IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **44** (2011) 275002 (20pp) doi:10.1088/1751-8113/44/27/275002

Tailored graph ensembles as proxies or null models for real networks II: results on directed graphs

E S Roberts[1,2], T Schlitt[3] and A C C Coolen[1,2,4]

[1] Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK
[2] Randall Division of Cell and Molecular Biophysics, King's College London, New Hunts House, London SE1 1UL, UK
[3] Department of Medical and Molecular Genetics, King's College London School of Medicine, 8th floor Guy's Tower, London SE1 9RT, UK
[4] London Institute for Mathematical Sciences, 35a South St, Mayfair, London W1K 2XF, UK

E-mail: ekaterina.roberts@kcl.ac.uk, ton.coolen@kcl.ac.uk and thomas.schlitt@kcl.ac.uk

Received 31 January 2011
Published 3 June 2011
Online at stacks.iop.org/JPhysA/44/275002

Abstract
We generate new mathematical tools with which to quantify the macroscopic topological structure of large directed networks. This is achieved via a statistical mechanical analysis of constrained maximum entropy ensembles of directed random graphs with prescribed joint distributions for in- and out-degrees and prescribed degree–degree correlation functions. We calculate exact and explicit formulae for the leading orders in the system size of the Shannon entropies and complexities of these ensembles, and for information-theoretic distances. The results are applied to data on gene regulation networks.

PACS numbers: 87.18.Vf, 89.70.Cf, 89.75.Fb, 64.60.aq

1. Introduction

There is a great demand, especially in cellular biology, for precise mathematical tools with which to quantify topological structure in large observed networks. Such tools can be used to: compare networks; distinguish between meaningful and random structural features; and, to define and generate tailored random graphs as null models or network proxies. In a previous paper [1], it was shown how a specific family of tailored random graph ensembles, with controlled degree distributions and controlled degree–degree correlation functions, is well suited for generating such tools. The authors of [1] applied techniques from statistical mechanics to calculate explicit formulae for the leading orders in the systems size of the Shannon entropy per node for these tailored graph ensembles, and related quantities such as complexity and information-theoretic distances. Subsequent papers were devoted

to the numerical generation of graphs [2] from the proposed ensemble families and the application in cellular biology of the resulting mathematical tools [3]. For an overview see e.g. [4]. The main limitation of [1] was that it only dealt with nondirected networks and graphs. In this paper we take the next step and develop the corresponding theory for directed ones.

Extending the methods in [1] to directed networks will enable their application to important new problems especially in cellular biology. Other applications could include the analysis and control of communication and computation networks. For example, to understand the processes driving a cell it is necessary to go beyond studying individual genes; one needs to study their interactions. Information on how genes interact within the cell is commonly represented by a directed graph: the gene regulation network. High-throughput methods have generated a wealth of data on gene regulation. We now need powerful mathematical tools to analyse these data. By focussing on which properties are the most important to the structure of the biological signalling network, we can envisage being able to postulate mechanisms for how the network evolved and came to fulfil its function, and build better models for such networks. Evaluating the fit of a network model to network data is often seen as a formidable computational challenge [5], which is usually overcome by looking at fit based on comparing network properties. Our approach gives a rigorous quantitative method for prioritising network properties; this is important as different properties might promote different potential models.

The use of statistical mechanics to quantify the information content of network structure is well established; see e.g. [1, 6–8]. Most work so far has focused on undirected networks. The network properties most frequently studied are degree distributions, clustering coefficients, assortativities and path length statistics. There has also been research on occurrences of motifs and subgraphs, motivated by the idea that if a network favours specific local topological patterns then these might reflect common local processes. A particular benefit of the approach followed here and in [1] is the compact and explicit nature of the final formulae. Although their derivations are involved in places, the final results are compact. They take easily measured topological observables as input, avoid the need for numerical simulations or approximations, and are easy and efficient to use as our (biological) datasets grow. We therefore imagine that this line of research will continue to develop, by adding further macroscopic network observables, beyond degree statistics and degree correlation functions. Each addition will make the method more powerful and useful.

The specific quantities calculated in this paper are: the Shannon entropy and complexity of directed graph ensembles with controlled degree distributions; the Shannon entropy and complexity of directed graph ensembles with controlled degree distributions and controlled degree–degree correlation functions; and, the symmetrised Kullback–Leibler distance between pairs of such ensembles. For each of these we calculate the leading orders in the network size, expressed in terms of the controlled degree distributions and degree–degree correlation functions of the ensembles concerned. We illustrate the use of our results in section 5 with applications to experimental data on gene regulation networks.

We adopt the following notation conventions. Each directed graph with N nodes is defined by a matrix $\boldsymbol{c} = \{c_{ij}\}$, with entries $c_{ij} \in \{0, 1\}$ indicating whether ($c_{ij} = 1$) or not ($c_{ij} = 0$) there is a directed arc from node j to node i. For each node i we define the so-called in- and out-degrees, viz. $k_i^{\rm out}(\boldsymbol{c}) = \sum_j c_{ji}$ and $k_i^{\rm in}(\boldsymbol{c}) = \sum_j c_{ij}$; in nondirected graphs such as in [1] one would have had $k_i^{\rm in}(\boldsymbol{c}) = k_i^{\rm out}(\boldsymbol{c})$ for all i. We write the pair of degrees at a site i as $\vec{k}_i(\boldsymbol{c}) = (k_i^{\rm in}(\boldsymbol{c}), k_i^{\rm out}(\boldsymbol{c}))$. Boldface letters will represent ordered sets with N elements, such as $\boldsymbol{k}^{\rm in} = (k_1^{\rm in}, \ldots, k_N^{\rm in})$, or $\boldsymbol{k}^{\rm in}(\boldsymbol{c}) = (k_1^{\rm in}(\boldsymbol{c}), \ldots, k_N^{\rm in}(\boldsymbol{c}))$.

2. Directed graphs with controlled in- and out-degree distributions

Here we calculate the Shannon entropy of an ensemble of directed random graphs constrained by a common joint distribution of in- and out-degrees. Via suitable adaptations of the methods developed for nondirected networks, we achieve a standard path-integral form to which we can apply the method of steepest descent. This leads to an elegant analytical expression for the entropy of the ensemble in the leading orders in N. The key term takes the form of a Kullback–Leibler distance between the imposed joint degree distribution and the Poissonnian one that would have been found upon generating directed arcs independently.

2.1. Definition of the problem

We consider an ensemble of directed random graphs, where degree pairs $\vec{k}_i = \left(k_i^{\rm in}, k_i^{\rm out}\right)$ are for each node i drawn independently from a specified joint degree distribution $p(\vec{k})$:

$$p(\boldsymbol{c}) = \sum_{\vec{k}_1 \ldots \vec{k}_N} \Big[\prod_i p(\vec{k}_i) \Big] p(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N) \tag{2.1}$$

$$p(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N) = \frac{\prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}}{Z(\vec{k}_1 \ldots \vec{k}_N)}, \qquad Z(\vec{k}_1 \ldots \vec{k}_N) = \sum_{\boldsymbol{c}} \prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}. \tag{2.2}$$

For this ensemble we want to find the Shannon entropy per node $S = -N^{-1} \sum_{\boldsymbol{c}} p(\boldsymbol{c}) \log p(\boldsymbol{c})$, which informs us about the effective number $\mathcal{N} = \exp(NS)$ of graphs in the ensemble and the complexity of directed graphs with the imposed degree statistics $p(\vec{k})$. Upon substituting (2.2) into the entropy formula, and after some simple manipulations and use of the law of large numbers, one finds that the entropy per node takes the form

$$S = \frac{1}{N} \sum_{\vec{k}_1 \ldots \vec{k}_N} \Big[\prod_i p(\vec{k}_i) \Big] \log Z(\vec{k}_1 \ldots \vec{k}_N) - \sum_{\vec{k}} p(\vec{k}) \log p(\vec{k}) + \epsilon_N, \tag{2.3}$$

where $\epsilon_N \to 0$ as $N \to \infty$. To make the first term in this expression more tractable, we transform $Z(\vec{k}_1 \ldots \vec{k}_N)$ into an average involving an alternative measure. If we denote the average degree by $\bar{k} = N^{-1} \sum_i k_i^{\rm in} = N^{-1} \sum_i k_i^{\rm out}$, we may define the measure

$$\begin{aligned} w(\boldsymbol{c}|\bar{k}) &= \prod_{ij} \left[\frac{\bar{k}}{N} \delta_{c_{ij},1} + \left(1 - \frac{\bar{k}}{N}\right) \delta_{c_{ij},0} \right] \\ &= \left[1 - \frac{\bar{k}}{N}\right]^{N(N-1)} \left[\frac{\bar{k}/N}{1-\bar{k}/N}\right]^{N\bar{k}(\boldsymbol{c})} \equiv W(\bar{k}, \bar{k}(\boldsymbol{c})). \end{aligned} \tag{2.4}$$

Since this measure depends on the graph $\boldsymbol{c}$ via $\bar{k}(\boldsymbol{c})$ only, we can write the partition function $Z(\vec{k}_1 \ldots \vec{k}_N)$ in terms of an average over the measure (2.4), viz.

$$Z(\vec{k}_1 \ldots \vec{k}_N) = \frac{1}{W(\bar{k}, \bar{k})} \sum_{\boldsymbol{c}} w(\boldsymbol{c}|\bar{k}) \prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}. \tag{2.5}$$

Introducing the notation $\langle f(\boldsymbol{c}) \rangle_\kappa = \sum_{\boldsymbol{c}} w(\boldsymbol{c}|\kappa) f(\boldsymbol{c})$ to represent averages over the measure (2.4) with average connectivity κ, the entropy per node can be written as

$$\begin{aligned} S = &\frac{1}{N} \sum_{\vec{k}_1 \ldots \vec{k}_N} \Big[\prod_i p(\vec{k}_i) \Big] \log \Big\langle \prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})} \Big\rangle_{\bar{k}} - \sum_{\vec{k}} p(\vec{k}) \log p(\vec{k}) \\ &- \frac{1}{N} \sum_{\vec{k}_1 \ldots \vec{k}_N} \Big[\prod_i p(\vec{k}_i) \Big] \log \left[\left[1 - \frac{\bar{k}}{N}\right]^{N(N-1)} \left[\frac{\bar{k}/N}{1-\bar{k}/N}\right]^{N\bar{k}} \right] + \epsilon_N \end{aligned}$$

$$= \frac{1}{N}\sum_{\vec{k}_1\ldots\vec{k}_N}\Big[\prod_i p(\vec{k}_i)\Big]\log\Big\langle\prod_i \delta_{\vec{k}_i,\vec{k}_i(c)}\Big\rangle_{\vec{k}} - \sum_{\vec{k}} p(\vec{k})\log p(\vec{k})$$
$$+ \langle k\rangle\Big[\log(N/\langle k\rangle)+1\Big] + \varepsilon_N \tag{2.6}$$

with $\lim_{N\to\infty}\varepsilon_N = 0$, and with $\langle k\rangle = \sum_{\vec{k}} k^{\rm in}p(\vec{k}) = \sum_{\vec{k}} k^{\rm out}p(\vec{k})$. All the complexity of the problem is thus contained in the first term of (2.6):

$$\phi = \frac{1}{N}\sum_{\vec{k}_1\ldots\vec{k}_N}\Big[\prod_i p(\vec{k}_i)\Big]\log\Big\langle\prod_i \delta_{\vec{k}_i,\vec{k}_i(c)}\Big\rangle_{\vec{k}}. \tag{2.7}$$

