Handbook of Statistical Systems Biology

CONTENTS

1

Modelling biological networks via tailored random graphs

ACC Coolen, F Fraternali, A Annibale, L Fernandes, and J Kleinjung

King's College London

1.1 Introduction

The nature of biological research has changed irreversibly in the last decade. Experimental advances have generated unprecedented amounts of genetic and structural information at molecular levels of resolution, and to understand the biological systems that are now being studied, knowing the list of parts no longer suffices. The parts have become too complicated and too numerous. It is often not even clear how best to represent the new high-dimensional experimental data in a way that helps us making sense of them. We need to integrate our knowledge of the individual biological events and components in mathematical and computational models that can capture the complexity of the data at a systems level.

Data collected on metabolic, regulatory or signalling processes in the cell are usually represented in the form of networks, with nodes representing dynamical variables (metabolites, enzymes, RNA or protein concentrations) and links between nodes representing pairs of variables that have been observed to interact with each other. Some of these networks are directed (e.g. metabolic and gene regulatory networks, GRN), and some are undirected (e.g. protein-protein interaction networks, PPIN). The idea behind this representation is that functional properties of complex cellular processes will have fingerprints in the structure of their networks. Most of the observed biological networks, however, are too complex to allow for direct interpretations; to proceed we need precise tools for quantifying their topologies.

Ensembles of tailored random graphs with controlled topological properties are a natural and rigorous language for describing biological networks. They suggest precise definitions of structural features, they allow us to classify networks and obtain precise (dis)similarity measures, they provide 'null models' for hypothesis testing, and they can serve as efficient proxies for real networks in process modelling. In this chapter we explain the connection between biological networks and tailored random graphs, we show how this connection can be exploited, and we discuss exact algorithms for generating such graphs numerically.

1.2 Quantitative Characterization of Network Topologies

We consider networks with N nodes ('vertices') labelled by Roman indices. For each node pair (i,j) we write $c_{ij}=1$ if a link ('edge') $j\to i$ is present, and $c_{ij}=0$ if not. The set of N^2 link variables $\{c_{ij}\}$ specifies the network in full, and is abbreviated as $\mathbf{c}$. We limit ourselves in this chapter to non-directed networks (such as PPINs), where $c_{ij}=c_{ji}$ for all (i,j), and we assume that $c_{ii}=0$ for all i. We denote the set of all such non-directed networks as $G=\{0,1\}^{\frac{1}{2}N(N-1)}$. The specific biological networks we are interested in tend to be large, containing of the order $N\sim 10^4$ nodes, but with a small average number $\langle k\rangle=N^{-1}\sum_{ij}c_{ij}$ of links per node. The current estimate for e.g. the human PPIN is $\langle k\rangle\sim 7$.

1.2.1 Local Network Features and Their Statistics

To characterize networks quantitatively a natural first step is to inspect simple local quantities and their distributions (Albert and Barabási 2002, Dorogovtsev *et al.* 2008, Newman 2003), such as the degrees $k_i(\mathbf{c})=\sum_j c_{ij}$ (the number of partners of node i) or the clustering coefficients $C_i(\mathbf{c})=[\sum_{j\neq k}c_{ij}c_{ik}c_{jk}]/[\sum_{j\neq k}c_{ij}c_{ik}]$ (the fraction of the partners of i that are themselves connected). For instance, the distribution of the N degrees[1],

$$p(k|\mathbf{c})=\frac{1}{N}\sum_i \delta_{k,k_i(\mathbf{c})} \tag{1.1}$$

gives us a simple and transparent characterisation of the network's topology. Often we would in addition like to capture correlations between local properties of different nodes, especially between connected nodes, which prompts us to define also distributions such as

$$W(k,k'|\mathbf{c})=\frac{1}{N\langle k\rangle}\sum_{ij}\delta_{k,k_i(\mathbf{c})}c_{ij}\delta_{k',k_j(\mathbf{c})} \tag{1.2}$$

i.e. the fraction of *connected* node pairs (i,j) with degrees (k,k'). From (1.2) follows the assortativity (Newman 2002), the overall correlation between the degrees of connected nodes:

$$a(\mathbf{c})=[\langle kk'\rangle_w-\langle k\rangle_w^2]/[\langle k^2\rangle_w-\langle k\rangle_w^2] \tag{1.3}$$

with the short-hand $\langle f(k,k')\rangle_w=\sum_{kk'}f(k,k')W(k,k'|\mathbf{c})$, and where we used the symmetry of $W(k,k'|\mathbf{c})$. Both (1.1) and (1.2) are global measures of network structure, and they provide complementary information. They have the advantage of not depending explicitly upon the network size N (such a dependence would be undesirable, as most biological data sets are known to represent incomplete samples and the sizes of available biological networks continue to increase). However, (1.1) and (1.2) are not independent, since

$$W(k|\mathbf{c})=\sum_{k'}W(k,k'|\mathbf{c})=\frac{k}{\langle k\rangle}p(k|\mathbf{c}) \tag{1.4}$$

More generally one could have for each node i a list $\mathbf{k}_i(\mathbf{c})=(k_{i1}(\mathbf{c}),\ldots,k_{ir}(\mathbf{c}))$ of local quantities. Choosing e.g. $k_{i\ell}(\mathbf{c})=(\mathbf{c}^{\ell+1})_{ii}$ would give $k_{i2}(\mathbf{c})=\sum_{jm}c_{ij}c_{jm}c_{mi}$ (the

[1] Here $\delta_{ab}=1$ if $a=b$, and $\delta_{ab}=0$ if $a\neq b$.

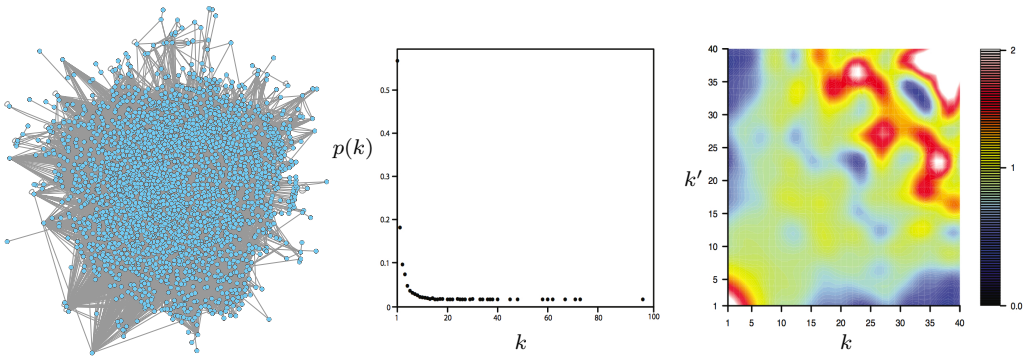


Figure 1.1 Windows on the *H. sapiens* PPIN data in (Prasad *et al.* 2009), with $N = 9306$ proteins and $\langle k\rangle = 7.53$ interactions/node on average. Left: data shown as a non-directed network, with proteins as nodes and physical interactions as links. Middle: degree distribution $p(k)$, defined in (1.5). Right: re-scaled joint degree statistics of connected nodes $\Pi(k,k') = W(k,k')/W(k)W(k')$, following (1.6). One would have $\Pi(k,k') = 1$ for all (k,k') if connected nodes had uncorrelated degrees, so deviations from light green suggest nontrivial structural properties. Figure taken from (Fernandes *et al.* 2010).

number of length-3 paths through i)[2], followed by counters of longer loops. The choice $k_{i\ell}(\mathbf{c}) = \sum_j (\mathbf{c}^\ell)_{ij}$ would give observables that count the number of paths through each node of a given length (open or closed)[3]. The distributions (1.1,1.2) would then generalize to

$$p(\mathbf{k}|\mathbf{c}) = \frac{1}{N}\sum_i \delta_{\mathbf{k},\mathbf{k}_i(\mathbf{c})} \tag{1.5}$$

$$W(\mathbf{k},\mathbf{k}'|\mathbf{c}) = \frac{1}{N\langle k\rangle}\sum_{ij} \delta_{\mathbf{k},\mathbf{k}_i(\mathbf{c})} c_{ij} \delta_{\mathbf{k}',\mathbf{k}_j(\mathbf{c})} \tag{1.6}$$

with (1.5) giving the overall fraction of nodes in the network with local properties $\mathbf{k}$, and (1.6) giving the fraction of *connected* node pairs (i,j) with local properties $(\mathbf{k},\mathbf{k}')$.

In this chapter we will mainly work with structure characterisations of the above form. However, we note that there are alternatives. One is the network spectrum $\varrho(\mu|\mathbf{c}) = N^{-1}\sum_i \delta[\mu - \mu_i(\mathbf{c})]$, where the $\mu_i(\mathbf{c})$ are the eigenvalues of the matrix $\mathbf{c}$; from it one obtains the joint distribution of loops of all lengths, see e.g. (Kühn 2008, Rogers *et al.* 2010) and references therein. Another alternative is the spectrum of the network Laplacian $L(\mathbf{c})$, a matrix defined as $L_{ij}(\mathbf{c}) = k_i(\mathbf{c})\delta_{ij} - c_{ij}$; it contains information on sub-graph statistics, modularity, and typical path distances, see e.g. (Mohar 1991) and references therein.

