IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **42** (2009) 485001 (25pp) doi:10.1088/1751-8113/42/48/485001

Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure

A Annibale[1], **A C C Coolen**[1,2], **L P Fernandes**[2], **F Fraternali**[2] **and J Kleinjung**[3]

[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 Hunt's House, London SE1 1UL, UK
[3] MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

E-mail: alessia.annibale@kcl.ac.uk and ton.coolen@kcl.ac.uk

Received 14 August 2009, in final form 6 October 2009
Published 11 November 2009
Online at stacks.iop.org/JPhysA/42/485001

Abstract
We study the tailoring of structured random graph ensembles to real networks, with the objective of generating precise and practical mathematical tools for quantifying and comparing network topologies macroscopically, beyond the level of degree statistics. Our family of ensembles can produce graphs with any prescribed degree distribution and any degree–degree correlation function; its control parameters can be calculated fully analytically, and as a result we can calculate (asymptotically) formulae for entropies and complexities and for information-theoretic distances between networks, expressed directly and explicitly in terms of their measured degree distribution and degree correlations.

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

1. Introduction

In the study of natural or synthetic signaling networks, one of the key questions is how network structure relates to the execution of the process which it supports. This is especially true in systems biology, where, for instance, our understanding of how the structure of protein–protein interaction networks (PPIN) relates to their biological functionality is vital in the design of a new generation of intelligent and personalized medical interventions. In recent years, high-throughput proteomics has allowed for the drafting of large PPIN data sets, for different organisms and with different experimental techniques and degrees of accuracy. With this accumulation of information, we now face the challenge of analysing these data from a complex network's perspective, and using them optimally in order to increase our understanding of how PPIN control the functioning of cells, both in healthy and in diseased conditions. A prerequisite for achieving this is the availability of precise mathematical tools with which to quantify topological structure in large observed networks, to compare network instances and distinguish between meaningful and 'random' structural features. These tools have to be both systematic, i.e. with a sound statistical or information-theoretic basis, and practical, i.e. preferably formulated in terms of explicit formulae as opposed to tedious numerical simulations.

Many quantities have been proposed for characterizing the structure of networks, such as degree distributions [1], degree sequences [2], degree correlations [3] and assortativity [4], clustering coefficients [5] and community structures [6]. To assess the relevance of an observed topological feature in a network, a common strategy is to compare it against similar observations in so-called null models, defined as randomized versions of the original network which retain some features of the original one. The choice of the topological features to be conserved in the randomized models was mostly limited to degree distributions and degree sequences. Such null models were used to assess the statistical relevance of network motifs in real networks, namely patterns which were observed significantly more often in the real networks than in their randomized counterparts [7–9]. Whether any such proposed motif is indeed functionally important and/or represent (evolutionary) arisen principles is however not obvious; topological deviations from randomized networks could also be merely irrelevant consequences of some neglected structural property of the network, i.e. the result of an inappropriate null hypothesis rather than of a distinctive feature of the process [10, 11]. The definition and generation of good null models for benchmarking topological measures of real world graphs (and the dynamical processes which they enable) is a nontrivial issue. Similarly, in comparing observed networks (which, as a result of experimental noise, will usually not even have identical nodes), one would seek to focus on the values of macroscopic topological observables, and know the typical properties of networks with the observed features.

In recent years there have been efforts to define and generate random graphs whose topological features can be controlled and tailored to experimentally observed networks. In [12] a parametrized random graph ensemble was defined where graphs have a prescribed degree sequence, and links are drawn in a way that allows for preferential attachment on the basis of arbitrary two-degree kernels. In this paper we generalize the definition of this ensemble, and show that it can be tailored asymptotically to the generation of graphs with any prescribed degree distribution and any prescribed degree correlation function (and that it is a maximum entropy ensemble, given the degree correlations). Moreover, in spite of its parameter space being in principle infinitely large, in contrast to most random graph ensembles used to mimic real networks, we can derive explicit analytical formulae for the parameters of the ensemble, to leading order in system size, expressed directly in terms of the observed characteristics of the network given. Graphs from this ensemble are thus ideally suited to be used as either proxies or null models for observed networks, depending on the question to be answered.