2.2. Entropy evaluation

Using Fourier representations of the Kronecker deltas in (2.7) and some straightforward manipulations brings us to

$$\phi = \frac{1}{N}\sum_{\vec{k}_1\ldots\vec{k}_N}\left[\prod_i p(\vec{k}_i)\right]\log\int_{-\pi}^{\pi}\prod_i\left[\frac{{\rm d}\omega_i\,{\rm d}\psi_i}{4\pi^2}{\rm e}^{{\rm i}[\omega_i k_i^{\rm in}+\psi_i k_i^{\rm out}]}\right]L(\boldsymbol{\omega},\boldsymbol{\psi}) \tag{2.8}$$

$$L(\boldsymbol{\omega},\boldsymbol{\psi}) = \exp\left[\bar{k}N\left(\frac{1}{N}\sum_i {\rm e}^{-{\rm i}\omega_i}\right)\left(\frac{1}{N}\sum_j {\rm e}^{-{\rm i}\psi_j}\right) - \bar{k}N + \mathcal{O}(N^0)\right]. \tag{2.9}$$

Introducing the quantities $R(\boldsymbol{\omega}) = N^{-1}\sum_i {\rm e}^{-{\rm i}\omega_i}$ and $S(\boldsymbol{\psi}) = N^{-1}\sum_i {\rm e}^{-{\rm i}\psi_i}$, and inserting $\int {\rm d}R\,{\rm d}S\delta[R-R(\boldsymbol{\omega})]\delta[S-S(\boldsymbol{\psi})]$ with δ-functions written in integral form, allows us to write

$$L(\boldsymbol{\omega},\boldsymbol{\psi}) = \int\frac{{\rm d}R\,{\rm d}\hat{R}\,{\rm d}S\,{\rm d}\hat{S}}{4\pi^2/N^2}{\rm e}^{N[{\rm i}(\hat{R}R+\hat{S}S)+\bar{k}(RS-1)]+\mathcal{O}(N^0)}\prod_i {\rm e}^{-{\rm i}[\hat{R}{\rm e}^{-{\rm i}\omega_i}+\hat{S}{\rm e}^{-{\rm i}\psi_i}]}. \tag{2.10}$$

Substituting this back into ϕ, using the law of large numbers, then gives

$$\phi = \frac{1}{N}\sum_{\vec{k}_1\ldots\vec{k}_N}\Big[\prod_i p(\vec{k}_i)\Big]\log\int{\rm d}R\,{\rm d}\hat{R}\,{\rm d}S\,{\rm d}\hat{S}\,{\rm e}^{N\Psi(R,\hat{R},S,\hat{S})+\mathcal{O}(\log N)} \tag{2.11}$$

where

$$\Psi(R,\hat{R},S,\hat{S}) = {\rm i}(\hat{R}R+\hat{S}S) + \bar{k}(RS-1) + \sum_{k^{\rm in}} p(k^{\rm in})\log\int_{-\pi}^{\pi}\frac{{\rm d}\omega}{2\pi}{\rm e}^{{\rm i}[\omega k^{\rm in}-\hat{R}{\rm e}^{-{\rm i}\omega}]}$$
$$+ \sum_{k^{\rm out}} p(k^{\rm out})\log\int_{-\pi}^{\pi}\frac{{\rm d}\psi}{2\pi}{\rm e}^{{\rm i}[\psi k^{\rm out}-\hat{S}{\rm e}^{-{\rm i}\psi}]}. \tag{2.12}$$

The average in (2.11) over degree sequences is now obsolete since the argument depends in leading order in N on their distribution only, and (2.11) can be evaluated by steepest descent:

$$\lim_{N\to\infty}\phi = {\rm extr}_{R,\hat{R},S,\hat{S}}\Psi(R,\hat{R},S,\hat{S}]. \tag{2.13}$$

We can simplify Ψ by doing the remaining integrals, using

$$\int_{-\pi}^{\pi}\frac{{\rm d}\omega}{2\pi}{\rm e}^{{\rm i}[\omega k - A{\rm e}^{-{\rm i}\omega}]} = \sum_{m\geqslant 0}\frac{(-{\rm i}A)^m}{m!}\int_{-\pi}^{\pi}\frac{{\rm d}\omega}{2\pi}{\rm e}^{{\rm i}\omega(k-m)} = \frac{(-{\rm i}A)^k}{k!}. \tag{2.14}$$

Hence

$$\Psi(R,\hat{R},S,\hat{S}) = {\rm i}(\hat{R}R+\hat{S}S) + \bar{k}(RS-1) + \sum_{k^{\rm in}} p(k^{\rm in})\log[(-{\rm i}\hat{R})^{k^{\rm in}}/k^{\rm in}!]$$
$$+ \sum_{k^{\rm out}} p(k^{\rm out})\log[(-{\rm i}\hat{S})^{k^{\rm out}}/k^{\rm out}!]. \tag{2.15}$$

Differentiation of Ψ gives the following saddle-point equations:

$$-\mathrm{i}\hat{R} = \bar{k}S, \qquad -\mathrm{i}\hat{S} = \bar{k}R \tag{2.16}$$

$$\mathrm{i}R\hat{R} + \bar{k} = 0, \mathrm{i}S\hat{S} + \bar{k} = 0. \tag{2.17}$$

We conclude that $RS = 1$, and hence at the saddle-point we have

$$\Psi(R, \hat{R}, S, \hat{S}) = \sum_{k^{\mathrm{in}}} p(k^{\mathrm{in}}) \log \pi_{\bar{k}}(k^{\mathrm{in}}) + \sum_{k^{\mathrm{out}}} p(k^{\mathrm{out}}) \log \pi_{\bar{k}}(k^{\mathrm{out}}) \tag{2.18}$$

with the Poissonnian degree distribution $\pi_{\bar{k}}(k) = \mathrm{e}^{-\bar{k}}\bar{k}^k/k!$.

2.3. Final analytical expression for the entropy of the ensemble

The intermediate result (2.18) can now be substituted back into the expression for the entropy of the constrained random graph ensemble defined in (2.6), giving

$$S = \bar{k}[\log(N/\bar{k}) + 1] - \sum_{k^{\mathrm{in}},k^{\mathrm{out}}} p(k^{\mathrm{in}}, k^{\mathrm{out}}) \log\left(\frac{p(k^{\mathrm{in}}, k^{\mathrm{out}})}{\pi_{\bar{k}}(k^{\mathrm{in}})\pi_{\bar{k}}(k^{\mathrm{out}})}\right) + \zeta_N, \tag{2.19}$$

where $\bar{k}$ is the average connectivity, N is the number of nodes in the network, $p(k^{\mathrm{in}}, k^{\mathrm{out}})$ is its degree distribution that constrained the random graph ensemble, and $\lim_{N\to\infty} \zeta_N = 0$.

The compact form of (2.19) enables us to interpret and understand this result for the entropy per node. For example, we can consider what the result would have been if the constraint on the ensemble had been less restrictive. If our ensemble was a maximum entropy ensemble on the space of all directed graphs, but now constrained by the average degree only (as opposed to the full joint in- and out-degree distribution), then the entropy per node would have been $S = \bar{k}[\log(N/\bar{k}) + 1]$. We see that this is identical to what we would obtain from (2.19) if the constraining degree distribution was $p(k^{\mathrm{in}}, k^{\mathrm{out}}) = \pi(k^{\mathrm{in}})\pi(k^{\mathrm{out}})$; a trivial calculation confirms that in the maximum entropy ensemble with constrained average degree one indeed has $p(k^{\mathrm{in}}, k^{\mathrm{out}}) = \pi(k^{\mathrm{in}})\pi(k^{\mathrm{out}})$ for $N \to \infty$. Similarly, if we had chosen a maximum entropy ensemble of directed graphs constrained by a prescribed degree sequence (as opposed to a joint degree distribution), then the entropy would have taken the form

$$S = \bar{k}\big[\log(N/\bar{k}) + 1\big] + \sum_{k^{\mathrm{in}},k^{\mathrm{out}}} p(k^{\mathrm{in}}, k^{\mathrm{out}}) \log[\pi_{\bar{k}}(k^{\mathrm{in}})\pi_{\bar{k}}(k^{\mathrm{out}})] + \zeta_N. \tag{2.20}$$

This value is seen to be simply (2.19) minus the Shannon entropy of the joint degree distribution $p(k^{\mathrm{in}}, k^{\mathrm{out}})$, reflecting the possible ways to relabel sites in the original ensemble; this freedom is removed once we specify the individual degrees rather than their distribution.

3. Directed graphs with controlled degree distributions and degree–degree correlation functions

We extend our calculation to directed graph ensembles that are constrained further, by imposing a degree–degree correlation function in addition to a degree distribution. Degree-degree correlations in networks are known to carry valuable information. They can give rise to properties such as 'assortativity' or 'disassortativity' and often reflect the algorithm responsible for a network's generation. One such algorithm, 'preferential attachment', is well illustrated by the World Wide Web, where pages are more likely to be 'linked' to if they already have

many pages linking to them. Preferential attachment models such as [6] gained credibility by reproducing the typical fat tails often found in the degree distributions of real networks.

3.1. Definition of the problem

We now wish to generate graphs with degree pairs $\left(k_i^{\rm in}, k_i^{\rm out}\right)$ again drawn independently from the distribution $p(\vec{k}) = p(k^{\rm in}, k^{\rm out})$, but now the link probabilities are modified by some function $Q(\vec{k}_i, \vec{k}_j|\bar{p})$ of the degrees of the nodes concerned, and their distribution, with $\vec{k}_i = \left(k_i^{\rm in}, k_i^{\rm out}\right)$:

$$p(\boldsymbol{c}|p, Q) = \sum_{\vec{k}_1 \ldots \vec{k}_N} \left[\prod_i p(\vec{k}_i)\right] p(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N, Q) \tag{3.1}$$

$$\begin{aligned} p(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N, Q) &= \frac{w(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N, Q) \prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}}{Z(\vec{k}_1 \ldots \vec{k}_N, Q)} \\ Z(\vec{k}_1 \ldots \vec{k}_N, Q) &= \sum_{\boldsymbol{c}} w(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N, Q) \prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}. \end{aligned} \tag{3.2}$$