1.2.2 Examples

Figure 1.1 illustrates the topology characterization (1.1,1.2) for the PPIN of *H. sapiens*. It is clear that from the network image itself (on the left) one cannot extract much useful information. Instead we characterise the network structure hierarchically by measuring increasingly sophisticated degree-related quantities. The zero-th level is to measure the

[2] From $k_{i1}(\mathbf{c})$ and $k_{i2}(\mathbf{c})$ the clustering coefficient $C_i(\mathbf{c})$ follows via $C_i(\mathbf{c}) = k_{i2}(\mathbf{c})/k_{i1}(\mathbf{c})[k_{i1}(\mathbf{c}) - 1]$.
[3] These so-called generalized degrees were to our knowledge first proposed in (Skantzos, 2005).

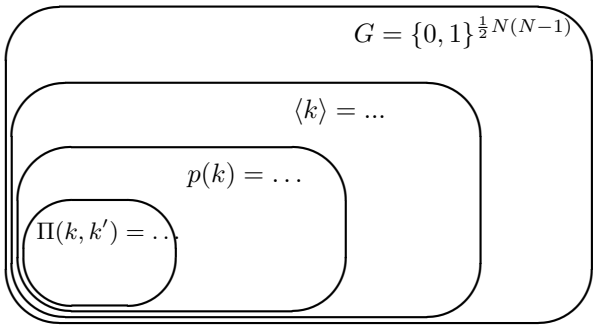


Figure 1.2 Classification of networks into hierarchically organized families. Starting from the set of all non-directed graphs of size N, we first classify networks according to their average connectivities, followed by class sub-division for each $\langle k \rangle$ according to the degree distribution $p(k)$, and by further class sub-division for each $p(k)$ according to the degree correlation profile $\Pi(k,k') = W(k,k')/W(k)W(k')$, etc. By construction, the subsets will decrease in size with each sub-division, i.e. the specification of networks becomes increasingly detailed and prescriptive.

average degree $\langle k \rangle$. The next level is measuring the degree distribution $p(k|\mathbf{c})$ (1.1). Finally, we collect degree statistics of connected nodes by probing $W(k,k'|\mathbf{c})$ (1.2). To aid our interpretation of the latter distribution we plot, rather than $W(k,k'|\mathbf{c})$ itself, the ratio

$$\Pi(k,k'|\mathbf{c}) = W(k,k'|\mathbf{c})/W(k|\mathbf{c})W(k'|\mathbf{c}) \tag{1.7}$$

where the marginals $W(k|\mathbf{c})$ follow via (1.4). A weak Gaussian smoothening is applied to $\Pi(k,k')$, to prevent trivial pathologies in calculations. Since $W(k|\mathbf{c})$ can be written directly in terms of $p(k|\mathbf{c})$, only the ratio $\Pi(k,k'|\mathbf{c})$ can reveal topological information (if any) that is not already contained in the degree distribution. Any significant deviation from $\Pi(k,k'|\mathbf{c}) = 1$ tells us that the network wiring contains nontrivial regularities beyond those encoded in the degree distribution, which manifest themselves in either a higher (red) or a lower (blue) than expected tendency of degree pairs (k,k') to interact.

In dealing with real data one should be aware, however, that these are generally incomplete samples of a true underlying biological network, and that sampling impacts on the shape of distributions such as (1.5,1.6) (Han *et al.* 2005, Stumpf and Wiuf 2005). Both links and nodes will be under-sampled, and the sampling is likely to be biased (Hakes *et al.* 2008). For instance, PPIN data sets are influenced by experimentalists' focus on proteins that are 'interesting' or easy to measure, by experimental protocol, and even by data processing.

1.3 Network Families and Random Graphs

1.3.1 Network Families, Hypothesis Testing and Null Models

Any quantitative characterization of networks via increasingly detailed structure measurements, of which $\langle k \rangle$, $p(k)$ and $W(k,k')$ are specific examples, induces automatically a hierarchical classification of all networks into families, see Figure 1.2. This is not a deep insight, but it does aid our formulation of practical concepts and questions. Let us denote with

$G[p] \subseteq G$ the subset of networks with topological features characterized by a given choice for the function $p(k)$ (see Figure 1.2), and let us write the number of networks in that subset as $|G[p]|$. Similarly, let us denote with $G[p, \Pi] \subseteq G[p]$ the subset of networks with topological features characterized by specified choices for both $p(k)$ and $\Pi(k, k')$, with $|G[p, \Pi]|$ giving the number of such networks. It is then natural to define network comparison, observation interpretation, and hypothesis testing along the following lines:

- *A network with features* $\{p, \Pi\}$ *is more complex than a network with features* $\{p', \Pi'\}$ *if* $|G[p, \Pi]| < |G[p', \Pi']|$.

  The rationale is this: the smaller the number of networks with given features $\{p, \Pi\}$, i.e. the smaller the associated compartment in Figure 1.2, the more difficult it will be to find or construct a network with these specific features.

- *Measuring a value* Ω *for some observable* $\Omega(\mathbf{c})$ *in a network* $\mathbf{c}^\star \in G[\ldots]$ *is trivial for the family* $G[\ldots]$ *if most of the networks* $\mathbf{c} \in G[\ldots]$ *exhibit* $\Omega(\mathbf{c}) = \Omega$.

  Especially in large networks, where usually $|G[p, \Pi]| \ll |G[p]|$, an observation may be nontrivial for $G[p]$ but trivial once we limit ourselves to $G[p, \Pi]$. For instance, in the set of *all graphs* with average degree $\langle k \rangle$ the vast majority will have Poissonian degree statistics, so observing $\langle k^2 \rangle > 2\langle k^2 \rangle$ becomes highly unlikely for large N. Yet, once we limit ourselves further to networks with power-law degree distributions the previously unlikely event becomes ordinary.

- *To test a hypothesis that an observation* $\Omega(\mathbf{c}^\star) = \Omega$ *in a network* $\mathbf{c}^\star$ *is atypical, we must define a null-hypothesis in terms of one of the above sets* $G[\ldots]$. *The* p*-value of the test is then the probability to observe* $\Omega(\mathbf{c}) = \Omega$ *(or a more extreme value) if we pick graphs* $\mathbf{c}$ *randomly from* $G[\ldots]$.

  In analogy with our previous observations, an observation may have a very small p-value (and be interpreted as important) if we choose a large and diverse family of networks, say $G[p]$, but may be recognized as trivial once we limit ourselves to the subset $G[p, \Pi]$ to which $\mathbf{c}^\star$ belongs. In the latter case we would say that the observation $\Omega(\mathbf{c}^\star) = \Omega$ is a strict consequence of its degree correlations, as measured by $\Pi(k, k')$.

Many important questions relating to quantifying network structures and to interpretation of observations apparently involve calculating averages over constrained sets of randomly generated networks, and counting the number of networks with specific structural features (Milo *et al.* 2002, Holme and Zhao 2007, Foster *et al.* 2007). This is the connection between biological networks and tailored random graph ensembles.

1.3.2 Tailored Random Graph Ensembles

Random graph ensembles (see e.g. Erdös and Rényi 1959, Molloy and Reed 1995, Watts and Strogatz 1998, Barabási and Albert 1999) give us a mathematical framework within which to make our ideas precise, and allow us to apply methods from information theory and statistical mechanics (Park and Newman 2004, Garlaschelli and Loffredo 2008). Random graph ensembles are defined by a set of allowed graphs, here taken to be (a subset of) G, and a measure $p(\mathbf{c})$ that tells us how likely each $\mathbf{c} \in G$ is to be generated. The ensembles found in the previous section were all of the following form. We prescribed as constraints

the values for specific observables, i.e. $\Omega_\mu(\mathbf{c}) = \Omega_\mu$ for $\mu = 1 \ldots p$, and demanded that only graphs that met the constraints were included, each with uniform weight:

$$p_{\rm h}(\mathbf{c}|\mathbf{\Omega}) = Z_{\rm h}^{-1}(\mathbf{\Omega})\, \delta_{\mathbf{\Omega}(\mathbf{c}),\mathbf{\Omega}}, \qquad Z_h(\mathbf{\Omega}) = \sum_{\mathbf{c}\in G} \delta_{\mathbf{\Omega}(\mathbf{c}),\mathbf{\Omega}} \tag{1.8}$$

with $\mathbf{\Omega} = (\Omega_1, \ldots, \Omega_p)$. Such ensembles have maximum Shannon entropy (Cover and Thomas 1991), given the imposed 'hard' constraints, so in an information-theoretic sense the *only* structural information built into our graphs is that imposed by the constraints. Alternatively, one could relax the constraints and instead of $\mathbf{\Omega}(\mathbf{c}) = \mathbf{\Omega}$ for all $\mathbf{c} \in G$ demand that these constraints are satisfied *on average*, i.e. that $\sum_{\mathbf{c}\in G} p(\mathbf{c})\mathbf{\Omega}(\mathbf{c}) = \mathbf{\Omega}$. Maximising Shannon's entropy under the latter 'soft' constraint give the so-called exponential family

$$p_{\rm s}(\mathbf{c}|\mathbf{\Omega}) = Z_s^{-1}(\mathbf{\Omega})\, {\rm e}^{\sum_\mu \omega_\mu(\mathbf{\Omega})\Omega_\mu(\mathbf{c})}, \qquad Z_s(\mathbf{\Omega}) = \sum_{\mathbf{c}\in G} {\rm e}^{\sum_\mu \omega_\mu(\mathbf{\Omega})\Omega_\mu(\mathbf{c})} \tag{1.9}$$

where the parameters $\omega_\mu(\mathbf{\Omega})$ must be solved from the equations $\sum_{\mathbf{c}\in G} p_s(\mathbf{c}|\mathbf{\Omega})\Omega_\mu(\mathbf{c}) = \Omega_\mu$. In contrast to (1.8), not all graphs generated by (1.9) will exhibit the properties $\mathbf{\Omega}(\mathbf{c}) = \mathbf{\Omega}$, but for observables $\mathbf{\Omega}(\mathbf{c})$ that are macroscopic in nature one will generally find even in (1.9) deviations from the 'hard' condition $\mathbf{\Omega}(\mathbf{c}) = \mathbf{\Omega}$ to tend to zero as N becomes large.

