4 Emergence

While it is tempting to focus on individual molecules, such as hemoglobin, or combinations of a few molecules, as in the case of motor–microtubule–ATP, or transcription factor–DNA, it is clear that the complexity of an organism or a cell cannot be fully captured from such a perspective. If limited to one role per protein, the roughly 30,000 Human genes would have limited utility. The key to diversity of behavior is *(i)* the combinatorial power from many genes acting in concert; *(ii)* the time profile of expressing and suppressing genes, *(iii)* localization/compartmentalization of proteins in different locations, and *(iv)* interactions with the resources and stimuli from the environment. Various forms of behavior can then *emerge* from a palette of few elements.

Faced with the complexity of *emergent* phenomena, one approach is to focus on the elements, say picking out individual molecules and studying their properties, and possibly interactions with a few other molecules, in great detail. While undoubtedly important, this is akin to study of human history through biographies of important individuals– certainly illuminating, but with the risk of losing sight of the forest for its trees. It certainly increasingly difficult to obtain the same level of complete understanding upon increasing the number of interacting components. Even for simple point particles in classical mechanics, there is no general (deterministic) solution for three bodies; even simple orbits becoming chaotic and unpredictable at long times. However, this is no reason to despair as statistical mechanics teaches us that such complexity can actually be helpful, provided that the deterministic perspective is abandoned for a probabilistic one. Thus while we cannot follow the trajectories of particles in a gas, we can quite accurately account for emergent properties, such as the pressure and temperature of the gas, and how their spatio-temporal variations lead to passage of sound waves. Indeed one of the lessons of statistical physics is that many macroscopic properties of the gas are independent of its constituent molecules. It is tempting to apply similar reasoning to emergent phenomena in life sciences, and there are such attempts, such as cellular automata as representations of evolution and life. One has to be careful with such approaches as the constituents are far more complex than point particles, and their number is (relatively) much smaller. For example, one can legitimately wonder how the course of history would be modified if Newton and Watson changed place. What the productive medium is between including more details of individual constituents, and focusing on collective behavior of many simplified ingredients, is not clear-cut. We shall focus on examples of the latter, starting with descriptions of networks.

4.1 Networks

The primary elements of a network are its *nodes*. These can be a set of genes or proteins in the cell, the interconnected neurons of the brain, or organisms in an ecosystems. *Links* between nodes indicate a direct interaction, for example between proteins that bind, neurons connected by synapses, or organisms in a predator/prey relationship. In its most basic form, the network can be represented by nodes $i = 1, 2, \cdots, N$ as points of a graph, and links L_{ij} as edges between pairs of points. Excluding self-connections, the maximal number of

possible links is $N(N-1)$ with directed connections (e.g. as in a predator/prey relation), and $N(N-1)/2$ for undirected links (as in binding proteins). A *subgraph* is a portion of the total network, say with n nodes and l links. Some types of subgraphs have specific names; e.g. a *cycle* is a path starting and ending at the same node, while a *tree* is a branching structure without cycles.

4.2 The random graph

Analyzing biological data from the perspective of networks has gained interest recently. Much is known about the interplay of proteins that control expression of genes, the connections of the few hundred neurons in the roundworm *C. elegans,* and other example. One possible route to extracting information from such data is to look for specific *motifs*, subgroups of several nodes, that can cooperate in simple functions (e.g. a feedforward loop). A particular motif can be significant if it appears more (or less) frequently than expected. We thus need a simple model whose expectations can be compared with biological data. *Random graphs*, introduced by Erdös and Rényi, serve this purpose. The model consists of N nodes, with any pair connected at random and independently, with probability p.