The difference with the graph ensemble in the previous section is the appearance of a new measure $w(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N, Q)$, defined as

$$w(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N, Q) = \prod_{i \neq j} \left[\frac{\bar{k}}{N} Q(\vec{k}_i, \vec{k}_j|\bar{p}) \delta_{c_{ij},1} + \left(1 - \frac{\bar{k}}{N} Q(\vec{k}_i, \vec{k}_j|\bar{p})\right) \delta_{c_{ij},0}\right] \tag{3.3}$$

with $Q(\vec{k}_i, \vec{k}_j|\bar{p}) \geqslant 0$ for all $(\vec{k}_i, \vec{k}_j)$, and with the distribution $\bar{p}(\vec{k}) = N^{-1} \sum_i \delta_{\vec{k}, \vec{k}_i}$ and the average degree $\bar{k} = N^{-1} \sum_i k_i^{\rm in} = N^{-1} \sum_i k_i^{\rm out}$ of the imposed degree sequence. The objective of the measure (3.3) is to deform the graph probabilities such as to impose a specific correlation profile between the degrees of connected nodes, by a suitable choice of the kernel $Q(.,.)$. We take $Q(.,.)$ to be normalized such that $w(\boldsymbol{c}|\ldots)$ is asymptotically consistent with the average degree $\bar{k}$. This means that we demand $N^{-2} \sum_{ij} Q(\vec{k}_i, \vec{k}_j|\bar{p}) = 1$. Equivalently, $\sum_{\vec{k},\vec{k}'} \bar{p}(\vec{k}) \bar{p}(\vec{k}') Q(\vec{k}, \vec{k}'|\bar{p}) = 1$, which explains why $Q(.,.)$ depends on the distribution $\bar{p}$. The entropy per node S of our ensemble is

$$S = -\sum_{\boldsymbol{c}} p(\boldsymbol{c}|p, Q) \Omega(\boldsymbol{c}|p, Q) \tag{3.4}$$

$$\Omega(\boldsymbol{c}|p, Q) = N^{-1} \log p(\boldsymbol{c}|p, Q). \tag{3.5}$$

3.2. Entropy evaluation

In appendix A we calculate the quantity (3.5) in leading orders in N, resulting in formula (A.23). Substitution into expression (3.4) for the entropy, followed by doing the average over $p(\boldsymbol{c}|p, Q)$ and some simple re-arranging of terms, then gives us

$$\begin{aligned} S = \bar{k}\left[\log(N/\bar{k}) + 1\right] &- \sum_{\vec{k}} p(\vec{k}) \log\left[\frac{p(\vec{k})}{\pi_{\bar{k}}(k^{\rm in}) \pi_{\bar{k}}(k^{\rm out})}\right] \\ &- \bar{k} \sum_{\vec{k},\vec{k}'} W(\vec{k}, \vec{k}') \log\left[\frac{R(\vec{k}|p, Q) Q(\vec{k}, \vec{k}'|p) S(\vec{k}'|p, Q)}{W_1(\vec{k}) W_2(\vec{k}')}\right] + \tilde{\zeta}_N \end{aligned} \tag{3.6}$$

with $\lim_{N\to\infty}\tilde{\zeta}_N = 0$, $\pi_{\bar{k}}(k) = \mathrm{e}^{-\bar{k}}\bar{k}^k/k!$, and $\bar{k} = \sum_{\vec{k}} p(\vec{k})k^{\mathrm{in}} = \sum_{\vec{k}} p(\vec{k})k^{\mathrm{out}}$. The kernel $W(\vec{k},\vec{k}')$ and its two marginals $W_{1,2}(\vec{k})$ in this expression are as defined in (A.8, A.9, A.10), but now calculated for graphs from our ensemble (3.1). Similarly, the quantities $R(\vec{k}|p, Q)$ and $Q(\vec{k}|p, Q)$ are now solved from

$$R(\vec{k}) = \frac{p(\vec{k})k^{\mathrm{in}}}{\bar{k}\sum_{\vec{k}'} Q(\vec{k},\vec{k}'|p)S(\vec{k}')}, \qquad S(\vec{k}) = \frac{p(\vec{k})k^{\mathrm{out}}}{\bar{k}\sum_{\vec{k}'} Q(\vec{k}',\vec{k}|p)R(\vec{k}')} \tag{3.7}$$

in which the distribution $p(\vec{k})$, its associated average $\bar{k}$, as well as the kernel $Q(\vec{k},\vec{k}'|p)$, correspond to ensemble (3.1). Thus the correct normalization of the kernel $Q(.,.)$ is $\sum_{\vec{k},\vec{k}'} p(\vec{k})p(\vec{k})Q(\vec{k},\vec{k}'|p) = 1$. What remains is to express the distribution $W(\vec{k},\vec{k}'|p, Q)$ for ensemble (3.1) in terms of $\{p, Q\}$. This is done in appendix B, resulting in (B.3):

$$\lim_{N\to\infty} W(\vec{k},\vec{k}') = R(\vec{k}|p, Q)Q(\vec{k},\vec{k}'|p)S(\vec{k}'|p, Q) \tag{3.8}$$

in which $R(\vec{k}|p, Q)$ and $S(\vec{k}|p, Q)$ are once more the solutions of (3.7), but now with $\tilde{p}(\vec{k})$ replaced by $p(\vec{k})$. Combination with (3.6) then gives us

$$S = \bar{k}[\log(N/\bar{k}) + 1] - \sum_{\vec{k}} p(\vec{k}) \log\left[\frac{p(\vec{k})}{\pi_{\bar{k}}(k^{\mathrm{in}})\pi_{\bar{k}}(k^{\mathrm{out}})}\right]$$
$$- \bar{k}\sum_{\vec{k},\vec{k}'} W(\vec{k},\vec{k}') \log\left[\frac{W(\vec{k},\vec{k}')}{W_1(\vec{k})W_2(\vec{k}')}\right] + \tilde{\epsilon}_N \tag{3.9}$$

with $\lim_{N\to\infty}\tilde{\epsilon}_N = 0$. Compared to the entropy per node (2.20) of ensembles where only the in-out degree distributions are imposed, we see that imposing in addition our new constraint, the specific degree–degree correlations as embodied by $W(\vec{k},\vec{k}')$, leads to a reduction of the entropy by an amount proportional to the mutual information of in-out degrees of connected nodes. An analogous result was derived in [1] for nondirected graphs. It can immediately be seen that if the in-out degrees of connected nodes are statistically independent, then the final nonvanishing term of 3.9 will be zero. Hence the entropy of the ensemble will in that case be the same as though the only constraint was the degree distribution.

4. Quantifying structural distance between networks

4.1. Derivation of the distance formula

In this section we define and calculate an information theoretic distance between two directed networks A and B, with in-out degree distributions $p_A(\vec{k})$ and $p_B(\vec{k})$ and with degree–degree correlation functions $W_A(\vec{k},\vec{k}')$ and $W_B(\vec{k},\vec{k}')$. We generalize to the present context of directed graphs the choice made in [1], viz. the Jeffreys divergence (i.e. symmetrized Kullback–Leibler distance) per node of the two associated ensembles from our family (3.1):

$$D_{AB} = \frac{1}{2N}\sum_{\boldsymbol{c}}\left\{p(\boldsymbol{c}|p_A, Q_A)\log\left[\frac{p(\boldsymbol{c}|p_A, Q_A)}{p(\boldsymbol{c}|p_B, Q_B)}\right]\right.$$
$$\left. + p(\boldsymbol{c}|p_B, Q_B)\log\left[\frac{p(\boldsymbol{c}|p_B, Q_B)}{p(\boldsymbol{c}|p_A, Q_A)}\right]\right\} \tag{4.1}$$

D_{AB} is non-negative and equals zero only when both networks A and B belong to the same tailored graph ensemble (i.e. have equivalent constraints). Upon writing the Shannon entropies per node of the ensembles A and B as S_A and S_B, we have

$$D_{AB} = \tfrac{1}{2}(S_{AB} + S_{BA} - S_{AA} - S_{BB}) \tag{4.2}$$

where, using the abbreviation (3.5),

$$S_{AB} = -\frac{1}{N}\sum_{\boldsymbol{c}} p(\boldsymbol{c}|p_A, Q_A)\log p(\boldsymbol{c}|p_B, Q_B)$$
$$= -\sum_{\boldsymbol{c}} p(\boldsymbol{c}|p_A, Q_A)\Omega(\boldsymbol{c}|p_B, Q_B) \tag{4.3}$$

with $\Omega(\boldsymbol{c}|p, Q)$ as defined in (3.5). We may now use result (A.23) of appendix A, but in doing so it is vital to keep track carefully of the labels (A, B) of the degree distributions and kernels. In particular, according to (4.3) we must make in (A.23) the substitutions $p(\vec{k}|\boldsymbol{c}) \to p_A(\vec{k})$, $W(\vec{k}, \vec{k}'|\boldsymbol{c}) \to W_A(\vec{k}, \vec{k}')$, $p(\vec{k}) \to p_B(\vec{k})$, and $Q(\vec{k}, \vec{k}'|\tilde{p}) \to Q_B(\vec{k}, \vec{k}'|p_A)$. This leads us to

$$\lim_{N\to\infty} S_{AB} = -\sum_{\vec{k}} p_A(\vec{k})\log p_B(\vec{k}) - \bar{k}_A\left[1 + \log\left(\frac{\bar{k}_A}{N}\right)\right] - \sum_{\vec{k}} p_A(\vec{k})\log(k^{\text{in}}!k^{\text{out}}!)$$
$$+ \sum_{\vec{k}} p_A(\vec{k})k^{\text{in}}\log\left[\frac{p_A(\vec{k})k^{\text{in}}}{R(\vec{k}|p_A, Q_B)}\right] + \sum_{\vec{k}} p_A(\vec{k})k^{\text{out}}\log\left[\frac{p_A(\vec{k})k^{\text{out}}}{S(\vec{k}|p_A, Q_B)}\right]$$
$$- \bar{k}_A\sum_{\vec{k},\vec{k}'} W_A(\vec{k}, \vec{k}')\log Q_B(\vec{k}, \vec{k}'|p_A) \tag{4.4}$$

in which $R(\vec{k}|p_A, Q_B)$ and $S(\vec{k}|p_A, Q_B)$ are to be solved from

$$R(\vec{k}) = \frac{p_A(\vec{k})k^{\text{in}}}{\bar{k}_A\sum_{\vec{k}'} Q_B(\vec{k}, \vec{k}'|p_A)S(\vec{k}')}, \qquad S(\vec{k}) = \frac{p_A(\vec{k})k^{\text{out}}}{\bar{k}_A\sum_{\vec{k}'} Q_B(\vec{k}', \vec{k}|p_A)R(\vec{k}')}. \tag{4.5}$$

Hence, upon assembling and combining the various terms in (4.2) and upon using relations such as (A.9, A.10) and (B.3) to simplify the result, we find

$$D_{AB} = \frac{1}{2}\sum_{\vec{k}} p_A(\vec{k})\log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right] + \frac{1}{2}\sum_{\vec{k}} p_B(\vec{k})\log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$$
$$+ \frac{1}{2}\bar{k}_A\sum_{\vec{k},\vec{k}'} W_A(\vec{k}, \vec{k}')\log\left[\frac{W_A(\vec{k}, \vec{k}')}{R(\vec{k}|p_A, Q_B)Q_B(\vec{k}, \vec{k}'|p_A)S(\vec{k}'|p_A, Q_B)}\right]$$
$$+ \frac{1}{2}\bar{k}_B\sum_{\vec{k},\vec{k}'} W_B(\vec{k}, \vec{k}')\log\left[\frac{W_B(\vec{k}, \vec{k}')}{R(\vec{k}|p_B, Q_A)Q_A(\vec{k}, \vec{k}'|p_B)S(\vec{k}'|p_B, Q_A)}\right]. \tag{4.6}$$

