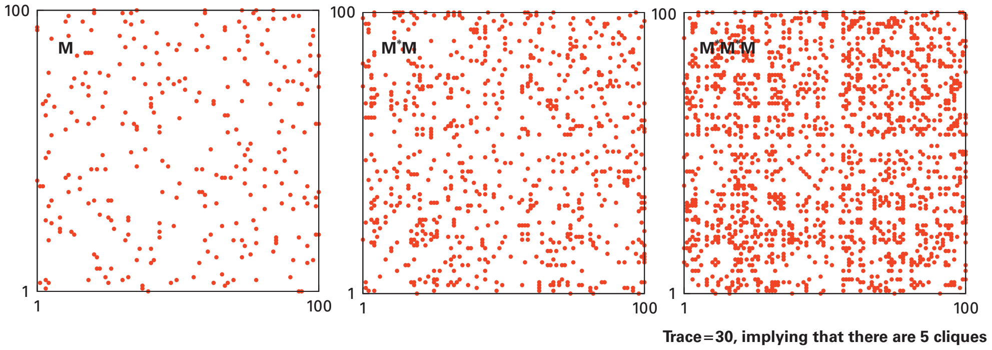

Figure 9.1 A $N = 100$ network with an average connectivity three. The matrix for the network is in the left-hand panel, the two subsequent panels show nodes connected by two links, and nodes three links, respectively.

a very long time be equal the average number of transitions from node j to node i. If this was not the case, there would be a net flux across the link. that is incompatible with random motion. As the net flux out of a node i through one of its k_i links must equal $1/k_i$ (all outgoing links chosen with equal probability), the steady-state condition gives:

$$\frac{\rho_i}{k_i} = \frac{\rho_j}{k_j} \Rightarrow \rho_i \propto k_i$$

where ρ_i is the probability that the walker is at node i at a given time. Thus, highly connected nodes are visited more often.

9.2.6 *Generate a random network of size* $N = 200$ *with 150 links (average degree* $\langle k \rangle = 3$*) and monitor the size of the largest component as nodes are subsequently removed. Do the same when removing links subsequently, but maintaining all nodes. Hint: the central part of this question is to make an algorithm that detects all parts of a network that is connected indirectly to a given node.*

Answer The central part in this question is to make an algorithm that detects all parts of a network that are connected to a given node i. This is done iteratively, making a growing list of nodes starting from i. At each step of the algorithm the list grows with all neighbors to all nodes on the list, under the condition that these neighbors are not already on the list. The growing list terminates when it does not contain any nodes whose neighbors have not been added to the list.

© K. Sneppen

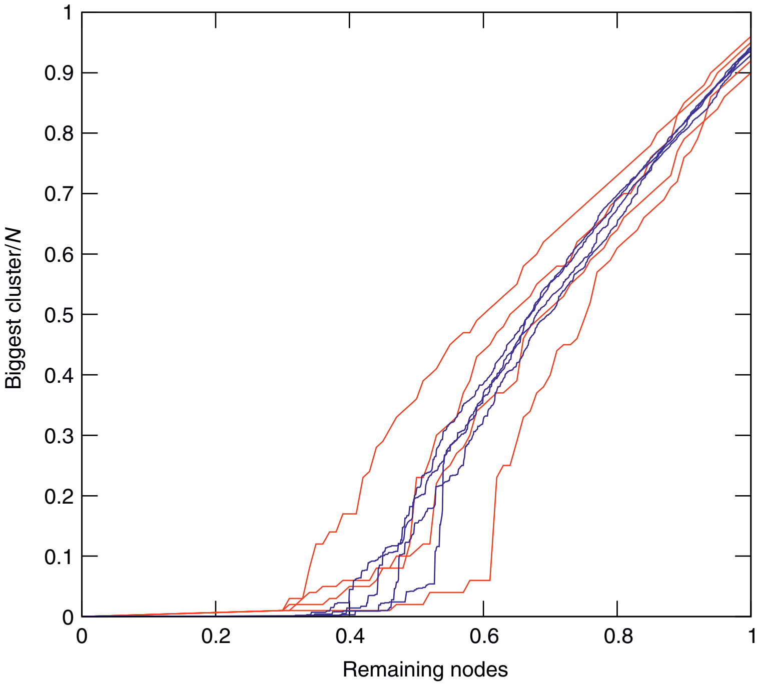

Figure 9.2 Pruning of an N=100, and an N=1000 Erdos-Reynei network, both with average connectivity three. The remaining nodes are counted in units of the size of the network before pruning. For each network size, we have sampled four different realizations.