We shall explore several features of Erdös-Rényi networks in the following sections. For the time being, we note that the expected number of subgraphs of n nodes and l links,

$$\mathcal{N}(n,l) = \left(\begin{array}{c} N \\ n \end{array} \right) p^l \times \frac{n!}{(\text{symmetry factors})}\,, \tag{4.1}$$

is obtained as a product of the number of ways of picking n points and connecting them with l links, and a factor that accounts for the number of ways of connecting the points into the desired graph. For example, there are $n!/2$ ways to string n points along a straight line with $l = (n-1)$, and the expected number of such linear pathways is

$$\mathcal{N}(n \text{ in a line}) = \frac{N!}{(N-n)!}\frac{p^{n-1}}{2}\,, \tag{4.2}$$

while there are $n!/(2n)$ ways to make a cycle of n nodes and $l = n$ links, such that

$$\mathcal{N}(n \text{ in a cycle}) = \frac{N!}{(N-n)!}\frac{p^n}{2n}\,. \tag{4.3}$$

There is also a single way to make a *complete graph* in which any pair of nodes is connected by a link, i.e. $l = n(n-1)/2$, and

$$\mathcal{N}(n \text{ in complete graph}) = \frac{N!}{(N-n)!n!}p^{n(n-1)/2}\,. \tag{4.4}$$

4.2.1 Percolation

A network can display two types of global connectivity. With few connections amongst nodes, there will be many disjoint clusters, with their typical size (but not necessarily number)

increasing with the number of connections. At high connectivity there will be one very large cluster, and potentially a number of smaller clusters. In the limit of $N \to \infty$, a well defined *percolation transition* separates the two regimes in the random graph, as the probability p is varied.

Above the percolation transition, the number of nodes M in the largest cluster also goes to infinity, proportionately to the number of nodes, such that there is a finite *percolation probability*

$$P(p) = \lim_{N\to\infty} \frac{M}{N} = \text{Probability to belong to the infinite cluster.}$$

For the random graph $P(p)$ can be calculated from a self-consistency argument: Take a particular site and consider the probability that it is *not* connected to the infinite cluster. This is the case if none of the $(N-1)$ edges emanating from this site connect it to the large cluster. A particular edge connects to the infinite cluster with probability $pP(p)$ (that the edge exists and that the adjoining site is on the large cluster), and hence

$$\begin{aligned} 1 - P(p) &= (\text{prob. of no connections via any edge})^{N-1} \\ &= (1 - pP)^{N-1} \,. \end{aligned} \tag{4.5}$$

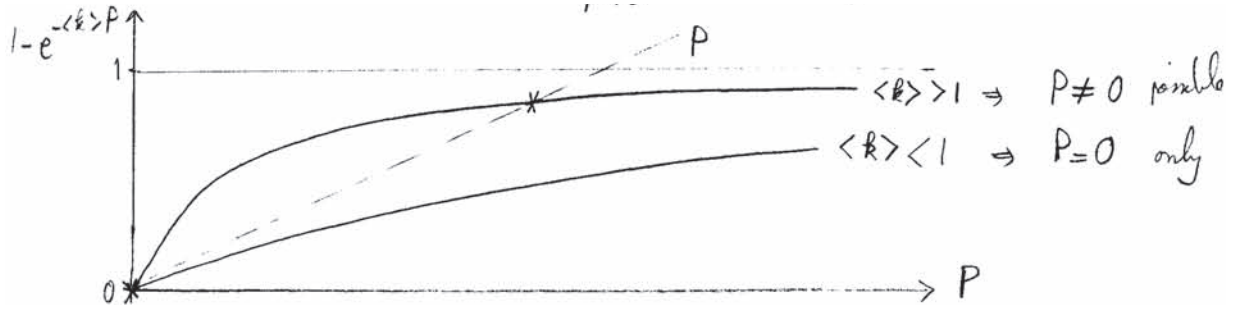
There is a phase transition in the limit $N \to \infty$, provided that $p \to 0$, such that

$$p(N-1) = \langle k \rangle \,, \tag{4.6}$$

where $\langle k \rangle$, the number of expected edges per node, is finite. In this limit, we can re-express Eq. (4.5) as