According to (B.3), the product $W_{AB}(\vec{k}, \vec{k}') = R(\vec{k}|p_A, Q_B)Q_B(\vec{k}, \vec{k}'|p_A)S(\vec{k}'|p_A, Q_B)$ equals the joint distribution of in- and out- degrees of connected nodes in an ensemble of the family (3.1) that would have been obtained upon choosing the hybrid combination $\{p_A, Q_B\}$ of degree distribution and wiring kernel, where Q_B is normalized

according to $\sum_{\vec{k},\vec{k}'} p_A(\vec{k})p_A(\vec{k}')Q_B(\vec{k},\vec{k}'|p_A) = 1$. Similarly, the product $W_{BA}(\vec{k},\vec{k}') = R(\vec{k}|p_B, Q_A)Q_A(\vec{k},\vec{k}'|p_B)S(\vec{k}'|p_B, Q_A)$ would have been obtained for the ensemble $\{p_B, Q_A\}$. Thus we may write

$$\lim_{N\to\infty} D_{AB} = \frac{1}{2}\sum_{\vec{k}} p_A(\vec{k}) \log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right] + \frac{1}{2}\sum_{\vec{k}} p_B(\vec{k}) \log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$$
$$+ \frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log\left[\frac{W_A(\vec{k},\vec{k}')}{W_{AB}(\vec{k},\vec{k}')}\right]$$
$$+ \frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log\left[\frac{W_B(\vec{k},\vec{k}')}{W_{BA}(\vec{k},\vec{k}')}\right]. \tag{4.7}$$

This appealing formula shows that $D_{AB} \geqslant 0$ for all choices of (A, B), with equality if and only if $W_A = W_B$; in the later case one automatically will have $W_{AB} = W_{BA} = W_A = W_B$. In the case where degree–degree correlations are absent from both networks one will find $W_{AB}(\vec{k},\vec{k}') = W_A(\vec{k},\vec{k}') = W_{1A}(\vec{k})W_{2A}(\vec{k}')$, and formula (4.7) reduces to the Jeffreys divergence between the degree distributions p_A and p_B.

4.2. Practical form of the distance formula

In contrast to W_A and W_B, which correspond to the two given networks $\boldsymbol{c}_A$ and $\boldsymbol{c}_B$, we cannot measure W_{AB} and W_{BA}; the later would correspond to hypothetical hybrid networks. Hence in order to use (4.7) in practice it will be convenient to write it in an alternative form:

$$\lim_{N\to\infty} D_{AB} = \frac{1}{2}\sum_{\vec{k}} p_A(\vec{k}) \log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right] + \frac{1}{2}\sum_{\vec{k}} p_B(\vec{k}) \log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$$
$$+ \frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log\left[\frac{W_A(\vec{k},\vec{k}')}{W_B(\vec{k},\vec{k}')}\right] + \frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log\left[\frac{W_B(\vec{k},\vec{k}')}{W_A(\vec{k},\vec{k}')}\right]$$
$$+ \frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log\left[\frac{W_B(\vec{k},\vec{k}')}{R(\vec{k}|p_A, Q_B)Q_B(\vec{k},\vec{k}'|p_A)S(\vec{k}'|p_A, Q_B)}\right]$$
$$+ \frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log\left[\frac{W_A(\vec{k},\vec{k}')}{R(\vec{k}|p_B, Q_A)Q_A(\vec{k},\vec{k}'|p_B)S(\vec{k}'|p_B, Q_A)}\right]. \tag{4.8}$$

If we choose Q_A and Q_B to be the canonical kernels for the two ensembles A and B, i.e. $Q_A(\vec{k},\vec{k}'|\bar{p}) = W_A(\vec{k},\vec{k}')/\bar{p}(\vec{k})\bar{p}(\vec{k}')$ and $Q_B(\vec{k},\vec{k}'|\bar{p}) = W_B(\vec{k},\vec{k}')/\bar{p}(\vec{k})\bar{p}(\vec{k}')$, expression (4.8) simplifies to

$$\lim_{N\to\infty} D_{AB} = \frac{1}{2}\sum_{\vec{k}} p_A(\vec{k}) \log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right] + \frac{1}{2}\sum_{\vec{k}} p_B(\vec{k}) \log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$$
$$+ \frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log\left[\frac{W_A(\vec{k},\vec{k}')}{W_B(\vec{k},\vec{k}')}\right] + \frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log\left[\frac{W_B(\vec{k},\vec{k}')}{W_A(\vec{k},\vec{k}')}\right]$$

$$+\frac{1}{2}\bar{k}_A\left\{\sum_{\vec{k}}W_{1A}(\vec{k})\log\left[\frac{p_A(\vec{k})}{R(\vec{k}|p_A,Q_B)}\right]+\sum_{\vec{k}'}W_{2A}(\vec{k}')\log\left[\frac{p_A(\vec{k}')}{S(\vec{k}'|p_A,Q_B)}\right]\right\}$$

$$+\frac{1}{2}\bar{k}_B\left\{\sum_{\vec{k}}W_{1B}(\vec{k})\log\left[\frac{p_B(\vec{k})}{R(\vec{k}|p_B,Q_A)}\right]+\sum_{\vec{k}'}W_{2B}(\vec{k}')\log\left[\frac{p_B(\vec{k}')}{S(\vec{k}'|p_B,Q_A)}\right]\right\} \tag{4.9}$$

with $R(\vec{k}|p_A,Q_B)$ and $S(\vec{k}|p_A,Q_B)$ to be solved from

$$R(\vec{k})/p_A(\vec{k})=\frac{W_{1A}(\vec{k})}{\sum_{\vec{k}'}W_B(\vec{k},\vec{k}')[S(\vec{k}')/p_A(\vec{k}')]}, \tag{4.10}$$

$$S(\vec{k})/p_A(\vec{k}')=\frac{W_{2A}(\vec{k})}{\sum_{\vec{k}'}W_B(\vec{k}',\vec{k})[R(\vec{k}')/p_A(\vec{k}')]}. \tag{4.11}$$

Next we rewrite the arguments of the logarithms in the second line of (4.8) in terms of the two degree correlation ratios $\Pi_A(\vec{k},\vec{k}')=W_A(\vec{k},\vec{k}')/W_{1A}(\vec{k})W_{2A}(\vec{k}')$ and $\Pi_B(\vec{k},\vec{k}')=W_B(\vec{k},\vec{k}')/W_{1B}(\vec{k})W_{2B}(\vec{k}')$. We also transform the order parameters $R(\vec{k}|p_A,Q_B)$ and $S(\vec{k}|p_A,Q_B)$ to new functions $\rho_{AB}(\vec{k})$ and $\sigma_{AB}(\vec{k})$ via

$$\rho_{AB}(\vec{k})=\frac{p_A(\vec{k})W_{1A}(\vec{k})}{R(\vec{k}|p_A,Q_B)W_{1B}(\vec{k})},\qquad \sigma_{AB}(\vec{k})=\frac{p_A(\vec{k})W_{2A}(\vec{k})}{S(\vec{k}|p_A,Q_B)W_{2B}(\vec{k})}. \tag{4.12}$$

Our distance then becomes

$$\begin{aligned}\lim_{N\to\infty}D_{AB}=&\frac{1}{2}\sum_{\vec{k}}p_A(\vec{k})\log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right]+\frac{1}{2}\sum_{\vec{k}}p_B(\vec{k})\log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]\\&+\frac{1}{2}\bar{k}_A\sum_{\vec{k},\vec{k}'}W_A(\vec{k},\vec{k}')\log\left[\frac{\Pi_A(\vec{k},\vec{k}')}{\Pi_B(\vec{k},\vec{k}')}\right]+\frac{1}{2}\bar{k}_B\sum_{\vec{k},\vec{k}'}W_B(\vec{k},\vec{k}')\log\left[\frac{\Pi_B(\vec{k},\vec{k}')}{\Pi_A(\vec{k},\vec{k}')}\right]\\&+\frac{1}{2}\bar{k}_A\sum_{\vec{k}}W_{1A}(\vec{k})\log\rho_{AB}(\vec{k})+\frac{1}{2}\bar{k}_A\sum_{\vec{k}}W_{2A}(\vec{k})\log\sigma_{AB}(\vec{k})\\&+\frac{1}{2}\bar{k}_B\sum_{\vec{k}}W_{1B}(\vec{k})\log\rho_{BA}(\vec{k})+\frac{1}{2}\bar{k}_B\sum_{\vec{k}}W_{2B}(\vec{k})\log\sigma_{BA}(\vec{k})\end{aligned} \tag{4.13}$$

in which $\rho_{AB}(\vec{k})$ and $\sigma_{AB}(\vec{k})$ are to be solved from

$$\rho_{AB}(\vec{k})=\sum_{\vec{k}'}\Pi_B(\vec{k},\vec{k}')W_{2A}(\vec{k}')\sigma_{AB}^{-1}(\vec{k}') \tag{4.14}$$

$$\sigma_{AB}(\vec{k})=\sum_{\vec{k}'}\Pi_B(\vec{k}',\vec{k})W_{1A}(\vec{k}')\rho_{AB}^{-1}(\vec{k}'). \tag{4.15}$$

Whenever $p_A=p_B$ or $\Pi_A=\Pi_B$ (or both), the solution of (4.14, 4.15) will be $\rho_{AB}(\vec{k})=\sigma_{AB}(\vec{k})=1$ for all $\vec{k}$. Hence the last two lines of (4.13) represent corrections to the distance formula, that reflect interference between the constraints imposed by prescribed degree statistics and those imposed by presecribed degree correlations[5].

[5] A similar interference term was erroneously omitted from [1], which can be confirmed by retracing the above arguments and the calculations in appendix A for nondirected graphs. We will summarize and compare our results for directed and nondirected graphs below.

We note, finally, that although definition (4.1) requires that the networks A and B have the same number of nodes, the final form (4.13) of our formula does not depend on the (relative) network sizes. Hence we will apply the result (4.1) also to networks of different sizes, provided both are sufficiently large, which makes (4.1) more widely applicable to real networks (which will in general be large, but of different sizes).

5. Tests, comparisons, and applications

5.1. Simple special cases

If the in-degrees are statistically independent of the out-degrees, i.e. $p(\vec{k}) = p(k^{\rm in})p(k^{\rm out})$, the entropy per node (2.19) of the ensemble (2.1) with prescribed degree statistics but no degree correlations simplifies to

$$S = \bar{k}\left[\log\left(\frac{N}{\bar{k}}\right)+1\right] - \sum_{k^{\rm in}} p(k^{\rm in}) \log\left[\frac{p(k^{\rm in})}{\pi_{\bar{k}}(k^{\rm in})}\right] - \sum_{k^{\rm out}} p(k^{\rm out}) \log\left[\frac{p(k^{\rm out})}{\pi_{\bar{k}}(k^{\rm out})}\right] + \zeta_N \qquad (5.1)$$

with $\lim_{N\to\infty} \zeta_N = 0$. This, according to [1], is the sum of the individual entropies of the 'out-graph' ensemble and the 'in-graph' ensemble, calculated as though they were considered as two separate undirected networks. In ensembles with degree correlations, i.e. (3.1), with entropy per node (3.9), the additional term that represents the entropy reduction imposed by the degree correlations does not simplify as a result of assuming $p(\vec{k}) = p(k^{\rm in})p(k^{\rm out})$; the degree correlations can generate statistical relations between in- and out-degrees that are not visible in $p(\vec{k})$.