Statistical mechanics approaches have been proposed to quantify the information content of network structures. Especially the (Shannon or Boltzmann) entropy has been instrumental in characterizing the complexity of network ensembles [13–15]. Here, the crucial availability of analytical expressions for the parameters of our ensemble will enable us to derive explicit formulae, in the thermodynamic limit (based on combinatorial and saddle-point arguments), for our ensemble's Shannon entropy, and hence also for the complexity of its typical graphs. These formulae are compact and transparent, and expressed solely and explicitly in terms of the degree distribution and the degree correlations that our ensemble is targeting. Finally, along similar lines we can obtain an information theoretic distance between networks, again expressed solely in terms of their degree distributions and degree correlations. A companion paper [16] will be devoted to large scale applications to PPIN data of these complexity and distance measures; here we focus on their mathematical derivation. Although there is no need for numerical sampling in our derivations (all results can be obtained analytically), we note that exact algorithms for generating random graphs from the proposed ensemble exist [17].

2. Definitions and properties of network topology characterizations

2.1. Networks, degree distributions and degree correlation functions

We study networks (or graphs) of N nodes (or vertices), labelled by Roman indices $i, j, \ldots,$ where every vertex can be connected to other vertices by undirected links (or 'edges'). The microscopic structure of such a network is defined in full by an $N \times N$ matrix of binary variables $c_{ij} \in \{0, 1\}$, where the nodes i and j are connected by a link if and only if $c_{ij} = 1$. We define $c_{ij} = c_{ji}$ and $c_{ii} = 0$ for all (i, j), and we abbreviate $\mathbf{c} = \{c_{ij}\}$. Henceforth, unless indicated otherwise, any summation over Roman indices will always run over the set $\{1, \ldots, N\}$.

A standard way of characterizing the topology of a network $\mathbf{c}$, as e.g. observed in a biological or physical system under study, is to measure for each vertex i the degree $k_i(\mathbf{c}) = \sum_j c_{ij}$, the number of links to this vertex. From these numbers then follow the empirical degree distribution $p(k|\mathbf{c})$ and the observed average connectivity $\bar{k}(\mathbf{c})$:

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

(using the Kronecker δ-symbol for $n, m \in \mathbb{N}$, defined as $\delta_{nm} = 1$ for $n = m$ and $\delta_{nm} = 0$ otherwise). Objects such as $p(k|\mathbf{c})$ have the advantage of being macroscopic in nature, allowing for size-independent characterization of network topologies, and for comparing networks that differ in size. However, networks with the same degree distribution can still differ profoundly in their microscopic structures. We need observables that capture additional topological information, in order to discriminate between different networks with the same degree distribution (1).

To construct macroscopic observables that quantify network topology beyond the level of degree statistics, it is natural to consider how the likelihood for two nodes of a network $\mathbf{c}$ to be connected depends on their degrees, which is measured by the degree correlation function

$$\tilde{\Pi}(k, k'|\mathbf{c}) = \frac{\mathcal{P}[\text{conn}|\mathbf{c}, k, k']}{\mathcal{P}[\text{conn}|\mathbf{c}]}. \tag{2}$$

Here $\mathcal{P}[\text{conn}|\mathbf{c}, k, k']$ is the probability for two randomly drawn nodes with degrees (k, k') to be connected, and $\mathcal{P}[\text{conn}|\mathbf{c}]$ is the overall probability for two randomly drawn nodes to be connected, irrespectively of their degrees, namely

$$\mathcal{P}[\text{conn}|\mathbf{c}, k, k'] = \frac{\sum_{i\neq j} c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}}{\sum_{i\neq j}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}} \tag{3}$$

$$\mathcal{P}[\text{conn}|\mathbf{c}] = \frac{1}{N(N-1)}\sum_{i\neq j} c_{ij} = \frac{\bar{k}(\mathbf{c})}{N-1}. \tag{4}$$

By definition, $\tilde{\Pi}(k, k'|\mathbf{c})$ is symmetric under exchanging k and k'. For simple networks $\mathbf{c}_0$, with some degree distribution $p(k)$ but without any micro-structure beyond that required by $p(k)$,[4] it is known (see e.g. [18] and references therein) that in the limit $N \to \infty$ one finds

$$\tilde{\Pi}(k, k'|\mathbf{c}_0) = kk'/\overline{k}^2(\mathbf{c}). \tag{5}$$

It follows that those topological properties of a given (large) network $\mathbf{c}$, that manifest themselves at the level of degree correlations and cannot be attributed simply to its degree statistics, can be quantified by a deviation from the simple law (5); see also [7, 19, 20]. One is therefore led in a natural way to the introduction of the *relative* degree correlations

$$\Pi(k, k'|\mathbf{c}) = \frac{\tilde{\Pi}(k, k'|\mathbf{c})}{\tilde{\Pi}(k, k'|\mathbf{c}_0)} = \frac{\mathcal{P}[\text{conn}|\mathbf{c}, k, k']}{\mathcal{P}[\text{conn}|\mathbf{c}]} \frac{\overline{k}^2(\mathbf{c})}{kk'}. \tag{6}$$