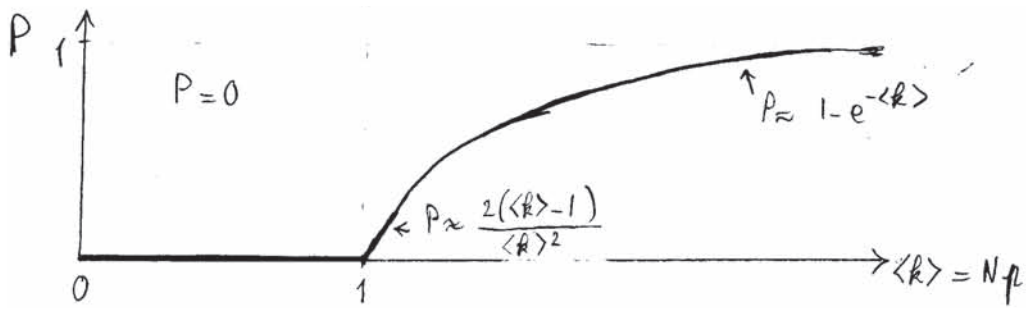
$$1 - P = e^{-\langle k \rangle P} \,, \quad \Rightarrow \quad P = 1 - e^{-\langle k \rangle P} \,. \tag{4.7}$$

The above equation can be solved self-consistently, for example graphically, For $\langle k \rangle \leq 1$, the



only solution is $P = 0$, while for $\langle k \rangle > 1$ a finite P is possible, indicating the appearance of an infinite cluster. Close to the percolation transition at $\langle k \rangle_c = 1$, P is small and we can expand Eq. (4.7) as

$$P = \langle k \rangle \, P - \frac{\langle k \rangle^2}{2} P^2 + \mathcal{O}(P^3) \,, \quad \Rightarrow \quad P \approx \frac{2\left(\langle k \rangle - 1\right)}{\langle k \rangle^2} \approx 2\left(\langle k \rangle - 1\right) \,. \tag{4.8}$$



4.2.2 Distance, Diameter, & Degree Distribution

There are typically several ways to traverse from a node i to a node j. The *distance* between any pair of nodes is defined as the number of edges along the *shortest path* between the nodes. For the entire network, we can define a *diameter* as the largest of all distances between pairs of nodes.

Distances to a particular node can be obtained efficiently by the following simple (burn and move) algorithm. In the first step, label the nodes connected to the starting point ($d = 1$), and then remove it from the network. Consider a random graph with $\langle k \rangle \gg 1$, such that $P \approx 1$. (Distances cannot be defined to disconnected clusters.) In the random graph, the number of sites with $d = 1$ will be around $p(N-1) = \langle k \rangle$. In the second step identify all sites connected to the set labeled before (and thus at $d = 2$), and then remove all sites with $d = 1$ from the network. From each site with $d = 1$, there are of the order of $p(N - \langle k \rangle - 1) \approx \langle k \rangle$ accessible sites, since $\langle k \rangle \ll N$. There are thus around $\langle k \rangle^2$ sites labelled with $d = 2$. This burn and move process can be repeated, with $N_p \lessapprox \langle k \rangle^p$ sites tagged at distance $d = p$. (Note that each step we have overestimated the number of sites by ignoring connections leading to sites already removed.) The procedure has to be stopped when all sites belonging to the cluster have been removed, i.e. for

$$\langle k \rangle^D \lessapprox N\,, \quad \Rightarrow \quad D \lessapprox \frac{\ln N}{\ln \langle k \rangle}\,, \tag{4.9}$$

where D is a rough measure of the diameter of the network. Note that the diameter of a random network is quite small, justifying the popular lore of "six degrees of separation." In a population of a few billion, with each individual knowing a few thousand, Eq. (4.9) in fact predicts a distance of three or four between any two. Clearly segregation by geographical and social barriers increases this distance. The model of "small world networks" considers mostly segregated communities, but shows that even a small fraction of random links is sufficient to reintroduce a logarithmic behavior ala Eq. (4.9).

For $\langle k \rangle < 1$, the typical situation is of disjoint clusters. We can then inquire about the probability p_k that there are exactly k links emanating from a site. Since there are a total of $(N-1)$ potential connections from a site, in a random graph the probability that k such

links are active is given by the binomial probability

$$p_k = \left(\begin{array}{c} N-1 \\ k \end{array} \right) p^k (1-p)^{N-1-k} \,. \tag{4.10}$$

Taking the limits $N \to \infty$ and $p \to 0$ with $pN = \langle k \rangle$ as before, we obtain

$$\begin{aligned} p_k &= \frac{N^k}{k!} \frac{p^k}{(1-p)^k} (1-p)^{N-1} \\ &= \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle} \,, \end{aligned} \tag{4.11}$$

i.e. a Poisson distribution with mean $\langle k \rangle$. The above results for random graphs can be used as a potential model for assessing significance of putative anomalies in the the degree distributions of social and biological networks.