A regular directed graph is one where each node has the same in- and the same out-degree. Since for a well-defined directed graph, we also have $\sum_{\vec{k}} p(\vec{k})k^{\rm in} = \sum_{\vec{k}} p(\vec{k})k^{\rm out} = \bar{k}$, any regular directed graph must have $p(\vec{k}) = \delta_{\vec{k},(\bar{k},\bar{k})}$. This, in turn, implies also that $W(\vec{k},\vec{k}') = \delta_{\vec{k},(\bar{k},\bar{k})}\delta_{\vec{k}',(\bar{k},\bar{k})}$. So it is impossible to have degree correlations, and both equation (2.19) and (3.9) reduce to

$$S = \bar{k}[\log(N\bar{k}) - 1] - 2\log(\bar{k}!) + \zeta_N. \qquad (5.2)$$

5.2. Comparison of formulae for undirected versus directed networks

It is instructive to give an overview of the similarities and differences between directed and nondirected graphs. Instead of entropies per node, we will also compare entropic results in terms of complexities. The degree complexity per node $\mathcal{C}_{\rm deg}$ of a graph $\boldsymbol{c}$ is the difference between the entropy per node of the associated ensemble (2.1) and the value $S_0[\bar{k}]$ that is found for the entropy per node if only the average connectivity $\bar{k}$ is prescribed (i.e. for an ensemble with Poisson distributed degrees). The wiring complexity $\mathcal{C}_{\rm wir}$ is the further entropy reduction that results if we go from the ensemble (2.1) to the ensemble (3.1) where also the degree–degree correlations are imposed. Our results can then be summarized as in table 1.

Similarly we can compare the formulae for the information-theoretic distance D_{AB} between two networks $\boldsymbol{c}_A$ and $\boldsymbol{c}_B$, for directed versus nondirected ones. This gives in both cases $\lim_{N\to\infty} D_{AB} = D_{AB}^{\rm deg} + D_{AB}^{\rm wir} + D_{AB}^{\rm int}$, where $D_{AB}^{\rm deg}$ is the direct contribution from degree distribution dissimilarity, $D_{AB}^{\rm wir}$ is the direct contribution from degree-correlation dissimilarity, and $D_{AB}^{\rm int}$ accounts for the interference between degree statistics and the possible degree correlations that could be achieved. Our distance results can then be summarized in table 2.

Table 1. Comparison of entropies and complexities of directed versus nondirected graphs. The entropy per node is given by $S[p,W] = S_0[\bar{k}] - \mathcal{C}_{\rm deg}[p] - \mathcal{C}_{\rm wir}[p,W]$, modulo finite size corrections. For ensembles in which only the average connectivity $\bar{k}$ is prescribed one would find the value $S_0[\bar{k}]$. The quantities $\mathcal{C}_{\rm deg}[p]$ and $\mathcal{C}_{\rm wir}[p,W]$ measure the entropy reductions caused by subsequently imposing a degree distribution p, and the joint distribution W of connected nodes, and can therefore be identified with the degree complexity and the wiring complexity of the typical graphs in our ensembles. In directed graphs $\vec{k} = (k^{\rm in}, k^{\rm out})$, where $k_i^{\rm in}(\boldsymbol{c}) = \sum_j c_{ij}$ and $k_i^{\rm out}(\boldsymbol{c}) = \sum_j c_{ji}$, and $W(\vec{k},\vec{k}') = (N\bar{k})^{-1}\sum_{ij} c_{ij}\delta_{\vec{k},\vec{k}_i}\delta_{\vec{k}',\vec{k}_j}$. In nondirected graphs one has only $k_i(\boldsymbol{c}) = \sum_j c_{ij}$, and $W(k,k') = (N\bar{k})^{-1}\sum_{ij} c_{ij}\delta_{k,k_i}\delta_{k',k_j}$.

	directed graphs	nondirected graphs
$S_0[\bar{k}]:$	$\bar{k}[\log(N/\bar{k})+1]$	$\frac{1}{2}\bar{k}[\log(N/\bar{k})+1]$
$\mathcal{C}_{\rm deg}[p]:$	$\sum_{\vec{k}} p(\vec{k}) \log\left[\frac{p(\vec{k})}{\pi_{\vec{k}}(k^{\rm in})\pi_{\vec{k}}(k^{\rm out})}\right]$	$\sum_k p(k)\log\left[\frac{p(k)}{\pi_k(k)}\right]$
$\mathcal{C}_{\rm wir}[p,W]:$	$\bar{k}\sum_{\vec{k},\vec{k}'} W(\vec{k},\vec{k}')\log\left[\frac{W(\vec{k},\vec{k}')}{W_1(\vec{k})W_2(\vec{k}')}\right]$	$\frac{1}{2}\bar{k}\sum_{k,k'} W(k,k')\log\left[\frac{W(k,k')}{W(k)W(k')}\right]$

Table 2. Comparison of the contributions to the distance $\lim_{N\to\infty} D_{AB} = D_{AB}^{\rm deg} + D_{AB}^{\rm wir} + D_{AB}^{\rm int}$, between graphs $\boldsymbol{c}_A$ and $\boldsymbol{c}_B$. Notation conventions are mostly as in the caption of table 1. The degree correlation ratios Π are defined as $\Pi(\vec{k},\vec{k}') = W(\vec{k},\vec{k}')/W_1(\vec{k})W_2(\vec{k}')$ (for directed graphs) and $\Pi(k,k') = W(\vec{k},\vec{k}')/W(k)W(k')$ (for nondirected graphs). The functions $\rho_{AB}(\vec{k})$ and $\sigma_{AB}(\vec{k})$ (for directed graphs) are the solutions of equations (4.14, 4.15). The functions $\rho_{AB}(k)$ (for nondirected graphs) are to be solved from equation (5.3).

	directed graphs	nondirected graphs
$D_{AB}^{\rm deg}:$	$\frac{1}{2}\sum_{\vec{k}} p_A(\vec{k})\log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right]$	$\frac{1}{2}\sum_k p_A(k)\log\left[\frac{p_A(k)}{p_B(k)}\right]$
	$+\frac{1}{2}\sum_{\vec{k}} p_B(\vec{k})\log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$	$+\frac{1}{2}\sum_k p_B(k)\log\left[\frac{p_B(k)}{p_A(k)}\right]$
$D_{AB}^{\rm wir}:$	$\frac{1}{2}\bar{k}_A\sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}')\log\left[\frac{\Pi_A(\vec{k},\vec{k}')}{\Pi_B(\vec{k},\vec{k}')}\right]$	$\frac{1}{4}\bar{k}_A\sum_{k,k'} W_A(k,k')\log\left[\frac{\Pi_A(k,k')}{\Pi_B(k,k')}\right]$
	$+\frac{1}{2}\bar{k}_B\sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}')\log\left[\frac{\Pi_B(\vec{k},\vec{k}')}{\Pi_A(\vec{k},\vec{k}')}\right]$	$+\frac{1}{4}\bar{k}_B\sum_{k,k'} W_B(k,k')\log\left[\frac{\Pi_B(k,k')}{\Pi_A(k,k')}\right]$
$D_{AB}^{\rm int}:$	$\frac{1}{2}\bar{k}_A\sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}')\log[\rho_{AB}(\vec{k})\sigma_{AB}(\vec{k}')]$	$\frac{1}{2}\bar{k}_A\sum_k W_A(k)\log\rho_{AB}(k)$
	$+\frac{1}{2}\bar{k}_B\sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}')\log[\rho_{BA}(\vec{k})\sigma_{BA}(\vec{k}')]$	$+\frac{1}{2}\bar{k}_B\sum_k W_B(k)\log\rho_{BA}(k)$

The functions $\rho_{AB}(\vec{k})$ and $\sigma_{AB}(\vec{k})$ are solved from (4.14, 4.15). Repeating the calculation for nondirected graphs shows that there only one function $\rho_{AB}(k)$ is required (or equivalently, $\rho_{AB} = \sigma_{AB}$), which is the solution of

$$\rho_{AB}(k) = \sum_{k'} \Pi_B(k,k')W_A(k')\rho_{AB}^{-1}(k'). \tag{5.3}$$

5.3. Application to gene regulation networks

A gene regulation network can be viewed as a directed graph, where the nodes represent genes and the arcs indicate whether ($c_{ij} = 1$) or not ($c_{ij} = 0$) the protein synthesized from gene j acts as a regulator of gene i. In the present binary set-up, where $c_{ij} \in \{0,1\}$, one disregards information on the nature of regulation, i.e. whether it involves repression or activation.

In tables 3 and 4, we show the results of calculating the various contributions to the entropy of the ensemble associated with the networks of [9] and [10], respectively. Imposing only

Table 3. The tailoring of random graph ensembles by imposing as constraints the values of increasingly prescriptive macroscopic topological features measured in the gene regulation network of [9]. This tailoring reduces the entropy per node S in the ensemble in stages, and thereby the effective number of graphs $\mathcal{N} = \exp[NS]$ compatible with the network of [9]. We observe that, in this example, refining the tailoring of the graph ensemble from imposing only the correct average degree to imposing the correct degree distribution is more significant than the further refinement of imposing the correct degree–degree correlations. Hence the degree complexity of this network is significantly larger than the wiring complexity.

Imposed topological property	Entropy per node	Entropy per arc
Gene regulation network of Hughes et al (2000)		
average degree $\bar{k}$	44.5	7.9
degree distribution $p(\vec{k})$	19.5	3.5
degree–degree correlations $\Pi(\vec{k}, \vec{k}')$	17.9	3.2

Table 4. The tailoring of random graph ensembles by imposing as constraints the values of increasingly prescriptive macroscopic topological features measured in the gene regulation network of [10]. The tailoring reduces the entropy per node S in the ensemble in stages, and thereby the effective number of graphs $\mathcal{N} = \exp[NS]$ compatible with the network of [10]. As in the previous example, refining the tailoring of the graph ensemble from imposing only the correct average degree to imposing the correct degree distribution is more significant than the further refinement of imposing the correct degree–degree correlations. Hence the degree complexity of this network is again significantly larger than the wiring complexity.

Imposed topological property	Entropy per node	Entropy per arc
Gene regulation network of Harbison et al (2004)		
average degree $\bar{k}$	23.2	8.2
degree distribution $p(\vec{k})$	12.8	4.5
degree–degree correlations $\Pi(\vec{k}, \vec{k}')$	11.6	4.1

the correct average degree gives the entropy $S_0[\bar{k}]$. Imposing in addition the correct degree distribution (i.e. representing the network by ensemble (2.1)) gives the entropy $S_0[\bar{k}] - \mathcal{C}_{\rm deg}[p]$. Imposing additionally the correct degree–degree correlations (i.e. representing the network by ensemble (3.1)) reduces the entropy still further to $S_0[\bar{k}] - \mathcal{C}_{\rm deg}[p] - \mathcal{C}_{\rm wir}[p, W]$.

In both tables we also show the entropies per arc, defined as $S' = S/\bar{k}$. The latter are normalised for the average degree. This fits in with the 'arc centric' view that the calculations in this paper and its predecessor [1] seem to have steered us in, where the final answers are consistently found to be most elegantly formulated in terms of the joint distribution W of degrees at either end of an arc.