To find the largest cluster, one needs to start at any remaining node in the network.

Alternatively, if the reader uses matlab, one may use the exponential of the adjacency matrix, and simply measure the largest component as the maximum number of non-zero elements in a row in the matrix.

To measure how the largest cluster decreases by removing nodes, one has to repeat the investigation from start, each time removing one more node in the network.

The results for an $N = 100$ and an $N = 1000$ network are shown in Fig. 9.2. One can observe a sharp decline until about half the nodes are removed. At this point the network is effectively disconnected.

9.2.7 *Use the following equation for the fraction F of the nodes that remain in the largest cluster of an ER network, after a fraction f is removed [477]:*

$$F = (1 - f) \cdot (1 - e^{-\langle k \rangle \cdot F}) \tag{9.1}$$

to determine the critical value of f *where* $F = 0$ *and examine* F *as* f *approaches this critical point. Hint: expand the exponential in the above equation using* $\langle k \rangle F << 1$ *close to the critical point.*

Answer Expand $\exp(-\langle k \rangle \cdot F) = 1 - \langle k \rangle \cdot F$ to obtain

$$\begin{aligned} F &= (1 - f) \cdot (\langle k \rangle \cdot F) \\ \Rightarrow 1 &= (1 - f) \cdot \langle k \rangle \\ \Rightarrow f &= 1 - \frac{1}{\langle k \rangle} \end{aligned}$$

The behavior of F close to the critical point will be given by a second-order expansion $\exp(-\langle k \rangle \cdot F) = 1 - \langle k \rangle \cdot F + (\langle k \rangle \cdot F)^2/2$:

$$\begin{aligned} F &= (1 - f) \cdot (\langle k \rangle \cdot F - (\langle k \rangle \cdot F)^2/2) \\ \Rightarrow 1 &= (1 - f) \cdot \langle k \rangle (1 - \langle k \rangle \cdot F/2) \\ \Rightarrow F &= \frac{2}{\langle k \rangle^2} \cdot (\langle k \rangle - \frac{1}{(1 - f)}) \end{aligned}$$

which for $f = f_c + \delta$ with $f_c = 1 - 1/\langle k \rangle$ changes as:

$$F = \frac{2}{\langle k \rangle} \cdot (1 - \frac{1}{(1 + \delta \langle k \rangle)}) \sim 2 \cdot \delta \tag{9.2}$$

Accordingly, F increases linearly away from the critical point.

9.3 Large Scale Molecular Networks

9.4 Analysis of network topologies

9.4.1 *Construct a network* $N = 100$ *nodes subdivided into 10 different classes with 10 nodes in each. Generate a random network where each node has about three links, and nodes within the same class have 10 times larger probability of being connected than nodes of different classes. Calculate the number of loops, and compare this with the number of loops when all the links are randomized.*

Answer Generate the networks in the form of sparse matrices, where the link density within blocks of 10 is 10 times larger than the link density p, outside. For a homogeneous network use density $q = 3/(N-1) = 0.03$, giving each node $q \cdot (N-1) = 3$ links. For a modular network with parameter p, the number of estimated links $10 \cdot p \cdot (10-1) + p \cdot 90$, p is given by:

$$(10 \cdot p \cdot (10-1) + p \cdot (N-10)) \cdot N = q \cdot N \Rightarrow$$
$$p \cdot (80 + N) = q \cdot N \Rightarrow p = \frac{q}{1 + 80/N} = 0.3/1.8 = 0.0166$$

Triangles counted for 1/6 of the trace of the network matrix. We sample over 25 networks, both with an average connectivity of three:
For a modular network size 100, the number of triangles $n(\text{tri}) = 6.8$.
For a homogeneous network size 100, the number of triangles $n(\text{tri}) = 4.4$.
For a modular network size 200, the number of triangles $n(\text{tri}) = 6.3$.
For a homogeneous network size 200, the number of triangles $n(\text{tri}) = 4.6$.

9.4.2 *Repeat the above procedure for a network matrix where nodes that are within ±10 of the diagonal have 10 times larger probability of being connected than nodes that are further away from each other.*

Answer Repeating the procedure from the last question, with the constraint being associated with nodes i and j where $\text{abs}(i-j) < 11$.