4.3 The Barabasi-Albert model

While examining a variety of networks, initially in the context of internet, Barabasi and Albert obtained degree distributions that were quite different from Poisson. Rather than the exponential behavior of Eq. (4.11) they observed a tendency for distributions to fall more slowly for large k, approximately as a power law $1/k^3$. To explain this observation, they noted that networks such as internet are obtained through a dynamic growth process, with new nodes added to a pre-existing network. They postulated that newly added nodes are more likely to link to popular pre-existing nodes, the latter thus becoming even more popular (the rich become richer) over time.

Consider an algorithm in which nodes are added to the network one at a time. Each new node makes exactly m links, but the probability that a link is made to a pre-existing node is proportional to the number of links already emanating from that node. Starting with no nodes or links, after t steps, the network will have $N(t) = t$ nodes, and $L(t) = mt$ links. A particular realization of this algorithm can be described by the set $\{N_k(t)\}$ of the number of nodes with k links. After the next node is added, these numbers change, such that

$$\begin{aligned} N_k(t+1) = N_k(t) & \\ &+ \#\text{ of nodes with } (k-1) \text{ edges connected to the new node} \\ &- \#\text{ of nodes with } k \text{ edges connected to the new node} \\ &+ \delta_{m,k} \,. \end{aligned} \tag{4.12}$$

The rule for *preferential attachment* states that the probability of attachment to a site with k links is $k/(\sum_{k'} k' N_{k'})$. Since there are m new links to be added, Eq. (4.12) implies that *on average*

$$\langle N_k(t+1) \rangle = \langle N_k(t) \rangle + \langle N_{k-1}(t) \rangle \cdot \frac{k-1}{\sum_{k'} k' N_{k'}(t)} \cdot m - \langle N_k(t) \rangle \cdot \frac{k}{\sum_{k'} k' N_{k'}(t)} \cdot m + \delta_{m,k} \,. \tag{4.13}$$

For large t, we hypothesize that the system evolves towards fixed probabilities p_k for finding nodes with k edges, such that

$$\lim_{t\gg 1} N_k(t) = N(t)p_k = tp_k\,. \tag{4.14}$$

Substituting Eq. (4.14) into Eq. (4.13), and noting that $\sum_{k'} k'N_{k'}(t) = L(t) = 2mt$, gives

$$(t+1)p_k = tp_k + p_{k-1}\frac{k-1}{2} - p_k\frac{k}{2} + \delta_{m,k}\,. \tag{4.15}$$

We see that terms involving t cancel out, justifying our assumption of a steady state, and yielding a recursion relation for the probabilities, as

$$\left(1+\frac{k}{2}\right)p_k = \left(\frac{k-1}{2}\right)p_{k-1} + \delta_{m,k}\,. \tag{4.16}$$

For $k > m$, successive recursions give

$$\begin{aligned} p_k &= \frac{k-1}{k+2}p_{k-1} = \frac{(k-1)(k-2)}{(k+2)(k+1)}p_{k-2} = \cdots \\ &= \frac{(k-1)(k-2)\cdot m}{(k+2)(k+1)\cdots(m+3)}p_m = \frac{(m+2)(m+1)m}{(k+2)(k+1)k}p_m\,. \end{aligned}$$

For $k = m$, the addition to probability is not from p_{k-1} but from the new nodes, and thus

$$\left(1+\frac{m}{2}\right)p_m = 1\,, \quad \Rightarrow \quad p_m = \frac{2}{m+2}\,. \tag{4.17}$$

We thus find the final distribution

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}\,, \tag{4.18}$$

which is properly normalized as $\sum_k p_k = 1$. The $k \gg 1$ the tail of this distribution indeed decays as k^{-3}, as empirically observed for the internet and a number of other cases.

8.592J / HST.452J Statistical Physics in Biology
Spring 2011

8.592J / HST.452J Statistical Physics in Biology
Spring 2011