The normalization factor in (1.8) equals the number of graphs with the property $\mathbf{\Omega}(\mathbf{c}) = \mathbf{\Omega}$, and can also be written in terms of the Shannon entropy of this ensemble via $Z_h(\mathbf{\Omega}) = \exp[-\sum_{\mathbf{c}\in G} p_h(\mathbf{c}|\mathbf{\Omega}) \log p_h(\mathbf{c}|\mathbf{\Omega})]$. For 'soft' constrained ensembles (1.9) all graphs $\mathbf{c} \in G$ could in principle emerge. However, some are much more likely than others and one can define in that case more generally an *effective* number of graphs $\mathcal{N}(\mathbf{\Omega})$, via the connection with entropy. So we have more generally for either ensemble

$$\mathcal{N}[\mathbf{\Omega}] = {\rm e}^{S[\mathbf{\Omega}]}, \qquad S[\mathbf{\Omega}] = -\sum_{\mathbf{c}\in G} p(\mathbf{c}|\mathbf{\Omega}) \log p(\mathbf{c}|\mathbf{\Omega}) \tag{1.10}$$

and one must expect for large N and macroscopic observables $\mathbf{\Omega}(\mathbf{c})$ that the leading order of $S(\mathbf{\Omega})$ does not depend on whether we use (1.8) or (1.9).

For instance, if we choose as our constraining observable only the average degree $N^{-1}\sum_{ij} c_{ij}$, we obtain the following maximum entropy ensembles (upon rewriting the measure for the soft constraint ensemble, and after solving its Lagrange parameter equation):

$$p_{\rm h}(\mathbf{c}|\langle k\rangle) = Z_{\rm h}^{-1}(\langle k\rangle)\, \delta_{\sum_{ij} c_{ij}, N\langle k\rangle} \tag{1.11}$$

$$p_{\rm s}(\mathbf{c}|\langle k\rangle) = \prod_{i<j} \Big[\frac{\langle k\rangle}{N} \delta_{c_{ij},1} + (1 - \frac{\langle k\rangle}{N}) \delta_{c_{ij},0} \Big] \tag{1.12}$$

The latter is the well known Erdös-Rényi random graph ensemble (Erdös and Rényi 1959). Both are tailored to the production of random graphs with average connectivity $\langle k\rangle$, and are otherwise strictly unbiased. Specializing further, in the spirit of Figure 1.2, we next constrain the full degree sequence $\mathbf{k} = (k_1, \ldots, k_N)$ (equivalent to fixing the degree distribution $p(k)$, apart from node permutation). We then obtain the following maximum entropy ensembles:

$$p_{\rm h}(\mathbf{c}|\mathbf{k}) = Z_{\rm h}^{-1}(\mathbf{k}) \prod_i \delta_{\sum_j c_{ij}, k_i} \tag{1.13}$$

$$p_{\rm s}(\mathbf{c}|\mathbf{k}) = \prod_{i<j} \Big[\frac{{\rm e}^{\omega_i+\omega_j}}{1+{\rm e}^{\omega_i+\omega_j}} \delta_{c_{ij},1} + \frac{1}{1+{\rm e}^{\omega_i+\omega_j}} \delta_{c_{ij},0} \Big] \tag{1.14}$$

with the $\{\omega_i\}$ to be solved from the N equations $k_i = \sum_{\ell\neq i}(1+\mathrm{e}^{-\omega_i-\omega_\ell})^{-1}$. Both ensembles (1.13,1.14) are tailored to the production of random graphs with degrees $\mathbf{k}$, and are otherwise strictly unbiased. Specializing further to the level where, for instance, all degrees $\mathbf{k}$ as well as the joint distribution $W(k,k')$ (1.2) are precribed gives us the graph ensembles

$$p_{\rm h}(\mathbf{c}|\mathbf{k},W) = \frac{\delta_{\mathbf{k},\mathbf{k}(\mathbf{c})}}{Z_{\rm h}(\mathbf{k},W)} \prod_{kk'} \delta_{\sum_{ij}\delta_{k,k_i(\mathbf{c})}c_{ij}\delta_{k',k_j(\mathbf{c})},N\langle k\rangle W(k,k')} \tag{1.15}$$

$$p_{\rm s}(\mathbf{c}|\mathbf{k},W) = \frac{1}{Z_{\rm s}(\mathbf{k},W)}\,\mathrm{e}^{\sum_{i<j}c_{ij}[\omega_i+\omega_j+\psi(k_i(\mathbf{c}),k_j(\mathbf{c}))+\psi(k_j(\mathbf{c}),k_i(\mathbf{c}))]} \tag{1.16}$$

with the $\{\omega_i\}$ and $\{\psi(k,k')\}$ to be solved from the equations $k_i = \sum_{\mathbf{c}\in G} p_{\rm s}(\mathbf{c}|\mathbf{k},W)\sum_j c_{ij}$ and $\sum_{\mathbf{c}\in G} p_{\rm s}(\mathbf{c}|\mathbf{k},W)\sum_{ij}\delta_{k,k_i(\mathbf{c})}c_{ij}\delta_{k',k_j(\mathbf{c})} = N\langle k\rangle W(k,k')$, respectively.

Increasing the complexity of our observables gives us increasingly sophisticated tailored random graphs, which share more and more features with the biological networks one aims to study or mimic, but the price paid is mathematical and computational complexity. For instance, for realistic network sizes it will be hard to solve the equations for the Lagrange parameters reliably in ensembles such as (1.16) by numerical sampling of the space G. Even for $N = 1000$ this space already contains $2^{\frac{1}{2}N(N-1)} \approx 10^{150,364}$ graphs (to put this number into perspective, there are estimated to be only around 10^{82} atoms in the universe ...).

1.4 Information-Theoretic Deliverables of Tailored Random Graphs

If we choose our ensembles carefully, and our networks are sufficiently large, it is possible to proceed analytically. We want to define our ensembles up to the complexity limit where the various sums over all $\mathbf{c}\in G$ can in leading order in N still be calculated mathematically. There is little point limiting ourselves to (1.11,1.12), since they typically generate graphs with Poissonian degree distributions which are very different from our biological networks (Barabási and Albert 1999). So we focus on (1.13,1.14) and (1.15,1.16). Since we know the degrees of our biological networks, and since it is not hard to handle hard degree constraints, we choose our ensembles such that $\mathbf{k} = \mathbf{k}(\mathbf{c})$ for all $\mathbf{c}$. However, incorporating the wiring information contained in $W(k,k')$ or in the ratio $\Pi(k,k') = W(k,k')/W(k)W(k')$ is easier using a soft constraint. The following choice was studied in detail in (Annibale *et al.* 2009):

$$p(\mathbf{c}|\mathbf{k},\Pi) = \frac{\delta_{\mathbf{k},\mathbf{k}(\mathbf{c})}}{Z(\mathbf{k},\Pi)}\prod_{i<j}\left[\frac{\langle k\rangle}{N}Q(k_i,k_j)\delta_{c_{ij},1} + \left(1-\frac{\langle k\rangle}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right] \tag{1.17}$$

with $Q(k,k') = \Pi(k,k')kk'/\langle k\rangle^2$. In fact, in (Annibale *et al.* 2009) the degree distribution $p(k)$ was constrained, instead of the sequence $\mathbf{k}$, giving the modestly different starting point

$$p(\mathbf{c}|p,\Pi) = \sum_{\mathbf{k}}[\prod_i p(k_i)]p(\mathbf{c}|\Pi,\mathbf{k}) \tag{1.18}$$

The ensemble (1.17) is tailored to the production of random graphs with degrees $\mathbf{k}$ (via a 'hard' constaint) and with joint degree statistics of connected nodes characterized by $\Pi(k,k')$ (via a 'soft' constraint), and is otherwise unbiased. It is the maximum entropy ensemble if we constrain the values of all degrees and the expectation value of the joint distribution $W(k,k')$ in (1.2). For $N\to\infty$ the 'soft' deviations from $W(k,k'|\mathbf{c}) = W(k,k')$ will vanish.