Again, limiting our investigation to the connectivity 3 case:
For the constrained network of size 100, use $p = 0.0112$ and obtain $n(\text{tria}) = 6.5$.
For a homogeneous network of size 100, one obtain $n(\text{tria}) = 4.4$.
For the constrained network of size 200, use $p = 0.0080$ and obtain $n(\text{tria}) = 6.6$.
For a homogeneous network size 200, $n(\text{tria}) = 4.6$.

9.5 Models for scale free molecular networks

9.5.1 *Simulate a threshold network for a $N = 1000$ system. With G_i assigned as in the text, good results should be obtained with a threshold of 7.*

Answer Assign a binding affinity to each node $G_i = -\ln(\text{random})$ where random is a random number selected uniformly from the interval $[0,1]$. A network matrix is generated by assigning links $M(i,j) = 1$ for all pairs i, j where $G_i + G_j > \theta$ and $\theta = 7$. Results are shown in Fig. 9.3. Noticeably, for the same θ there is substantial sample-to-sample variation, with $\langle k \rangle$ taking values between 2.3 and 6.7. When using the threshold $\theta = 8$, the average degree is only 0.5.

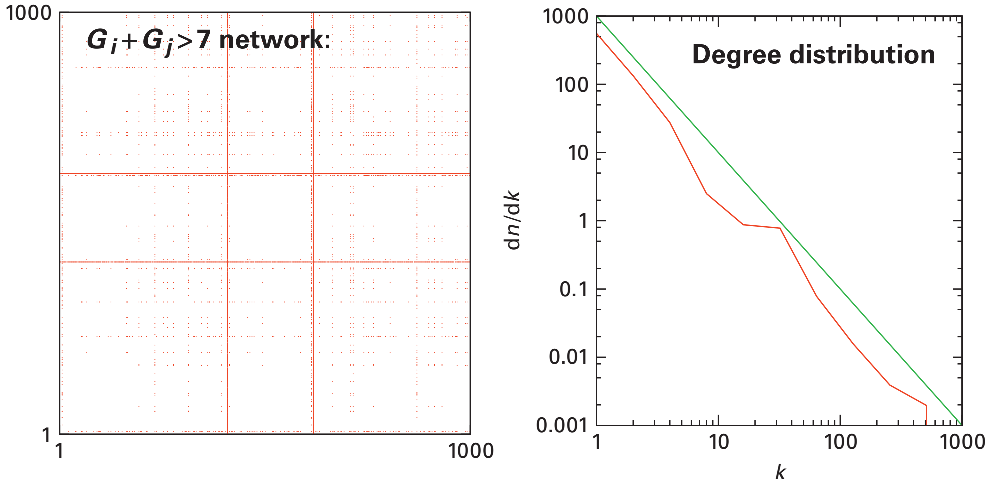

Figure 9.3 Threshold network visualized in terms of a sparse matrix, and the associated degree distribution. The threshold used was $\theta = 7$ and the average degree is 5.3.

9.5.2 *Simulate a threshold network for a* $N = 1000$ *system where* G_i *is selected from Gaussian with mean 0 and standard deviation 1. The threshold* $\theta = 3.5$ *should give reasonable results.*

Answer A Gausian number with mean 0 and standard deviation 1 is selected by

$$G_i = -6 + \sum_{i=1}^{12} \mathrm{ran}_i \tag{9.3}$$

where each of the 12 random numbers ran_i is selected uniformly between 0 and 1. A link between i and j is assigned when $G_i + G_j > \theta$.

Threshold $\theta = 4$ gives $\langle k \rangle = 2.7$.
Threshold $\theta = 3.5$ gives $\langle k \rangle = 7.5$.
Threshold $\theta = 3$ gives $\langle k \rangle = 18$.

The results are shown in Fig. 9.4. The degree distribution remains close to $1/k^2$, and thus the scale-free degree distribution is insensitive to details in the distribution of G_i.