In [9] Hughes *et al* used a two-color cDNA micro-array hybridization assay to generate expression profiles in yeast for 276 deletion mutants. We followed an approach published by Rung *et al* [11] to construct a network from this data. Two genes g1, g2 are connected by an arc from g1 to g2 if the ratio of the expression level in the mutant where gene g1 is deleted versus the background standard deviation in the wild-type strain is larger than a threshold. In this way, we arrived at a directed network with $N = 5654$ nodes (genes), with an average degree $\bar{k} \approx 5.6$. The degree distribution of this network is characterised by high frequency of occurrence of low degree nodes; the set of nodes with out-degree zero and in-degree less than 4 covers more than 50% of the set. However, the network also contains some nodes with very high out-degree.

The authors of [10], Harbison *et al* reported on a study of DNA binding transcriptional regulators in yeast. For each of the 203 transcription factors tested they report the genes where

the transcription factor bound to the putative promoter region. Similar to a previous study [12] we constructed a network by connecting gene g1, which encodes a transcription factor, to gene g2 if the measurements were statistically significant ($P \leqslant 0.001$). Their data were represented as a directed network of $N = 3865$ nodes, with an average degree of $\bar{k} \approx 2.81$. Compared with the data of [9], the network of [10] is more sparse. It does, however, show a similar degree distribution pattern — in fact over 50% of the nodes have zero out-degree and an in-degree of less than 2.

In practice, when the gene network data are collected, a decision has to be made about the cut-off point where the effect of one gene product on another gene is so small as to be considered insignificant. If there was no threshold and every small fluctuation was taken to be evidence of co-regulation, then it would appear that every gene regulated every other gene, and the network would be complete. Conversely, setting too strict a threshold will risk missing out on important but subtle interactions.

Changing the threshold would reduce the number of arcs, and hence make the network more sparse with lower average degree. Our base assumption would be that beyond that, the main qualitative features of the topology would be maintained. That is, the stricter threshold would remove arcs indiscriminately across the network. However, it is possible that, for example, a node would appear to be a 'hub' under a lenient criterion, but would lose a large number of interactions under stricter criteria, so that it is no longer a hub: this would be a qualitative change to the topology arising from the change in thresholds. The analysis proposed in this paper is measuring the topological properties of the network (rather than the network itself). We would expect these results to vary insofar as the topological properties varied. Figure 1 shows the results of repeating the analysis above for different values of the thresholds.

The above data all refer to the same organism, yeast; however, they present different aspects of gene interactions. Hence, even more than for protein–protein interaction networks, comparison must be done cautiously. The heterogeneity in the data sets emphasises the importance of developing a suite of tools and measures that can be used to study each network independently.

6. Discussion

In this paper we have derived several mathematical results for directed random graph ensembles tailored to match chosen properties of real-world networks. We have calculated the Shannon entropy of ensembles constrained by a prescribed degree distribution, and of ensembles constrained by a prescribed degree–degree correlation function (which contains more detailed topological information than the degree distribution). We have also defined a rational information-theoretic distance measurement for comparing networks based on their degree distribution and degree–degree correlation. All this complements and generalises earlier work done in [1] for nondirected networks. We also identified a correction term to the distance measure of nondirected graphs which was absent in [1]. A summary of our results and how they compare with the corresponding formula for nondirected networks is presented in tables 1 and 2.

Our growing suite of quantitative tools can be used to study the properties of large real world networks. These tools are precise in leading order in N, and take the form of explicit and transparent formulae which use easily measurable macroscopic parameters as input. The present generalization to nondirected networks enables their application to gene regulation networks. We trust that the benefits of having explicit formulae for network complexities and information-theoretic dissimilarity measures will increase, especially in bioinformatics,

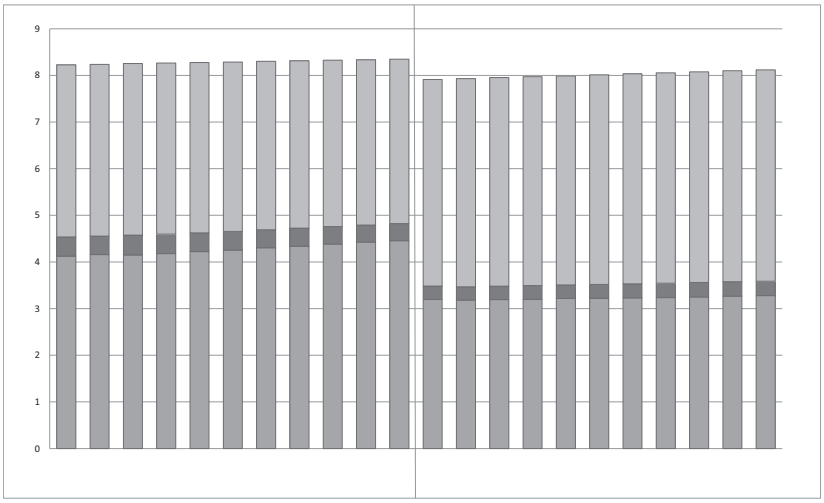

Figure 1. Each bar on the chart represents a different choice of threshold. Moving from left to right, the threshold is made progressively stricter so as to exclude approximately 3 percent of arcs at each step. The left half refers to Harbison *et al* [10] data; the right half refers to Hughes *et al* [9] data. Within a bar, the top line presents the entropy per bond when the constraint is 'average degree'; the next line shows the entropy per bond when the constraint is additionally 'degree distribution'; and, the final line gives the entropy per bond for the ensemble additionally targeting the 'degree–degree correlation'. Hence the top two shaded areas represent the degree complexity and the wiring complexity, respectively. Both datasets are plotted on the same axis in order to illustrate that, although there is some movement with different thresholds, the results for the two different networks remain distinct and distinguishable for any reasonable choice of threshold, and are not unduly sensitive to any reasonable choice of threshold.

as we gain experience with using and interpreting the method, and as we increase the range of topological properties to which we can tailor our graph ensembles.

The focus of our future work will be to increase the number of topological properties that we can characterise, measure, and impose upon tailored random graph ensembles. Significant progress has already been made towards including distributions of so-called generalised degrees, but our priority will be to focus on observables that measure the statistics of short loops. In the presence of such loops the methods and ideas that we applied so far will no longer suffice. However, short loops appear to be key biological motifs, so progress in this direction should yield substantial benefits in terms of applicability of the method in biological signalling.

Appendix A. Order parameter representation of the graph probabilities

In this section we derive a tool that is repeatedly used in this paper, being a formula in terms of simple observables and order parameters of the log-probability per node of graphs (3.5) given the ensemble definition (3.1), in leading orders in N. Upon substituting (3.1) into this formula, and after some simple manipulations and use of the law of large numbers, one finds

$$\Omega(\boldsymbol{c}|p,Q) = \sum_{\vec{k}} p(\vec{k}|\boldsymbol{c}) \log p(\vec{k}) + \phi_1(\boldsymbol{c}|Q) - \phi_2(\boldsymbol{c}|Q) + \epsilon_N \tag{A.1}$$

$$\phi_1(\boldsymbol{c}|Q) = \frac{1}{N}\log w(\boldsymbol{c}|\vec{k}_1,\ldots,\vec{k}_N,Q)\Big|_{\vec{k}_i=\vec{k}_i(\boldsymbol{c})\forall i} \tag{A.2}$$

$$\phi_2(\boldsymbol{c}|Q) = \frac{1}{N}\log Z(\vec{k}_1,\ldots,\vec{k}_N,Q)\Big|_{\vec{k}_i=\vec{k}_i(\boldsymbol{c})\forall i} \tag{A.3}$$

with $\epsilon_N \to 0$ as $N \to \infty$, and

$$Z(\vec{k}_1,\ldots,\vec{k}_N,Q) = \sum_{\boldsymbol{c}} w(\boldsymbol{c}|\vec{k}_1,\ldots,\vec{k}_N,Q)\prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})} \tag{A.4}$$

$$w(\boldsymbol{c}|\vec{k}_1,\ldots,\vec{k}_N,Q) = \prod_{i\neq j}\left[\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})\delta_{c_{ij},1} + \left(1-\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})\right)\delta_{c_{ij},0}\right]. \tag{A.5}$$

In these expressions $\bar{k} = N^{-1}\sum_i k_i^{\text{in}} = N^{-1}\sum_i k_i^{\text{out}}$, $\bar{p}(\vec{k}) = N^{-1}\sum_i \delta_{\vec{k},\vec{k}_i}$, and the kernel $Q(.,.)$ is normalized locally according to $\sum_{\vec{k},\vec{k}'}\bar{p}(\vec{k})\bar{p}(\vec{k}')Q(\vec{k},\vec{k}'|\bar{p}) = 1$.

A.1. Calculation of ϕ_1

The first contribution (A.2) to the entropy is calculated easily:

$$\begin{aligned}\phi_1(\boldsymbol{c}|Q) &= \frac{1}{N}\sum_{i\neq j}\left\{c_{ij}\log\left[\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})\right] - \frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})\right\}\Big|_{\vec{k}_i=\vec{k}_i(\boldsymbol{c})\,\forall i} + \mathcal{O}\left(\frac{1}{N}\right)\\ &= \bar{k}(\boldsymbol{c})\left\{\log\left[\frac{\bar{k}(\boldsymbol{c})}{N}\right] - 1 + \sum_{\vec{k},\vec{k}'}W(\vec{k},\vec{k}'|\boldsymbol{c})\log Q(\vec{k},\vec{k}'|p(.|\boldsymbol{c}))\right\} + \mathcal{O}\left(\frac{1}{N}\right).\end{aligned} \tag{A.6}$$

It involves the in- and out-degree distribution $p(\vec{k}|\boldsymbol{c})$, its degree average $\bar{k}(\boldsymbol{c})$, and the joint distribution $W(\vec{k},\vec{k}'|\boldsymbol{c})$ of in- and out-degrees of connected nodes. All are calculated for the graph $\boldsymbol{c}$ and defined as

$$p(\vec{k}|\boldsymbol{c}) = \frac{1}{N}\sum_i \delta_{\vec{k},\vec{k}(\boldsymbol{c})} \tag{A.7}$$

$$W(\vec{k},\vec{k}'|\boldsymbol{c}) = \frac{1}{N\bar{k}(\boldsymbol{c})}\sum_{ij} c_{ij}\delta_{\vec{k},\vec{k}_i(\boldsymbol{c})}\delta_{\vec{k}',\vec{k}_j(\boldsymbol{c})}. \tag{A.8}$$

They are related via the two identities

$$W_1(\vec{k}|\boldsymbol{c}) = \sum_{\vec{k}'}W(\vec{k},\vec{k}'|\boldsymbol{c}) = \frac{k^{\text{in}}}{\bar{k}(\boldsymbol{c})}p(\vec{k}|\boldsymbol{c}) \tag{A.9}$$

$$W_2(\vec{k}|\boldsymbol{c}) = \sum_{\vec{k}'}W(\vec{k}',\vec{k}|\boldsymbol{c}) = \frac{k^{\text{out}}}{\bar{k}(\boldsymbol{c})}p(\vec{k}|\boldsymbol{c}). \tag{A.10}$$

The kernel in (A.6) is normalized according to $\sum_{\vec{k},\vec{k}'}p(\vec{k}|\boldsymbol{c})p(\vec{k}'|\boldsymbol{c})Q(\vec{k},\vec{k}'|p(.|\boldsymbol{c})) = 1$.