1.4.1 Network Complexity

We saw that the complexity of graphs with given properties $\mathbf{\Omega}(\mathbf{c}) = \mathbf{\Omega}$ is related to the *number* of graphs that exist with these properties. This number is expressed in terms of the Shannon entropy of the random graph ensemble $p(\mathbf{c}|\mathbf{\Omega})$ via (1.10). It turns out that for the ensemble (1.18) one can calculate analytically the leading orders in N of the entropy, via statistical mechanical techniques (e.g. path integrals and saddle-point integration), and thus avoid the need for numerical sampling to find Lagrange parameters. The result is an exact expression for the Shannon entropy $S[p,\Pi]$ and for the effective number of graphs $\mathcal{N}[p,\Pi]$ with degree distribution $p(k)$ and degree statistics of connected nodes given by $\Pi(k,k')$:

$$S[p,\Pi]/N = S_0 - \mathcal{C}[p,\Pi] + \epsilon_N, \qquad \mathcal{N}[p,\Pi] = \mathrm{e}^{S[p,\Pi]} \tag{1.19}$$

in which ϵ_N represents a finite size correction term that will vanish as $N \to \infty$, and S_0 is the Shannon entropy per node one would have found for the trivial ensembles (1.11,1.12), viz.

$$S_0 = \frac{1}{2}\langle k\rangle\big[\log[N/\langle k\rangle + 1\big] \tag{1.20}$$

The most interesting term in (1.19) is $\mathcal{C}[p,\Pi]$, which tells us precisely when and how the imposition of the structural properties $p(k)$ and $\Pi(k,k')$ reduces the space of compatible graphs, in the spirit of figure (1.2). It contains two non-negative contributions:

$$\mathcal{C}[p,\Pi] = \sum_k p(k)\log\Big[\frac{p(k)}{\pi(k)}\Big] + \frac{1}{2\langle k\rangle}\sum_{kk'} p(k)p(k')kk'\Pi(k,k')\log\Pi(k,k') \tag{1.21}$$

Here $\pi(k) = e^{-\langle k\rangle}\langle k\rangle^k/k!$ is the Poissonian degree distribution with average degree $\langle k\rangle$ one would have found for the ensemble (1.12). If the degrees of connected nodes are uncorrelated, so $\Pi(k,k') = 1$ for all (k,k'), the second term of (1.21) will vanish. Hence the first term represents the complexity per node generated by the degree distribution alone; it is seen to increase as the degree distribution becomes more dissimilar from a Poissonian one (measured via a Kullback-Leibler distance). The second term of (1.21) represents the excess complexity per node generated by preferential wiring of the network, beyond the complexity induced by the imposed degrees. For the derivation of the above formulae we refer to (Annibale *et al.* 2009) and its precursors (Pérez-Vicente and Coolen 2008, Bianconi *et al.* 2008).

1.4.2 Information-Theoretic Dissimilarity

Information-theory also provides measures for the dissimilarity between networks $\mathbf{c}_A$ and $\mathbf{c}_B$, that take account of the probabilistic nature of network data by being formulated in terms of the associated random graph measures $p(\mathbf{c}|p_A,\Pi_A)$ and $p(\mathbf{c}|p_B,\Pi_B)$. One has a choice of definitions, but most are very similar and even identical when the underlying distributions become close. One of the simplest formulae is Jeffrey's divergence, which after a simple re-scaling leads to the following distance between networks $\mathbf{c}_A$ and $\mathbf{c}_B$:

$$D_{AB} = \frac{1}{2N}\sum_{\mathbf{c}\in G}\Big\{p(\mathbf{c}|p_A,\Pi_A)\log\Big[\frac{p(\mathbf{c}|p_A,\Pi_A)}{p(\mathbf{c}|p_B,\Pi_B)}\Big] + p(\mathbf{c}|p_B,\Pi_B)\log\Big[\frac{p(\mathbf{c}|p_B,\Pi_B)}{p(\mathbf{c}|p_A,\Pi_A)}\Big]\Big\} \tag{1.22}$$

Again the sums over all graphs in G can be calculated in leading order in the system size, and after taking $N \to \infty$ the end result is once more surprisingly simple, explicit and transparent:

$$
\begin{aligned}
D_{AB} = & \frac{1}{2}\sum_k p_A(k)\log\Big[\frac{p_A(k)}{p_B(k)}\Big] + \frac{1}{2}\sum_k p_B(k)\log\Big[\frac{p_B(k)}{p_A(k)}\Big] \\
& + \frac{1}{4\langle k\rangle_A}\sum_{kk'} p_A(k)p_A(k')kk'\Pi_A(k,k')\log\Big[\frac{\Pi_A(k,k')}{\Pi_B(k,k')}\Big] \\
& + \frac{1}{4\langle k\rangle_B}\sum_{kk'} p_B(k)p_B(k')kk'\Pi_B(k,k')\log\Big[\frac{\Pi_B(k,k')}{\Pi_A(k,k')}\Big] \\
& + \frac{1}{2}\sum_k p_A(k)k\log\rho_{AB}(k) + \frac{1}{2}\sum_k p_B(k)k\log\rho_{BA}(k) \qquad (1.23)
\end{aligned}
$$

The first line gives the degree statistics contribution to the dissimilarity of networks A and B. The second and third lines reflect wiring details beyond those imposed by the degree sequences. Line four is an interference term, involving quantities $\rho_{AB}(k)$ to be solved from a simple equation that is derived in (Roberts *et al.* 2010). The derivation of (1.23) from (1.22), apart from the interference term, is found in (Annibale *et al.* 2009). In contrast to dissimilarity measures of networks that are based on link overlap, the measure (1.23) is based strictly on macroscopic measures and has a precise information-theoretic basis[4].

1.5 Applications to Protein-Protein Interaction Networks

In contrast to genomic data, the available proteome data are still far from complete and of limited reproducibility (Hart *et al.* 2006, Stumpf *et al.* 2008). It is therefore vital that we understand the origin of the discrepancies between observed PPINs. Here we explore the use of information-theoretic random graph based tools for PPIN characterization and comparison, as described in the preceding sections, to shed light on this problem.

Figure 1.3 gives a table of various PPIN datasets, colour coded according to their experimental method. To get some feeling for these data we show in Figure 1.4 the degree correlations of connected nodes as measured by (1.7) for the bacterial species in our table. There appears to be nontrivial information in the degree correlations, giving rise to diverse patterns for different species. The most closely related bacteria in our table are *H. pylori* and *C. jejuni*, which both belong to the *Campylobacterales* genus, yet this is not reflected in their degree correlations. Similarly, comparing *E. coli* ,*C. jejuni*, *T. pallidum* and *H. pylori*, all belonging to the Proteobacteria Phylum family (the majority of gram-negative bacteria), does not reveal a consistent pattern either. More worryingly, fully consistent degree correlation fingerprints are not even observed for datasets of the same species. This is seen in Figure 1.5 which shows the degree correlations for yeast, the focus of most large-scale PPIN determinations so far, displayed in chronological order of experimental determination.

A hint at a possible explanation emerges if one compares only plots that refer to the same experimental technique. The degree correlation patterns then appear more similar, differing mostly in the strengths of the deviations from the random level, which increase roughly with the time of publication of the dataset. Compare e.g. *S. cerevisiae* II (core) to *S. cerevisiae* XII (both obtained via Y2H), and *S. cerevisiae* VIII to *S. cerevisiae* X (both obtained via AP–MS). The interactions reported in *S. cerevisiae* X were derived from the raw

[4]Link-by-link overlap is not a good measure of the (dis)similarity between two networks, just as the size in bits of a file does not generally give its true information content.

Species	NP	PCG	NI	<k>	kmax	DM	Ref
C.elegans	2528	20176	3864	2.96	99	Y2H	Simonis *et al.* 2008
C.jejuni	1324	1736	11796	17.5	207	Y2H	Parrish *et al.* 2007
H.pylori	724	1587	1403	3.87	55	Y2H	Rain *et al.* 2001
H.sapiens I	1499	21370	2530	3.37	125	Y2H	Rual *et al.* 2005
H.sapiens II	1655	21370	3076	3.71	95	Y2H	Stelzl *et al.* 2005
M.loti	1803	7343	3094	3.43	401	Y2H	Shimoda *et al.* 2008
P.falciparum	1267	5385	2709	4.17	51	Y2H	Lacount *et al.* 2005
S.cerevisiae I	991	6532	948	1.82	24	Y2H	Uetz *et al.* 2000
S.cerevisiae II	787	6532	806	1.91	55	Y2H	Ito *et al.* 2001(core)
S.cerevisiae III	3241	6532	4367	2.69	279	Y2H	Ito *et al.* 2001
S.cerevisiae XII	1544	6532	1809	2.34	86	Y2H	Yu *et al.* 2008
Synechocystis	1903	3725	3100	3.25	51	Y2H	Sato *et al.* 2007
T.pallidum	724	1039	3627	10.01	285	Y2H	Titz *et al.* 2008
E.coli	2457	4246	8664	7.05	641	AP-MS	Arifuzzaman *et al.* 2006
H.sapiens III	2268	21370	6433	5.67	314	AP-MS	Ewing *et al.* 2007
S.cerevisiae IV	1576	6532	3617	4.58	62	AP-MS	Ho *et al.* 2002
S.cerevisiae VI	1358	6532	3220	4.73	53	AP-MS	Gavin *et al.* 2002
S.cerevisiae VIII	2551	6532	21394	16.77	955	AP-MS	Gavin *et al.* 2006
S.cerevisiae IX	2708	6532	7121	5.25	141	AP-MS	Krogan *et al.* 2006
S.cerevisiae X	1630	6532	9089	11.15	127	AP-MS	Collins *et al.* 2007
S.cerevisiae XI	1078	6532	2770	4.7	58	PCA	Tarassov *et al.* 2008
D.melanogaster	7286	14141	25102	6.85	176	DD	Stark *et al.* 2006
H.sapiens IV	9306	21370	35021	7.52	247	DD	Prasad *et al.* 2009
S.cerevisiae V	2617	6532	11855	9.05	118	DI	Von Mering *et al.* 2002
S.cerevisiae VII	1379	6532	2493	3.61	32	DI	Han *et al.* 2004