9.5.3 *Derive Eq. (9.27) with pre-factors, as well as the expected number of metabolites in each added pathway using random stepping between available*

© K. Sneppen

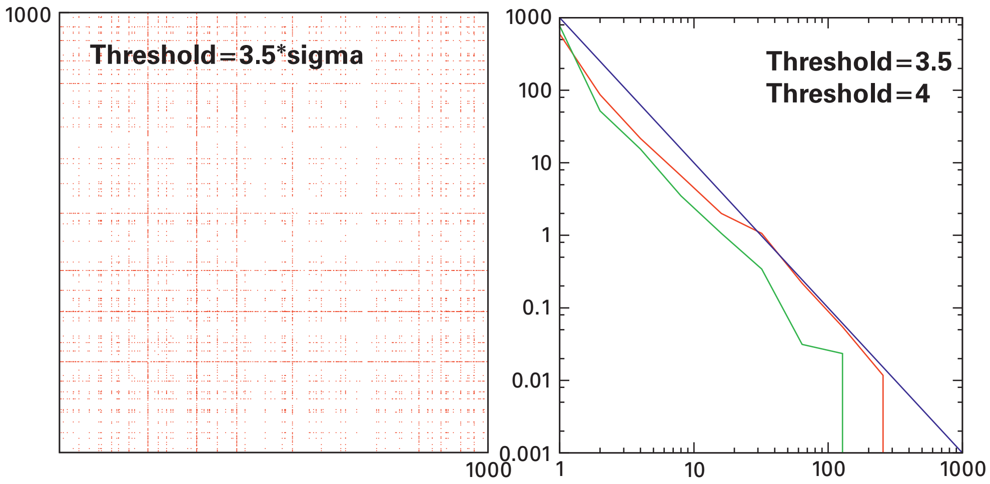

Figure 9.4 Threshold network visualized in terms of a sparse matrix, and the associated degree distribution. The G_i are selected from a Gaussian distribution with threshold for a link set at 3.5 standard deviations. On the right we also investigate the effect of increasing the threshold slightly.

reactions/enzymes. Reiterate the derivation of the scale-free distribution of the added number of enzymes for each new metabolite in the growth model. Eventually consult ref. [497].

Answer The model assumes that there is a total set of all possible enzymes in the world: N_{univ}. The primordial metabolism is assumed to consist of a core N_{core} consisting of a very small subsection of possible enzymes. Consider a growing metabolism. Metabolizing a new metabolite implies that the organism needs to connect the new enzyme to the already existing set of enzymes. In our simple model we assume this is done by adding a random sequence of enzymes drawn from the global pool. This sequence represents the best guess on a metabolic path. The sequence is terminated when an enzyme from the organism itself is chosen. The likelihood that this happens is:

$$p_{\text{end}} = \frac{N_{\text{enzymes}}}{N_{\text{univ}}} \tag{9.4}$$

The number of new enzymes that is added for each new metabolite is then:

$$\frac{N_{\text{univ}}}{N_{\text{enzymes}}} \tag{9.5}$$

© K. Sneppen

and, assuming that each new metabolite is at the end of a path, and thus needs one transcriptional regulator, we obtain:

$$\frac{\mathrm{d}N_{\mathrm{enzymes}}}{\mathrm{d}N_{\mathrm{R}}} = \frac{N_{\mathrm{univ}}}{N_{\mathrm{enzymes}}} \Rightarrow N_{\mathrm{R}} = \frac{N^2_{\mathrm{enzymes}}}{2N_{\mathrm{univ}}} \tag{9.6}$$

Answer to second part of question (same as main text, but with slightly more detail) The degree of the regulator added to the metabolite in previous question would be:

$$K_{\mathrm{out}} = N_{\mathrm{univ}}/N_{\mathrm{enzyme}} \tag{9.7}$$

As metabolism grows, still smaller metabolic branches are added. Pathways with length:

$$K_{\mathrm{out}} > K \tag{9.8}$$

are added when number of enzymes (= number of metabolites that can be utilized) is smaller than:

$$N_{\mathrm{enzymes}} < N_{\mathrm{univ}}/K \tag{9.9}$$

At that point the number of regulators is:

$$N_{\mathrm{R}} < \frac{N^2_{\mathrm{enzymes}}}{2N_{\mathrm{univ}}} = \frac{N_{\mathrm{univ}}}{2 \cdot K^2} \tag{9.10}$$

All these regulators have a high degree, in fact all exceed K:

$$P(K_{\mathrm{out}} > K) = \frac{N_{\mathrm{univ}}}{2 \cdot K^2} \Rightarrow \frac{\mathrm{d}P}{\mathrm{d}K} =\propto 1/K^3 \tag{9.11}$$

© K. Sneppen