A.2. Calculation of ϕ_2

In order to calculate (A.3) we first work out the following quantity, which will then have to be evaluated at $(\vec{k}_1,\ldots,\vec{k}_N)=(\vec{k}_1(\boldsymbol{c}),\ldots,\vec{k}_N(\boldsymbol{c}))$:

$$\begin{aligned}\tilde{\phi}_2(\vec{k}_1,\ldots,\vec{k}_N|Q) &= \frac{1}{N}\log Z(\vec{k}_1,\ldots,\vec{k}_N,Q)\\ &= \frac{1}{N}\log\sum_{\boldsymbol{c}}\prod_{i\neq j}\left[\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})\delta_{c_{ij},1}+\left(1-\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})\right)\delta_{c_{ij},0}\right]\\ &\quad\times\prod_i\delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}\\ &= \frac{1}{N}\log\int_{-\pi}^{\pi}\prod_i\left[\frac{\mathrm{d}\omega_i\,\mathrm{d}\psi_i}{4\pi^2}\,\mathrm{e}^{\mathrm{i}[\omega_i k_i^{\mathrm{in}}+\psi_i k_i^{\mathrm{out}}]}\right]L(\boldsymbol{\omega},\boldsymbol{\psi}|\bar{p},Q)\end{aligned}\tag{A.11}$$

with

$$\begin{aligned}L(\boldsymbol{\omega},\boldsymbol{\psi}|\bar{p},Q) &= \prod_{i\neq j}\left[1+\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|\bar{p})[\mathrm{e}^{-\mathrm{i}(\omega_i+\psi_j)}-1]\right]\\ &= \exp\left[\frac{\bar{k}}{N}\sum_{ij}Q(\vec{k}_i,\vec{k}_j|\bar{p})[\mathrm{e}^{-\mathrm{i}(\omega_i+\psi_j)}-1]+\mathcal{O}(N^0)\right].\end{aligned}\tag{A.12}$$

Upon introducing $R(\vec{k}|\boldsymbol{\omega})=N^{-1}\sum_i\delta_{\vec{k},\vec{k}_i}\mathrm{e}^{-\mathrm{i}\omega_i}$ and $S(\vec{k}|\boldsymbol{\psi})=N^{-1}\sum_i\delta_{\vec{k},\vec{k}_i}\mathrm{e}^{-\mathrm{i}\psi_i}$, and inserting $\int\prod_{\vec{k}}\big[\mathrm{d}R(\vec{k})\,\mathrm{d}S(\vec{k})\delta[R(\vec{k})-R(\vec{k}|\boldsymbol{\omega})]\delta[S(\vec{k})-S(\vec{k}|\boldsymbol{\psi})]\big]$ with δ-functions written in integral form, we can write

$$\begin{aligned}L(\boldsymbol{\omega},\boldsymbol{\psi}|\bar{p},Q) &= \int\prod_{\vec{k}}\left[\frac{\mathrm{d}R(\vec{k})\,\mathrm{d}\hat{R}(\vec{k})\,\mathrm{d}S(\vec{k})\,\mathrm{d}\hat{S}(\vec{k})}{4\pi^2/N^2}\,\mathrm{e}^{\mathrm{i}N[\hat{R}(\vec{k})R(\vec{k})+\hat{S}(\vec{k})S(\vec{k})]}\right]\mathrm{e}^{\mathcal{O}(N^0)}\\ &\quad\times\mathrm{e}^{-\mathrm{i}\sum_i[\hat{R}(\vec{k}_i)\mathrm{e}^{-\mathrm{i}\omega_i}+\hat{S}(\vec{k}_i)\mathrm{e}^{-\mathrm{i}\psi_i}]+\bar{k}N\sum_{\vec{k},\vec{k}'}R(\vec{k})Q(\vec{k},\vec{k}'|\bar{p})S(\vec{k}')-\bar{k}N}.\end{aligned}\tag{A.13}$$

Substituting this back into $\tilde{\phi}_2$, and using the law of large numbers, then gives

$$\tilde{\phi}_2(\ldots)=\frac{1}{N}\log\int\prod_{\vec{k}}[\mathrm{d}R(\vec{k})\,\mathrm{d}\hat{R}(\vec{k})\,\mathrm{d}S(\vec{k})\,\mathrm{d}\hat{S}(\vec{k})]\mathrm{e}^{N\Psi[R,\hat{R},S,\hat{S}|\bar{p},Q]+\mathcal{O}(\log N)}\tag{A.14}$$

where

$$\begin{aligned}\Psi[R,\hat{R},S,\hat{S}|\bar{p},Q] &= \mathrm{i}\sum_{\vec{k}}[\hat{R}(\vec{k})R(\vec{k})+\hat{S}(\vec{k})S(\vec{k})]+\bar{k}\sum_{\vec{k},\vec{k}'}R(\vec{k})Q(\vec{k},\vec{k}'|\bar{p})S(\vec{k}')-\bar{k}\\ &\quad+\sum_{\vec{k}}\bar{p}(\vec{k})\left\{\log\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{2\pi}\mathrm{e}^{\mathrm{i}[\omega k^{\mathrm{in}}-\hat{R}(\vec{k})\mathrm{e}^{-\mathrm{i}\omega}]}+\log\int_{-\pi}^{\pi}\frac{\mathrm{d}\psi}{2\pi}\mathrm{e}^{\mathrm{i}[\psi k^{\mathrm{out}}-\hat{S}(\vec{k})\mathrm{e}^{-\mathrm{i}\psi}]}\right\}.\end{aligned}\tag{A.15}$$

After doing the remaining integrals over ω and ψ we get

$$\begin{aligned}\Psi[R,\hat{R},S,\hat{S}|\bar{p},Q] &= \mathrm{i}\sum_{\vec{k}}[\hat{R}(\vec{k})R(\vec{k})+\hat{S}(\vec{k})S(\vec{k})]+\bar{k}\sum_{\vec{k},\vec{k}'}R(\vec{k})Q(\vec{k},\vec{k}'|\bar{p})S(\vec{k}')-\bar{k}\\ &\quad+\sum_{\vec{k}}\bar{p}(\vec{k})k^{\mathrm{in}}\log[-\mathrm{i}\hat{R}(\vec{k})]+\sum_{\vec{k}}\bar{p}(\vec{k})k^{\mathrm{out}}\log[-\mathrm{i}\hat{S}(\vec{k})]\\ &\quad-\sum_{\vec{k}}\bar{p}(\vec{k})\log(k^{\mathrm{in}}!k^{\mathrm{out}}!)\end{aligned}\tag{A.16}$$

For $N \to \infty$ the quantity $\tilde{\phi}_2(\vec{k}_1, \ldots, \vec{k}_N|Q)$ can be evaluated by steepest descent, giving $\lim_{N\to\infty} \tilde{\phi}_2(\ldots) = \text{extr}_{R,\hat{R},S,\hat{S}} \Psi[R, \hat{R}, S, \hat{S}|\bar{p}, Q]$. Differentiation of Ψ gives the following saddle-point equations:

$$-\mathrm{i}\hat{R}(\vec{k}) = \bar{p}(\vec{k})k^{\text{in}}/R(\vec{k}) = \bar{k}\sum_{\vec{k}'} Q(\vec{k}, \vec{k}'|\bar{p})S(\vec{k}') \tag{A.17}$$

$$-\mathrm{i}\hat{S}(\vec{k}) = \bar{p}(\vec{k})k^{\text{out}}/S(\vec{k}) = \bar{k}\sum_{\vec{k}'} Q(\vec{k}', \vec{k}|\bar{p})R(\vec{k}'). \tag{A.18}$$

At the saddle-point we deduce that $\sum_{\vec{k},\vec{k}'} R(\vec{k})Q(\vec{k}, \vec{k}'|\bar{p})S(\vec{k}') = 1$, and that

$$\Psi[R, \hat{R}, S, \hat{S}|\bar{p}, Q] = -2\bar{k} - \sum_{\vec{k}} \bar{p}(\vec{k}) \log(k^{\text{in}}!k^{\text{out}}!) + \sum_{\vec{k}} \bar{p}(\vec{k})k^{\text{in}} \log\left[\frac{\bar{p}(\vec{k})k^{\text{in}}}{R(\vec{k}|\bar{p}, Q)}\right] + \sum_{\vec{k}} \bar{p}(\vec{k})k^{\text{out}} \log\left[\frac{\bar{p}(\vec{k})k^{\text{out}}}{S(\vec{k}|\bar{p}, Q)}\right] \tag{A.19}$$

in which the functions $R(\vec{k}|\bar{p}, Q)$ and $S(\vec{k}|\bar{p}, Q)$ are the solutions of

$$R(\vec{k}) = \frac{\bar{p}(\vec{k})k^{\text{in}}}{\bar{k}\sum_{\vec{k}'} Q(\vec{k}, \vec{k}'|\bar{p})S(\vec{k}')}, \qquad S(\vec{k}) = \frac{\bar{p}(\vec{k})k^{\text{out}}}{\bar{k}\sum_{\vec{k}'} Q(\vec{k}', \vec{k}|\bar{p})R(\vec{k}')}. \tag{A.20}$$

Finally, the quantity (A.3) we aim to calculate is defined as the value of $\tilde{\phi}_2(\ldots)$ upon substituting $(\vec{k}_1, \ldots, \vec{k}^N) \to (\vec{k}_1(\boldsymbol{c}), \ldots, \vec{k}_N(\boldsymbol{c}))$. The only occurrences of the sequence $(\vec{k}_1, \ldots, \vec{k}_N)$ in the formula (A.19) are in the values of $\bar{p}(\vec{k})$ and $\bar{k}$, so we obtain $\phi_2(\boldsymbol{c}|Q)$ by making in (A.19) the substitutions $\bar{p}(\vec{k}) \to p(\vec{k}|\boldsymbol{c})$ and $\bar{k} \to \bar{k}(\boldsymbol{c})$. We conclude that

$$\phi_2(\boldsymbol{c}|Q) = -2\tilde{k} - \sum_{\vec{k}} \tilde{p}(\vec{k}) \log(k^{\text{in}}!k^{\text{out}}!) + \sum_{\vec{k}} \tilde{p}(\vec{k})k^{\text{in}} \log\left[\frac{\tilde{p}(\vec{k})k^{\text{in}}}{R(\vec{k}|\tilde{p}, Q)}\right] + \sum_{\vec{k}} \tilde{p}(\vec{k})k^{\text{out}} \log\left[\frac{\tilde{p}(\vec{k})k^{\text{out}}}{S(\vec{k}|\tilde{p}, Q)}\right] \tag{A.21}$$

in which $\tilde{p}(\vec{k}) = p(\vec{k}|\boldsymbol{c})$ and $\tilde{k} = \bar{k}(\boldsymbol{c})$, and in which $R(\vec{k}|\tilde{p}, Q)$ and $S(\vec{k}|\tilde{p}, Q)$ are the solutions of

$$R(\vec{k}) = \frac{\tilde{p}(\vec{k})k^{\text{in}}}{\tilde{k}\sum_{\vec{k}'} Q(\vec{k}, \vec{k}'|\tilde{p})S(\vec{k}')}, \qquad S(\vec{k}) = \frac{\tilde{p}(\vec{k})k^{\text{out}}}{\tilde{k}\sum_{\vec{k}'} Q(\vec{k}', \vec{k}|\tilde{p})R(\vec{k}')}. \tag{A.22}$$