Yeast-two-Hybrid Affinity Purification-Mass Spectrometry Protein Complementation Assay
Database Datasets Data Integration

Figure 1.3 Table of PPIN datasets corresponding to 11 species (nine eukaryotic, and six bacteria). Abbreviations: NP–Number of Proteins, NI–Number of Interactions, PCG–Number of Protein Coding Genes, AD–Average Degree, Kmax–Maximum Degree, and DM–Data collection Method. Most data were derived from high-throughput Y2H or AP-MS experiments. We added a recent PCA dataset and several consolidated datasets that combine high-throughput experimental data with literature mining. The Ito *et al.* data were divided in a high confidence set (core) and a low confidence set, as suggested by the authors. The Collins *et al.* data consist of the raw purifications in Krogan *et al.* and Gavin *et al.* , but re-analyzed differently. We also included two commonly used yeast datasets: the Han *et al.* network (a consolidated dataset referred to as the 'Filtered Yeast Interactome', consisting of experimentally determined and in silico predicted interactions), and the von Mering *et al.* dataset (assembled from two catalogs of yeast protein complexes, the MIPS and Yeast Protein Database catalogue).

data of two AP-MS datasets (*S. cerevisiae* VIII and *S. cerevisiae* IX), but processed using a different scoring and clustering protocol. The AP-MS datasets generally show stronger degree correlation patterns than the Y2H ones (this is also observed for *H. sapiens* data) although the regions where the main deviations from the random level occur are different.

1.5.1 PPIN Assortativity and Wiring Complexity

To assess the statistical significance of observed differences in degree correlation patterns we measure for the different datasets two quantities that are strongly dependent upon the degree correlations: the assortativity (1.3) and the wiring complexity, i.e. the second term of (1.21). We test the observed values against similar observations in appropriate null models. The latter are graphs generated randomly and with uniform probabilities from the set $G[p]$, all non-directed networks with size and degree distribution identical to those of the networks under study, but without degree correlations. For large N all graphs in $G[p]$ will have $\Pi(k,k') = 1$

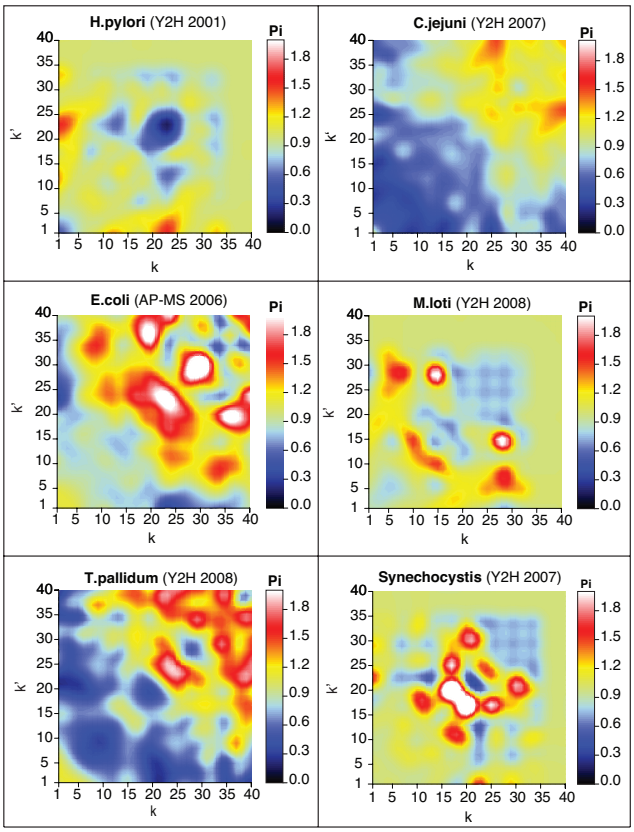


Figure 1.4 Re-scaled degree statistics $\Pi(k,k')$ of connected nodes in bacterial PPINs. Since one would have found $\Pi(k,k')=1$ for all (k,k') if connected nodes had uncorrelated degrees, deviations from light green suggest nontrivial structural features. Figure taken from (Fernandes *et al.* 2010).

for all (k,k'), and zero wiring complexity and assortativity. Any nonzero value reflects finite size effects, possibly complemented by imperfect equilibration during the generation of the null models (the numerical generation of such graphs is the subject of a subsequent section).

In Fig. 1.6 (top) we plot the assortativities of our PPIN datasets (original), together with those of their null models (reshuffled). Most sets have slightly negative assortativity values, i.e. weak preference for interactions between nodes with different degrees. The main deviant from this trend is *S. cerevisiae* X, with a strong positive assortativity. This is consistent with Fig. 1.5, where this dataset indeed exhibits high values of $\Pi(k,k')$ along the main diagonal, signalling a preference for interactions between nodes with similar degrees. The assortativities of the null models are expected to be closer to zero than those of the PPINs. This is indeed true for the majority of cases, and we may therefore conclude that the structures observed for $\Pi(k,k')$ in our PPIN data (as in Figures 1.4 and 1.5) cannot all be attributed to finite size fluctuations, and are hence statistically significant. The wiring complexity per node is the second term in (1.21). It measures topological information contained in

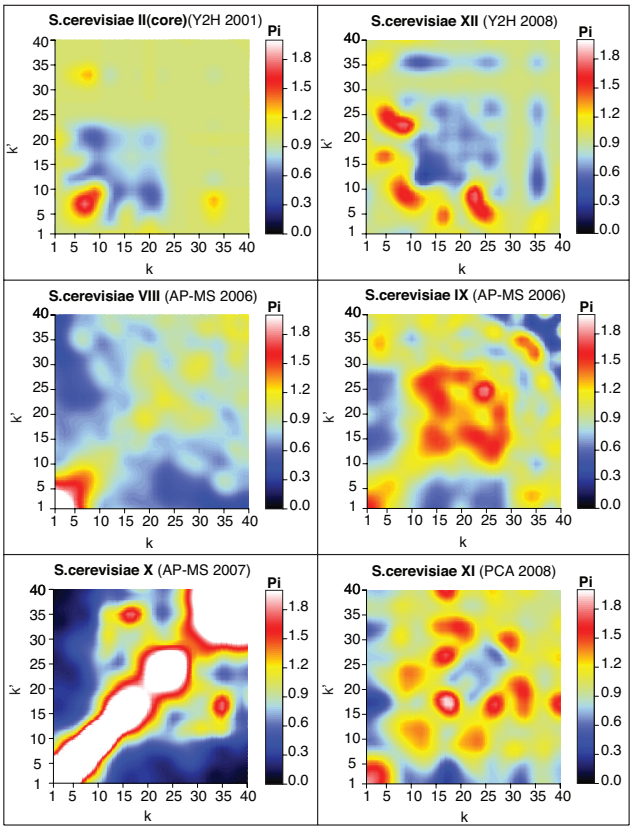


Figure 1.5 Re-scaled degree statistics $\Pi(k,k')$ of connected nodes in yeast PPINs. Since one would have found $\Pi(k,k')=1$ for all (k,k') if connected nodes had uncorrelated degrees, deviations from light green suggest nontrivial structural features. Figure taken from (Fernandes *et al.* 2010).

a network's degree correlations, beyond that contained in the degree distribution alone. However, given the considerable differences between the average connectivities in table 1.3, and the likelihood that these reflect sampling variability (if anything), we choose to measure and plot instead the wiring complexity *per link*, i.e.

$$\tilde{\mathcal{C}}[p,\Pi]_{\rm wiring}=\frac{1}{2}\sum_{kk'}p(k)p(k')kk'\Pi(k,k')\log\Pi(k,k') \tag{1.24}$$

(related to $\mathcal{C}[p,\Pi]_{\rm wiring}$ via division by $\langle k\rangle$). In Fig. 1.6 (bottom) we plot this quantity for our PPIN datasets (original), together with those of the corresponding null models (reshuffled). Interestingly, the AP–MS networks tend to have higher wiring complexities than the Y2H ones. The wiring complexities of the null models are expected to be closer to zero than those of the real PPINs, and this is again borne out by the data. Once more we conclude that the structures observed for $\Pi(k,k')$ in our PPIN data (as in Figures 1.4 and 1.5) cannot be attributed to finite size fluctuations; they are statistically significant.

Figure 1.6 Assortativity (top) and wiring complexity per link (bottom) for the biological PPIN of table 1.3 and their null models. Apart from having size and degree distributions identical to their biological counterparts, the null models are strictly random, and would for large N have zero assortativity and wiring complexity. Figures taken from (Annibale *et al.* 2009, Fernandes *et al.* 2010).