By definition, $\Pi(k, k'|\mathbf{c}_0) = 1$ for sufficiently large simple networks $\mathbf{c}_0$, whereas any statistically relevant deviation from $\Pi(k, k'|\mathbf{c}) = 1$ signals the presence in network $\mathbf{c}$ of underlying criteria for connecting nodes beyond its degrees. Just like $p(k|\mathbf{c})$, $\Pi(k, k'|\mathbf{c})$ is again a *macroscopic* observable that can be measured directly and at low computation cost. It is therefore a natural tool for quantifying and comparing network structures beyond the level of degree statistics.

2.2. Properties of the relative degree correlation function

To prepare the ground for proving some asymptotic mathematical properties of the relative degree correlation function $\Pi(k, k'|\mathbf{c})$, we first simplify the denominator of (3):

$$\begin{aligned}\sum_{i\neq j} \delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})} &= \sum_{ij} \delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})} - \delta_{kk'} \sum_i \delta_{k,k_i(\mathbf{c})} \\ &= N^2[p(k|\mathbf{c})p(k'|\mathbf{c}) - N^{-1}\delta_{kk'}p(k|\mathbf{c})].\end{aligned} \tag{7}$$

Upon inserting the result of (4) together with (3) into (6) we then find that

$$\Pi(k, k'|\mathbf{c}) = \frac{N^{-1}\sum_{i\neq j} c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}}{[p(k|\mathbf{c}) - N^{-1}\delta_{kk'}]p(k'|\mathbf{c})} \frac{\overline{k}(\mathbf{c})}{kk'}(1 - N^{-1}), \tag{8}$$

and hence, using $c_{ii} = 0$ for all i,

$$\lim_{N\to\infty} \Pi(k, k'|\mathbf{c}) = \lim_{N\to\infty} \frac{\overline{k}(\mathbf{c})\sum_{ij} c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}}{Np(k|\mathbf{c})p(k'|\mathbf{c})kk'}. \tag{9}$$

We are now in a position to establish three identities obeyed by $\Pi(k, k'|\mathbf{c})$. The first two of these, namely (10), (12), are the main ones; they are used frequently in mathematical manipulations of subsequent sections. The third provides the physical intuition behind (10), (12). It is assumed implicitly in all proofs that $\overline{k}(\mathbf{c})$ remains finite for $N \to \infty$ and that the limits $N \to \infty$ exist.

- *Linear constraints.*

$$\forall k \in \mathbb{N}: \quad \lim_{N\to\infty} \sum_{k'} \frac{k'p(k'|\mathbf{c})}{\overline{k}(\mathbf{c})} \Pi(k, k'|\mathbf{c}) = 1. \tag{10}$$

These are easily verified for simple graphs $\mathbf{c}_0$, for which $\Pi(k, k'|\mathbf{c}_0) = 1 \forall k, k'$. However, they turn out to hold for *any* graph $\mathbf{c}$, as can be proven using (9) as follows:

$$\begin{aligned}\lim_{N\to\infty} \sum_{k'} \frac{k'p(k'|\mathbf{c})}{\overline{k}(\mathbf{c})} \Pi(k, k'|\mathbf{c}) &= \lim_{N\to\infty} \frac{\sum_{k'}\sum_{ij} c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}}{Np(k|\mathbf{c})k} \\ &= \lim_{N\to\infty} \frac{\sum_i k_i(\mathbf{c})\delta_{k,k_i(\mathbf{c})}}{Np(k|\mathbf{c})k} = 1.\end{aligned} \tag{11}$$

[4] In section 2 we will give a precise and more general mathematical definition of 'simple networks', relative to some imposed macroscopic feature such as the degree distribution $p(k)$.

- *Normalization.*

$$\lim_{N\to\infty}\sum_{kk'}\frac{kp(k|\mathbf{c})}{\overline{k}(\mathbf{c})}\frac{k'p(k'|\mathbf{c})}{\overline{k}(\mathbf{c})}\Pi(k,k'|\mathbf{c})=1. \tag{12}$$

  This follows directly from (10) upon multiplying both sides by $kp(k)/\langle k\rangle$, followed by summation over all k.