A.3. Final analytical expression for Ω

The intermediate results (A.6, A.21) can now be substituted back into expression (A.1), which gives a formula that is seen to depend on $\boldsymbol{c}$ only via $W(\vec{k}, \vec{k}'|\boldsymbol{c})$ and $p(\vec{k}|\boldsymbol{c})$:

$$\Omega(\boldsymbol{c}|p, Q) = \left\{ \sum_{\vec{k}} \tilde{p}(\vec{k}) \log p(\vec{k}) + \tilde{k}[1+\log[\tilde{k}/N]] + \sum_{\vec{k}} \tilde{p}(\vec{k}) \log(k^{\text{in}}!k^{\text{out}}!) - \sum_{\vec{k}} \tilde{p}(\vec{k})k^{\text{in}} \log\left[\frac{\tilde{p}(\vec{k})k^{\text{in}}}{R(\vec{k}|\tilde{p}, Q)}\right] - \sum_{\vec{k}} \tilde{p}(\vec{k})k^{\text{out}} \log\left[\frac{\tilde{p}(\vec{k})k^{\text{out}}}{S(\vec{k}|\tilde{p}, Q)}\right] + \tilde{k}\sum_{\vec{k},\vec{k}'} \tilde{W}(\vec{k}, \vec{k}') \log Q(\vec{k}, \vec{k}'|\tilde{p}) \right\}_{\tilde{W}(\vec{k},\vec{k}')=W(\vec{k},\vec{k}'|\boldsymbol{c}), \tilde{p}(\vec{k})=p(\vec{k}|\boldsymbol{c})} + \varepsilon_N \tag{A.23}$$

with $\lim_{N\to\infty}\varepsilon_N = 0$, $\tilde{k} = \sum_{\vec{k}} k^{\rm in}\tilde{p}(\vec{k}) = \sum_{\vec{k}} k^{\rm out}\tilde{p}(\vec{k})$, and with the two functions $S(\vec{k}|\tilde{p}, Q)$ and $R(\vec{k}|\tilde{p}, Q)$ to be extracted from (3.7).

Appendix B. Calculation of the kernel W

For large N the kernel $W(\vec{k},\vec{k}') = (N\bar{k})^{-1}\sum_{ij} c_{ij}\delta_{\vec{k},\vec{k}_i}\delta_{\vec{k}',\vec{k}_j}$ will be self-averaging in the ensemble (3.1), i.e. with probability one any graph generated randomly according to (3.1) will exhibit the same kernel, modulo finite size effects. Thus we may for $N\to\infty$ calculate $W(\vec{k},\vec{k}')$ as an average over the ensemble (3.1):

$$
\begin{aligned}
W(\vec{k},\vec{k}') &= \frac{1}{N\bar{k}}\sum_{r\neq s}\sum_{\vec{k}_1\ldots\vec{k}_N}\frac{\delta_{\vec{k},\vec{k}_r}\delta_{\vec{k}',\vec{k}_s}\prod_i p(\vec{k}_i)}{Z(\vec{k}_1\ldots\vec{k}_N,Q)}\sum_{\boldsymbol{c}}\left[\prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}\right]c_{rs}\\
&\quad\times\prod_{i\neq j}\left[\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|p)\delta_{c_{ij},1}+\Big(1-\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|p)\Big)\delta_{c_{ij},0}\right]\\
&= \frac{1}{N^2}\sum_{r\neq s}\sum_{\vec{k}_1\ldots\vec{k}_N}\frac{\delta_{\vec{k},\vec{k}_r}\delta_{\vec{k}',\vec{k}_s}\prod_i p(\vec{k}_i)}{Z(\vec{k}_1\ldots\vec{k}_N,Q)}\int_{-\pi}^{\pi}\prod_i\left[\frac{{\rm d}\omega_i\,{\rm d}\psi_i}{4\pi^2}{\rm e}^{{\rm i}[\omega_i k_i^{\rm in}+\psi_i k_i^{\rm out}]}\right]\\
&\quad\times Q(\vec{k}_r,\vec{k}_s|p)[{\rm e}^{-{\rm i}(\omega_r+\psi_s)}\left[1+\mathcal{O}\left(\frac{1}{N}\right)\right]\\
&\quad\times\prod_{i\neq j}\left[1+\frac{\bar{k}}{N}Q(\vec{k}_i,\vec{k}_j|p)[{\rm e}^{-{\rm i}(\omega_i+\psi_j)}-1]\right]\\
&= Q(\vec{k},\vec{k}'|p)\sum_{\vec{k}_1\ldots\vec{k}_N}\frac{\prod_i p(\vec{k}_i)}{Z(\vec{k}_1\ldots\vec{k}_N,Q)}\int_{-\pi}^{\pi}\prod_i\left[\frac{{\rm d}\omega_i\,{\rm d}\psi_i}{4\pi^2}{\rm e}^{{\rm i}[\omega_i k_i^{\rm in}+\psi_i k_i^{\rm out}]}\right]\\
&\quad\times L(\boldsymbol{\omega},\boldsymbol{\psi}|p,Q)\left(\frac{1}{N}\sum_r\delta_{\vec{k},\vec{k}_r}{\rm e}^{-{\rm i}\omega_r}\right)\left(\frac{1}{N}\sum_s\delta_{\vec{k}',\vec{k}_s}{\rm e}^{-{\rm i}\psi_s}\right)\left[1+\mathcal{O}\left(\frac{1}{N}\right)\right]\\
&= Q(\vec{k},\vec{k}'|p)\sum_{\vec{k}_1\ldots\vec{k}_N}\frac{[1+\mathcal{O}\left(\frac{1}{N}\right)]\prod_i p(\vec{k}_i)}{Z(\vec{k}_1\ldots\vec{k}_N,Q)}\int\prod_{\vec{q}}\left[\frac{{\rm d}R(\vec{q})\,{\rm d}\hat{R}(\vec{q})\,{\rm d}S(\vec{q})\,{\rm d}\hat{S}(\vec{q})}{4\pi^2/N^2}\right]\\
&\quad\times{\rm e}^{{\rm i}N\sum_{\vec{q}}[\hat{R}(\vec{q})R(\vec{q})+\hat{S}(\vec{q})S(\vec{q})]+\bar{k}N\sum_{\vec{q},\vec{q}'}Q(\vec{q},\vec{q}'|p)R(\vec{q})S(\vec{q}')-\bar{k}N+\mathcal{O}(N^0)}\\
&\quad\times R(\vec{k})S(\vec{k}')\prod_i\int_{-\pi}^{\pi}\left[\frac{{\rm d}\omega\,{\rm d}\psi}{4\pi^2}{\rm e}^{{\rm i}\omega k_i^{\rm in}+{\rm i}\psi k_i^{\rm out}-{\rm i}\hat{R}(\vec{k}_i){\rm e}^{-{\rm i}\omega}-{\rm i}\hat{S}(\vec{k}_i){\rm e}^{-{\rm i}\psi}}\right].
\end{aligned}
\tag{B.1}
$$

We now write $Z(\vec{k}_1\ldots\vec{k}_N,Q)$ also as an integral over order parameters, as in our earlier derivation of (A.19), but noting that now the relevant degree distribution is that of our ensemble (3.1), i.e. $p(\vec{k})$ instead of $\bar{p}(\vec{k})$. This gives

$$
\begin{aligned}
W(\vec{k},\vec{k}') &= \left[1+\mathcal{O}\left(\frac{1}{N}\right)\right]Q(\vec{k},\vec{k}')\sum_{\vec{k}_1\ldots\vec{k}_N}\prod_i p(\vec{k}_i)\\
&\quad\times\frac{\int\prod_{\vec{q}}{\rm d}R(\vec{q})\,{\rm d}\hat{R}(\vec{q})\,{\rm d}S(\vec{q})\,{\rm d}\hat{S}(\vec{q})\,{\rm e}^{N\Psi[R,\hat{R},S,\hat{S}|p,Q]+\mathcal{O}(\log N)}R(\vec{k})S(\vec{k}')}{\int\prod_{\vec{q}}{\rm d}R(\vec{q})\,{\rm d}\hat{R}(\vec{q})\,{\rm d}S(\vec{q})\,{\rm d}\hat{S}(\vec{q})\,{\rm e}^{N\Psi[R,\hat{R},S,\hat{S}|p,Q]+\mathcal{O}(\log N)}},
\end{aligned}
\tag{B.2}
$$

where the non-extensive terms in the exponentials of numerator and denominator are fully identical, and with Ψ as defined in (A.15), modulo the replacement $\bar{p}\to p$. The summation

over degree sequences has now become obsolete, and for $N \to \infty$ we obtain

$$\lim_{N\to\infty} W(\vec{k}, \vec{k}') = R(\vec{k}|p, Q)Q(\vec{k}, \vec{k}'|p)S(\vec{k}'|p, Q) \tag{B.3}$$

in which $R(\vec{k}|p, Q)$ and $S(\vec{k}|p, Q)$ are to be solved from

$$R(\vec{k}) = \frac{p(\vec{k})k^{\text{in}}}{\bar{k}\sum_{\vec{k}'} Q(\vec{k}, \vec{k}'|p)S(\vec{k}')}, \qquad S(\vec{k}) = \frac{p(\vec{k})k^{\text{out}}}{\bar{k}\sum_{\vec{k}'} Q(\vec{k}', \vec{k}|p)R(\vec{k}')} \tag{B.4}$$

with the average degree of our ensemble, $\bar{k} = \sum_{\vec{k}} k^{\text{in}} p(\vec{k}) = \sum_{\vec{k}} k^{\text{out}} p(\vec{k})$.

References

[1] Annibale A, Coolen A C C, Fernandes L P, Fraternali F and Kleinjung J 2009 *J. Phys. A: Math. Gen.* **42** 485001
[2] Coolen A C C, Martino A D and Annibale A 2009 *J. Stat. Phys.* **136** 1035–67
[3] Fernandes L P, Annibale A, Kleinjung J, Coolen A C C and Fraternali F 2010 *PLoS ONE* **5** e12083
[4] Coolen A C C, Fraternali F, Annibale A, Fernandes L and Kleinjung J 2011 Modelling biological networks via tailored random graphs *Handbook of Statistical Systems* ed M Stumpf, D J Balding and M Girolami (Chichester: John Wiley and Sons Ltd)
[5] Memisević V, Milenković T and Pržulj N 2010 *J. Integrative Bioinformatics* **7** 120
[6] Albert R and Barabási A L 2002 *Rev. Mod. Phys.* **74** 47–97
[7] Dorogovtsev S N, Goltsev A V and Mendes J F F 2008 *Rev. Mod. Phys.* **80** 1275–335
[8] Bianconi G, Coolen A C C and Vicente C J P 2008 *Phys. Rev.* E **78** 016114
[9] Hughes T R *et al* 2000 *Cell* **102** 109–26
[10] Harbison C T *et al* 2004 *Nature* **431** 99–104
[11] Rung J, Schlitt T, Brazma A, Freivalds K and Vilo J 2002 *Bioinformatics (Oxford, England)* **18** (Suppl. 2) S202–10
[12] Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E and Brazma A 2003 *Genome Res.* **13** 2568–76