1.5.2 Mapping PPIN Data Biases

We saw that an efficient information-theoretic measure for the dissimilarity between two networks $\mathbf{c}_A$ and $\mathbf{c}_B$ is given by the Jeffrey's divergence between the probability measures of the associated random graph ensembles. If we work at the level of the sets $G[p, \Pi]$, the result is formula (1.23). Had we characterized networks only according to their degree distributions we would have worked with the sets $G[p]$, and would have found only the first line of (1.23). Given the observed assortativies and wiring complexities of our data, relative to those of null models, we take the degree correlations to be significant. We may then study the relations between the different biological networks by calculating their pairwise distances

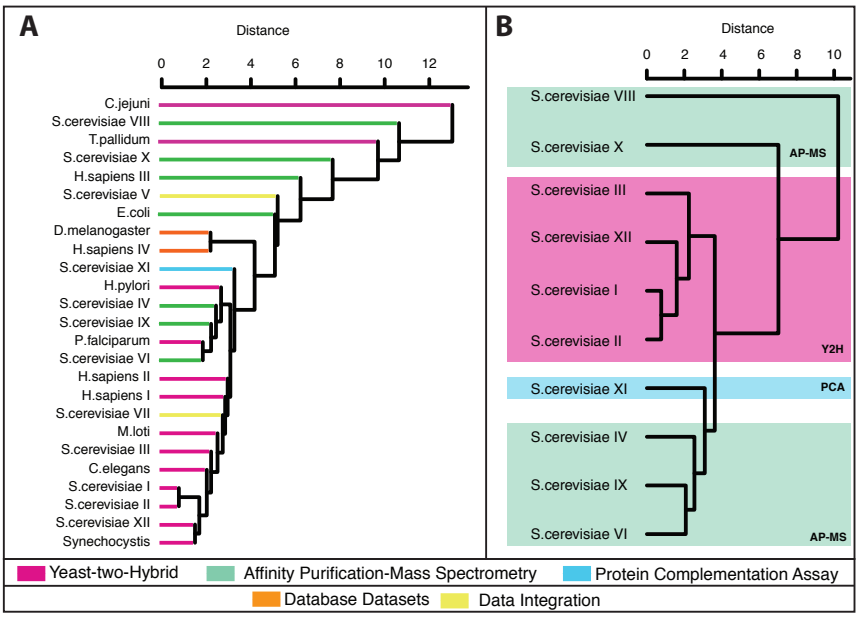


Figure 1.7 Network comparison by dendrogram clustering using the distance measure (1.23). Left (A): dendrogram for the full PPIN collection of table 1.3. Right (B): dendrogram for PPINs of *S. cerevisiae* only. Branch colours indicate the different experimental detection techniques. The integration datasets (*S. cerevisiae* V and *S. cerevisiae* VII were excluded from panel B, since they are based on a variety of techniques. Figure taken from (Fernandes *et al.* 2010).

via (1.23), use the distance table to cluster the datasets, and show the result in the form of a dendrogram. This gives Figure 1.7, which is quite revealing[5]. Those data sets which were most strongly criticized in the past for having worryingly small overlaps, e.g. the Y2H data sets *S. cerevisiae* I versus II and *H. sapiens* I versus II, are now unambiguously found to be topologically similar. However, our collection of PPINs group primarily by detection method; for the presently available PPIN datasets, any biological similarities are overshadowed by methodological biases. This is particularly evident in the central subgroup (central pink leaves in panel A), which clusters almost exclusively Y2H datasets and comprises a wide range of species. The methodological biases are also obvious in the intra-species comparison of *S. cerevisiae* shown in panel B. The largest sub–group distance within this tree is the one between two AP–MS datasets that have been post–processed differently (the top two within the green box). Also, the single PCA network is separated from the AP–MS and Y2H subgroups. We conclude: (i) protein-protein interaction networks of the same species and measured via the same experimental method are statistically similar, and more similar than networks measured via the same method but for different species, and (ii) protein-protein interaction networks measured via the same experimental method cluster together, revealing a bias introduced by the methods that is seen to overrule species-specific information.

[5]If one repeats this exercise using only the first line of (1.23), the resulting dendrogram is similar but less clear. The degree–degree correlations apparently contribute a valuable amount of information to PPIN comparisons.

1.6 Numerical Generation of Tailored Random Graphs

We now turn to the question of how to *generate* numerically graphs from ensembles such as (1.17). It is not difficult to build algorithms that sample the space of all graphs with a given degree sequence; the difficulty lies in generating each graph with the correct probability (Bender and Canfield 1978, Chung and Lu 2002, Stauffer and Barbosa 2005). One popular algorithm (Newman *et al.* 2001) is limited to the case where graphs are to be generated with equal probabilities, i.e. to (1.13) or (1.17) with $\Pi(k,k')=1$. This method cannot generate graphs with degree correlations. A second popular method for generating random graphs with a given degree sequence is 'edge swapping', which involves successions of ergodic graph randomizing moves of a type that leave the degrees $\mathbf{k}$ invariant (Seidel 1973, Taylor 1981). However, we will see that naive accept-all edge swapping can cause sampling biases which render this protocol unsuitable for generating null models. The reason is that the *number* of edge swaps that can be executed is not a constant, it depends on the graph $\mathbf{c}$ at hand.

1.6.1 Generating Random Graphs via Markov Chains

A general and exact method for generating graphs from the set $G[\mathbf{k}]=\{\mathbf{c}\in G|\ \mathbf{k}(\mathbf{c})=\mathbf{k}\}$ randomly, with specified probabilities $p(\mathbf{c})=Z^{-1}\exp[-H(\mathbf{c})]$ was developed in (Coolen *et al.* , 2009). It has the form of a Markov chain, viz. a discrete time stochastic process

$$\forall\mathbf{c}\in G[\mathbf{k}]:\quad p_{t+1}(\mathbf{c})=\sum_{\mathbf{c}'\in G[\mathbf{k}]}W(\mathbf{c}|\mathbf{c}')p_t(\mathbf{c}') \tag{1.25}$$

Here $p_t(\mathbf{c})$ is the probability of observing graph $\mathbf{c}$ at time t in the process, and $W(\mathbf{c}|\mathbf{c}')$ is the one-step transition probability from graph $\mathbf{c}'$ to $\mathbf{c}$. For any set Φ of ergodic[6] reversible elementary moves $F:G[\mathbf{k}]\to G[\mathbf{k}]$ we can choose transition probabilities of the form

$$W(\mathbf{c}|\mathbf{c}')=\sum_{F\in\Phi}q(F|\mathbf{c}')\Big[\delta_{\mathbf{c},F\mathbf{c}'}A(F\mathbf{c}'|\mathbf{c}')+\delta_{\mathbf{c},\mathbf{c}'}[1-A(F\mathbf{c}'|\mathbf{c}')]\Big] \tag{1.26}$$

The interpretation is as follows. At each step a candidate move $F\in\Phi$ is drawn with probability $q(F|\mathbf{c}')$, where $\mathbf{c}'$ denotes the current graph. This move is accepted (and the transition $\mathbf{c}'\to\mathbf{c}=F\mathbf{c}'$ executed) with probability $A(F\mathbf{c}'|\mathbf{c}')\in[0,1]$, which depends on the current graph $\mathbf{c}'$ and on the proposed new graph $F\mathbf{c}'$. If the move is rejected, which happens with probability $1-A(F\mathbf{c}'|\mathbf{c}')$, the system stays in $\mathbf{c}'$. We may always exclude from Φ the identity operation. One can prove that the process (1.25) will converge towards the equilibrium measure $p_\infty(\mathbf{c})=Z^{-1}\exp[-H(\mathbf{c})]$ upon making in (1.26) the choices

$$q(F|\mathbf{c})=I_F(\mathbf{c})/n(\mathbf{c}) \tag{1.27}$$

$$A(\mathbf{c}|\mathbf{c}')=\frac{n(\mathbf{c}')\mathrm{e}^{-\frac{1}{2}[H(\mathbf{c})-H(\mathbf{c}')]}}{n(\mathbf{c}')e^{-\frac{1}{2}[H(\mathbf{c})-H(\mathbf{c}')]}+n(\mathbf{c})\mathrm{e}^{\frac{1}{2}[H(\mathbf{c})-H(\mathbf{c}')]}} \tag{1.28}$$

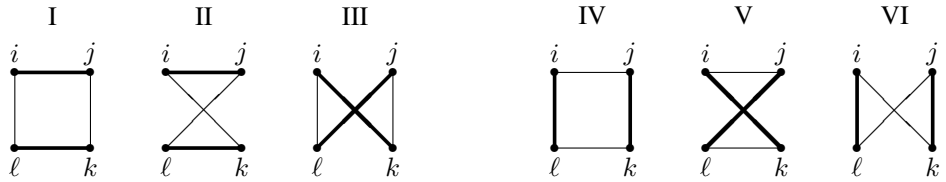
Here $I_F(\mathbf{c})=1$ if the move F can act on graph $\mathbf{c}$, with $I_F(\mathbf{c})=0$ otherwise, and $n(\mathbf{c})$ denotes the total number of moves that can act on a graph $\mathbf{c}$ (the 'mobility' of state $\mathbf{c}$):

$$n(\mathbf{c})=\sum_{F\in\Phi}I_F(\mathbf{c}). \tag{1.29}$$

[6] So we can go from any initial graph $\mathbf{c}\in G[\mathbf{k}]$ to any final graph $\mathbf{c}'\in G[\mathbf{k}]$ by a finite number of moves $F\in\Phi$.