- *Interpretation of the linear constraints.* The LHS of (10) can be rewritten as

$$\begin{aligned}\lim_{N\to\infty}\sum_{k'}\frac{k'p(k'|\mathbf{c})}{\overline{k}(\mathbf{c})}\Pi(k,k'|\mathbf{c}) &= \lim_{N\to\infty}\frac{\overline{k}(\mathbf{c})}{k}\sum_{k'}p(k'|\mathbf{c})\frac{\mathcal{P}[\text{conn}|\mathbf{c},k,k']}{\mathcal{P}[\text{conn}|\mathbf{c}]}\\ &= \lim_{N\to\infty}\frac{N}{k}\mathcal{P}[\text{conn}|\mathbf{c},k]\end{aligned} \tag{13}$$

  where we used (6) and (4), and $\mathcal{P}[\text{conn}|\mathbf{c},k]$ is the marginal probability of $\mathcal{P}[\text{conn}|\mathbf{c},k,k']$, which represents the probability that two randomly drawn nodes, one having degree k, are connected. We conclude that our first (proven) identity (10) boils down to the claim that for large N one has $\mathcal{P}[\text{conn}|\mathbf{c},k]=k/N$ (modulo irrelevant orders in N), which is easily understood.

We end this section with two further observations. First, the relations (10) involve the degree distribution, so one must expect that the possible values for $\Pi(k,k'|\mathbf{c})$ are dependent upon (or constrained by) $p(k|\mathbf{c})$. Second, several other useful properties of the kernel $\Pi(k,k'|\mathbf{c})$ can be extracted from (10). For instance, the only separable kernel Π is $\Pi(k,k'|\mathbf{c})=1$ for all (k,k'); a separable kernel is of the form $\Pi(k,k'|\mathbf{c})=G(k|\mathbf{c})G(k'|\mathbf{c})$ for some function $G(k|\mathbf{c})$ (Π being symmetric), and insertion of this form into (10) leads immediately to $G(k|\mathbf{c})=1$ for all $k,\mathbf{c}$.

3. Random graphs with controlled macroscopic structure

3.1. Definition of the random graph ensembles

To study the signalling properties of real-world networks, or generate 'null models' to assess the relevance of observed topological features, one needs random graph ensembles in which one can control the topological characteristics one is interested in and 'tune' these to match the characteristics of the observed networks. Most ensembles studied in the literature so far have focused on producing graphs with controlled degree statistics. The suggestion that (6) can be used for *identifying* network complexity beyond degree statistics goes back at least to [7, 18–20]. In contrast to these earlier studies, which were mostly limited to measuring (6) for real networks, here we take further mathematical steps that will allow us to use (6) as a systematic tool for *quantifying* complexity and distances in network structure beyond degree statistics. This requires generating random graphs in which we can control at will both the degree distribution $p(k)$ and the relative degree correlations $\Pi(k,k')$.

It will turn out that we can achieve our objectives with the following random graph ensembles, in which all degrees k_i are drawn randomly and independently from $p(k)$, and where in addition the edges are drawn in a way that allows for preferential attachment on the basis of an arbitrary symmetric function $Q(k,k')$ of the degrees of the two vertices concerned:

$$\text{Prob}(\mathbf{c}|p,Q)=\sum_{\mathbf{k}}\text{Prob}(\mathbf{c}|\mathbf{k},Q)\prod_i p(k_i) \tag{14}$$

$$\text{Prob}(\mathbf{c}|\mathbf{k},Q)=\frac{1}{\mathcal{Z}(\mathbf{k},Q)}\prod_{i<j}\left[\frac{\overline{k}}{N}Q(k_i,k_j)\delta_{c_{ij},1}+\left(1-\frac{\overline{k}}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right]\prod_i\delta_{k_i,k_i(\mathbf{c})}. \tag{15}$$

Here $\mathcal{Z}(\mathbf{k},Q)$ is a normalization constant that ensures $\sum_{\mathbf{c}}\text{Prob}(\mathbf{c}|\mathbf{k},Q)=1$ for all $(\mathbf{k},Q)$, $\overline{k}=N^{-1}\sum_i k_i$, and the function Q must obey $Q(k,k')\geqslant 0$ for all (k,k') and $N^{-2}\sum_{ij}Q(k_i,k_j)=1$. The ensemble (15) with prescribed degrees $\mathbf{k}=(k_1,\ldots,k_N)$ was defined and studied in [12, 21]. We note that in the above ensemble one will have $\overline{k}=\sum_k p(k)k+\mathcal{O}(N^{-1/2})$.