1.6.2 Degree-Constrained Graph Dynamics Based on Edge Swaps

We apply the above result to the case where our moves are edge swaps, which are the simplest graph moves that preserve all node degrees. They act on quadruplets of nodes and their mutual links, so we define the set $Q = \{(i,j,k,\ell) \in \{1,\ldots,N\}^4 |\, i<j<k<\ell\}$ of all ordered node quadruplets. The possible edge swaps to act on (i,j,k,ℓ) are the following, with thick lines indicating existing links and thin lines indicating absent links that will be swapped with the existing ones, and where (IV, V, VI) are the inverses of (I, II, III):



We group the edge swaps into the three pairs (I,IV), (II,V), and (III,VI), and label all three resulting auto-invertible operations for each ordered quadruple (i,j,k,ℓ) by adding a subscript α. Our auto-invertible edge swaps are from now on written as $F_{ijk\ell;\alpha}$, with $i < j < k < \ell$ and $\alpha \in \{1,2,3\}$. We define associated indicator functions $I_{ijk\ell;\alpha}(\mathbf{c}) \in \{0,1\}$ that detect whether (1) or not (0) the edge swap $F_{ijk\ell;\alpha}$ can act on state $\mathbf{c}$, so

$$I_{ijk\ell;1}(\mathbf{c}) = c_{ij}c_{k\ell}(1-c_{i\ell})(1-c_{jk}) + (1-c_{ij})(1-c_{k\ell})c_{i\ell}c_{jk} \tag{1.30}$$

$$I_{ijk\ell;2}(\mathbf{c}) = c_{ij}c_{k\ell}(1-c_{ik})(1-c_{j\ell}) + (1-c_{ij})(1-c_{k\ell})c_{ik}c_{j\ell} \tag{1.31}$$

$$I_{ijk\ell;3}(\mathbf{c}) = c_{ik}c_{j\ell}(1-c_{i\ell})(1-c_{jk}) + (1-c_{ik})(1-c_{j\ell})c_{i\ell}c_{jk} \tag{1.32}$$

If $F_{ijk\ell;\alpha}$ can indeed act, i.e. if $I_{ijk\ell;\alpha}(\mathbf{c}) = 1$, this edge swap will operate as follows:

$$F_{ijk\ell;\alpha}(\mathbf{c})_{qr} = 1 - c_{qr} \quad \text{for } (q,r) \in \mathcal{S}_{ijk\ell;\alpha} \tag{1.33}$$

$$F_{ijk\ell;\alpha}(\mathbf{c})_{qr} \quad = c_{qr} \quad \text{for } (q,r) \notin \mathcal{S}_{ijk\ell;\alpha} \tag{1.34}$$

where

$$\mathcal{S}_{ijk\ell;1} = \{(i,j),(k,\ell),(i,\ell),(j,k)\} \qquad \mathcal{S}_{ijk\ell;2} = \{(i,j),(k,\ell),(i,k),(j,\ell)\} \tag{1.35}$$

$$\mathcal{S}_{ijk\ell;3} = \{(i,k),(j,\ell),(i,\ell),(j,k)\} \tag{1.36}$$

Insertion of these definitions into the general recipe (1.26,1.27,1.28) then gives

$$W(\mathbf{c}|\mathbf{c}') = \sum_{i<j<k<\ell} \sum_{\alpha\leq 3} \frac{I_{ijk\ell;\alpha}(\mathbf{c}')}{n(\mathbf{c}')} \times \Big[\frac{\delta_{\mathbf{c},F_{ijk\ell;\alpha}\mathbf{c}'}\mathrm{e}^{-\frac{1}{2}[E(F_{ijk\ell;\alpha}\mathbf{c}')-E(\mathbf{c}')]} + \delta_{\mathbf{c},\mathbf{c}'}\mathrm{e}^{\frac{1}{2}[E(F_{ijk\ell;\alpha}\mathbf{c}')-E(\mathbf{c}')]}}{\mathrm{e}^{-\frac{1}{2}[E(F_{ijk\ell;\alpha}\mathbf{c}')-E(\mathbf{c}')]} + \mathrm{e}^{\frac{1}{2}[E(F_{ijk\ell;\alpha}\mathbf{c}')-E(\mathbf{c}')]}}\Big] \tag{1.37}$$

with $E(\mathbf{c}) = H(\mathbf{c}) + \log n(\mathbf{c})$. The graph dynamics algorithm described by (1.37) is the following. Given an instantaneous graph $\mathbf{c}'$: (i) pick uniformly at random a triplet (i,j,k,ℓ) of sites, (ii) if at least one of the three edge swaps $\mathbf{c}' \to F_{ijk\ell;\alpha}(\mathbf{c}')$ is possible, select one of these uniformly at random and execute it with an acceptance probability

$$A(\mathbf{c}|\mathbf{c}') = \big[1 + \mathrm{e}^{E(F_{ijk\ell;\alpha}\mathbf{c}')-E(\mathbf{c}')}\big]^{-1} \tag{1.38}$$

then return to (i). For this Markov chain recipe to be practical we finally need a formula for the mobility $n(\mathbf{c})$ of a graph. This could be calculated (Coolen *et al.* , 2009), giving[7]:

$$n(\mathbf{c}) = \frac{1}{4}\big(\sum_i k_i\big)^2 + \frac{1}{4}\sum_i k_i - \frac{1}{2}\sum_i k_i^2 - \frac{1}{2}\sum_{ij} k_i c_{ij} k_j + \frac{1}{4}\mathrm{Tr}(\mathbf{c}^4) + \frac{1}{2}\mathrm{Tr}(\mathbf{c}^3) \quad (1.39)$$

Naive 'accept-all' edge swapping would correspond to choosing $E(\mathbf{c}) = 0$ in (1.37), and upon equilibration it would give the *biased* graph sampling probabilities $p_\infty(\mathbf{c}) = n(\mathbf{c})/\sum_{\mathbf{c}'} n(\mathbf{c}')$. The graph mobility is seen to act as an entropic force, which can only be neglected if (1.39) is dominated by its first three terms; it was shown that a sufficient condition for this to be the case is $\langle k^2\rangle k_{\max}/\langle k\rangle^2 \ll N$. For networks with narrow degree sequences this condition would hold, and naive edge swapping would be acceptable. However, one has to be careful with scale-free degree sequences, where both $\langle k^2\rangle$ and $k_{\max}$ diverge as $N \to \infty$.

1.6.3 Numerical Examples

It is easy to construct example degree distributions where taking into account the entropic effects caused by nontrivial graph mobilities is vital in order to generate correct graph sampling probabilities. Here we show an example of the Markov chain (1.37) generating upon equilibration[8] graphs with controlled degree correlation structures of the form

$$\Pi(k,k') = \frac{(k-k')^2}{[\beta_1 - \beta_2 k + \beta_3 k^2][\beta_1 - \beta_2 k' + \beta_3 k'^2]} \quad (1.40)$$

(with the parameters β_i following from 1.4). An initial graph $\mathbf{c}_0$ was constructed with a non-Poissonian degree distribution and trivial relative degree correlations $\Pi(k,k'|\mathbf{c}_0) \approx 1$, corresponding to the flat ensemble (1.13); see Figure 1.8. After iterating until equilibrium the Markov chain (1.37) with move acceptance rates tailored to approaching (1.17) as an equilibrium measure, one finds indeed values for the degree correlations in very good agreement with their target values; (see the bottom panels of Figure 1.8).

1.7 Discussion

In this chapter we have discussed the fruitful connection between biological signalling networks and the theory of random graph ensembles with tailored structural features. We focused on two aspects of this connection: how random graph ensembles can be used to generate rigorous information-theoretic formulae with which to quantify the complexities and (dis)similarities of observed networks, and how to generate numerically tailored random graphs with controlled macroscopic structural properties, to serve e.g. as null models in hypothesis testing. We limited ourselves here to non-directed graphs, in view of space limitations and since these have so far been the focus of most research papers; similar analyses can be (and are being) undertaken for directed ones. The quantitative study of cellular signalling networks is still in its infancy, and as our mathematical tools continue to improve one can envisage many future research directions. These include e.g. the

[7] Here $\mathrm{Tr}\mathbf{A} = \sum_i A_{ii}$.

[8] The algorithm ran for a duration of 75,000 accepted edge swaps, and measurements of Hamming distances confirmed that with this duration the dynamics achieved maximum distance between initial and final graphs.

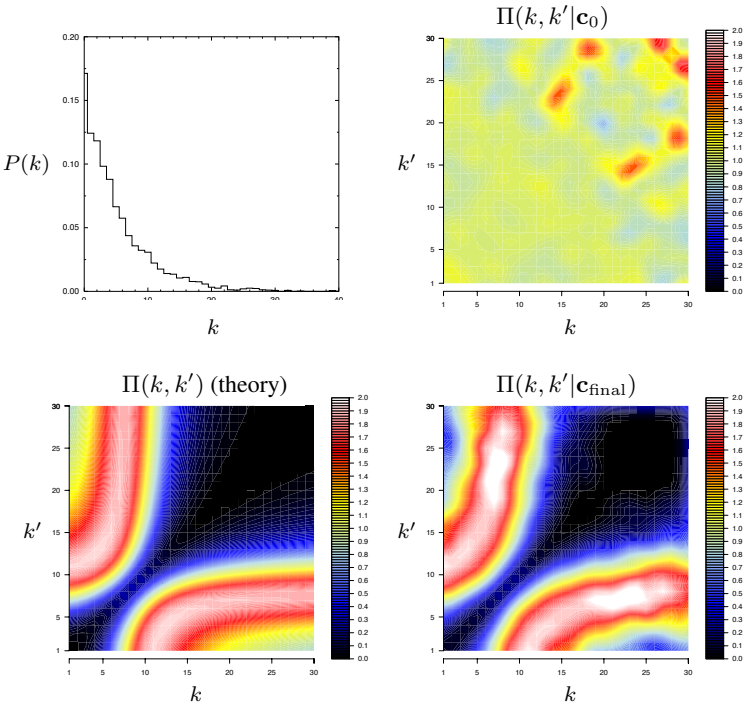


Figure 1.8 Results of canonical edge-swap Markov chain dynamics tailored to generating random graphs with the non-uniform measure (1.17). Top left: degree distribution of the (randomly generated) initial graph $\mathbf{c}_0$, with $N = 4000$ and $\langle k\rangle = 5$. Top right: relative degree correlations $\Pi(k,k'|\mathbf{c}_0)$ of the initial graph. Bottom left: the target relative degree correlations (1.40) chosen in (1.17). Bottom right: colour plot of the relative degree correlations $\Pi(k,k'|\mathbf{c}_{\rm final})$ in the final graph $\mathbf{c}_{\rm final}$, measured after 75,000 accepted moves of the Markov chain (1.37). Figure taken from (Coolen *et al.* 2009).

(biased) network sampling problem, where one could perhaps use the new information-theoretic formulae to predict unobserved nodes, and the study of integrated signalling networks that combine transcription and protein-protein interaction information. At the mathematical level the main new challenge to be confronted is to develop tools similar to the ones discussed in this chapter for measures of network structure that involve the statistics of loops. If observables include loop counters one cannot simply extend the existing mathematical techniques (sums over all graphs can no longer be made to factorize by existing manipulations); radically new ideas are required.

References

[1] Albert R and Barabási AL 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–96.

[2] Annibale A, Coolen ACC, Fernandes LP, Fraternali F and Kleinjung J 2009 Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure. *J. Phys. A: Math. Theor.* **42**, 485001 (25 pp).

[3] Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, *et al.* 2006 Large-scale identification of protein-protein interaction of *Escherichia* coli K-12. *Genome Res.* **16**, 686–691.

[4] Barabási AL and Albert R 1999 Emergence of scaling in random networks. *Science* **286**, 509–512.

[5] Bender E and Canfield E 1978 The asymptotic of labelled graphs with given degree sequences. *J. Comb. Theory, Ser. A* **24**, 296–307.

[6] Bianconi G, Coolen ACC and Pérez Vicente CJ 2008 Entropies of complex networks with hierarchically constrained topologies. *Phys. Rev. E* **78**, 016114 (11 pp).

[7] Chung F and Lu L 2002 The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 15879–15882.

[8] Collins SRR, Kemmeren P, Chu X, Greenblatt JFF, Spencer F, *et al.* 2007 Towards a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell Proteomics* **6**, 439–450.

[9] Coolen ACC, De Martino A and Annibale A 2009 Constrained Markovian dynamics of random graphs. *J. Stat. Phys.* **136**, 1035–1067.

[10] Cover TM and Thomas JA 1991 *Elements of information theory*. Wiley.

[11] Dorogovtsev SN, Goltsev AV and Mendes JFF 2008 Critical phenomena in complex networks. *Rev. Mod. Phys.* **80**, 1275–1335.

[12] Erdös P and Rényi A 1959 On random graphs I. *Publ. Math.* **6**, 290–297.

[13] Ewing RM, Chu P, Elisma F, Li H, Taylor P, *et al.* 2007 Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89.

[14] Fernandes LP, Annibale A, Kleinjung J, Coolen ACC and Fraternali F 2010 Protein networks reveal detection bias and species consistency when analysed by information-theoretic methods. *PLoS ONE* **5**, e12083.

[15] Foster JG, Foster DV, Grassberger P and Paczuski M 2007 Link and subgraph likelihoods in random undirected networks with fixed and partially fixed degree sequences. *Phys. Rev. E* **76**, 046112 (12 pp).

[16] Garlaschelli D and Loffredo MI 2008 Maximum likelihood: extracting unbiased information from complex networks. *Phys. Rev. E* **78**, 015101 (4 pp).

[17] Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.

[18] Gavin AC, Aloy P, Grandi P, Krause R, Bösche M, *et al.* 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636.

[19] Hakes L, Pinney JW, Robertson DL and Lovell SC 2008 Protein-protein interaction networks and biology-what's the connection? *Nature Biotechnology* **26**, 69–72.

[20] Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, *et al.* 2004 Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93.

[21] Han JDJ, Dupuy D, Bertin N, Cusick ME ands Vidal M 2005 Effect of sampling on topology predictors of protein-protein interaction networks. *Nature Biotechnology* **23**, 839–844.

[22] Hart GT, Ramani AK, and Marcotte EM 2006 How complete are current yeast and human protein-interaction networks? *Genome Biology* **7**, 120 (9 pp).

[23] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, *et al.* 2002 Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.

[24] Holme P and Zhao J 2007 Exploring the assortativity-clustering space of a network's degree sequence. *Phys. Rev. E* **75**, 046111 (10 pp).

[25] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, *et al.* 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574.

[26] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, *et al.* 2006 Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature* **440**, 637–643.

[27] Kühn R 2008 Spectra of sparse random matrices. *J. Phys. A: Math. Theor.* **41**, 295002 (21 pp).

[28] Lacount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, *et al.* 2005 A protein-protein interaction network of the malaria parasite *plasmodium falciparum*. *Nature* **438**, 103–107.

[29] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827.

[30] Mohar B 1991 The Laplacian spectrum of graphs. In *Graph theory, combinatorics, and applications* **2** (ed. Alavi Y, Chartrand G, Oellermann OR, and Schwenk AJ), 871-898, Wiley.

[31] Molloy M and Reed B 1995 A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–180.

[32] Newman MEJ, Strogatz SH and Watts DJ 2001 Random graphs with arbitrary degree distribution and their applications. *Phys. Rev. E* **64**, 026118 (17 pp).

[33] Newman MEJ 2002 Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (4 pp).

[34] Newman MEJ 2003 *Handbook of graphs and networks: from the genome to the internet*. (ed. Bornholdt S and Schuster HG). Wiley-VCH.

[35] Park J and Newman MEJ 2004 Statistical mechanics of networks. *Phys. Rev. E* **70**, 066117 (13 pp).

[36] Parrish JR, Yu J, Liu G, Hines JA, Chan JE, *et al.* 2007 A proteome-wide protein-protein interaction map for *campylobacter jejuni*. *Genome Biology* **8**, R131.

[37] Pérez-Vicente CJ and Coolen ACC 2008 Spin models on random graphs with controlled topologies beyond degree constraints. *J. Phys. A: Math. Theor.* **41**, 255003 (24 pp).

[38] Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, *et al.* 2009 Human protein reference database 2009 update. *Nucleic Acids Res.* **37**, D767-D772.

[39] Roberts ES, Coolen ACC and Schlitt T 2010 Tailored graph ensembles as proxies or null models for real networks II: results on directed graphs. submitted to *J. Phys. A: Math. Theor.*

[40] Rogers T, Pérez Vicente C, Takeda K and Pérez Castillo I 2010 Spectral density of random graphs with topological constraints. *J. Phys. A: Math. Theor.* **43**, 195002 (20 pp).

[41] Rogers T, Pérez Vicente C, Takeda K and Pérez Castillo I 2010 Spectral density of random graphs with topological constraints. *J. Phys. A: Math. Theor.* **43**, 195002 (20 pp).

[42] Rual JFF, Venkatesan K, Hao T, Kishikawa TH, Dricot A, *et al.* 2005 Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178.

[43] Sato S, Shimoda Y, Muraki A, Kohara M, Nakamura Y, *et al.* 2007 A large-scale protein-protein interaction analysis in *synechocystis sp. PCC6803*. *DNA Res.* **14**, 207–216.

[44] Seidel JJ 1973 A survey of two-graphs. In *Colloquio Internazionale sulle Teorie Combinatorie* (Atti dei Convegni Lincei, No. 17. Accad. Naz. Lincei, Rome), Tomo I, 481–511.

[45] Shimoda Y, Shinpo S, Kohara M, Nakamura Y, Tabata S, *et al.* 2008 A large scale analysis of protein-protein interactions in the nitrogen-fixing bacterium *mesorhizobium loti*. *DNA Res.* **15**, 13–23.

[46] Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, *et al.* 2008 Empirically controlled mapping of the *caenorhabditis elegans* protein-protein interactome network. *Nature Methods* **6**, 47–54.

[47] Skantzos NS 2005 *unpublished research report*

[48] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, *et al.* 2006 Biogrid: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539.

[49] Stauffer AO and Barbosa VC 2005 A study of the edge switching Markov-Chain method for the generation of random graphs. *arXiv:0512105*.

[50] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, *et al.* 2005 A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968.

[51] Stumpf MPH and Wiuf C 2005 Sampling properties of random graphs: the degree distribution. *Phys. Rev. E* **72**, 036118 (7 pp).

[52] Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, and Wiuf C 2008 Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959-6964.

[53] Taylor R 1981 Constrained switchings in graphs. In *Combinatorial Mathematics VIII* **884**, (ed. McAvaney KL), 314–336. Springer Lect. Notes Math.

[54] Tarassov K, Messier V, Landry CR, Radinovic S, Molina MM, *et al.* 2008 An in vivo map of the yeast protein interactome. *Science* **320**, 1465–1470.

[55] Titz B, Rajagopala SV, Goll J, Häuser R, Mckevitt MT, *et al.* 2008 The binary protein interactome of *treponema pallidum* – the syphilis spirochete. *PLoS ONE* **3**, e2292.

[56] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, *et al.* 2000 A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature* **403**, 623–627.

[57] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, *et al.* 2002 Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403.

[58] Watts DJ and Strogatz SH 1998 Collective dynamics of śmall worldńetworks. *Nature* **393**, 440–442.

[59] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, *et al.* 2008 High-quality binary protein-protein interaction map of the yeast interactome network. *Science* **322**, 104–110.