Upon making the simplest choice $Q(k,k')=1$ for all (k,k') one retrieves from (14) the 'flat' ensemble, where once the individual degrees are drawn randomly from $p(k)$, all graphs $\mathbf{c}$ with the prescribed degrees carry equal probability:

$$\text{Prob}(\mathbf{c})=\sum_{\mathbf{k}}\left[\prod_i p(k_i)\right]\frac{\prod_i\delta_{k_i,k_i(\mathbf{c})}}{\sum_{\mathbf{c}'}\prod_i\delta_{k_i,k_i(\mathbf{c}')}}. \tag{16}$$

This follows from the property that for $Q(k,k')=1$ the factor $\prod_{i<j}[\ldots\delta_{c_{ij},1}+\ldots\delta_{c_{ij},0}]$ in (14) depends on $\mathbf{c}$ via the degrees $\{k_i(\mathbf{c})\}$ only, and will consequently drop out of the measure (15):

$$\begin{aligned}\prod_{i<j}\left[\frac{\langle k\rangle}{N}\delta_{c_{ij},1}+\left(1-\frac{\langle k\rangle}{N}\right)\delta_{c_{ij},0}\right]&=\left(1-\frac{\overline{k}}{N}\right)^{\frac{1}{2}N(N-1)}\mathrm{e}^{\sum_{i<j}c_{ij}\log[\frac{\overline{k}}{N}(1-\frac{\overline{k}}{N})^{-1}]}\\&=\left(1-\frac{\overline{k}}{N}\right)^{\frac{1}{2}N(N-1)}\mathrm{e}^{\frac{1}{2}N\overline{k}\log[\frac{\overline{k}}{N}(1-\frac{\overline{k}}{N})^{-1}]}.\end{aligned} \tag{17}$$

3.2. Asymptotic properties of the ensembles

One should expect that macroscopic physical observables such as $p(k|\mathbf{c})$ (1) and $\Pi(k,k'|\mathbf{c})$ (8) are self-averaging, and can therefore be calculated, to leading order in N, in terms of their expectation values over the ensemble (14).[5] We should therefore find that each graph drawn from (14) will for sufficiently large N have as its degree distribution $p(k)$ and will have relative degree correlations identical to

$$\begin{aligned}\Pi(k,k')&=\lim_{N\to\infty}\sum_{\mathbf{c}}\text{Prob}(\mathbf{c}|p,Q)\Pi(k,k'|\mathbf{c})\\&=\frac{\langle k\rangle}{p(k)p(k')kk'}\lim_{N\to\infty}\frac{1}{N}\sum_{\mathbf{c}}\text{Prob}(\mathbf{c}|p,Q)\sum_{ij}c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}.\end{aligned} \tag{18}$$

It turns out that (18) can be calculated analytically, and expressed in terms of $p(k)$ and $Q(k,k')$. The first published result related to this connection in an appendix of [12] was unfortunately subject to an error; see the corrigendum [21] for the correct relation as given below, of which the actual derivation is given in appendix A of this present paper:

$$\Pi(k,k')=\frac{Q(k,k')}{F(k|Q)F(k'|Q)} \tag{19}$$

where the function $F(k|Q)$ is calculated self-consistently, for any $Q(k,k')$, as the solution of

$$\forall k:\quad F(k|Q)=\frac{1}{\langle k\rangle}\sum_{k'}p(k')k'\frac{Q(k,k')}{F(k'|Q)}. \tag{20}$$

It is satisfactory to observe, upon eliminating $Q(k,k')$ from (20) via (19), that (20) becomes identical to the set of relations (10) that we derived earlier for $\Pi(k,k')$ solely on the basis of the latter's microscopic definition. Clearly, (10) must indeed hold for every single graph of the

[5] Proving this self-averaging property explicitly for the ensemble (14) is trivial in the case of $p(k|\mathbf{c})$, and nontrivial but feasible in the case of $\Pi(k,k'|\mathbf{c})$.

ensemble (14), provided N is sufficiently large. On the other hand, for finite N a typical graph of the ensemble (14) will display deviations from (19) that are at least of order $\mathcal{O}(N^{-1})$ (the difference between definition (8) and its asymptotic form (9)), but possibly of order $\mathcal{O}(N^{-1/2})$ (the typical finite size corrections in empirical averages over $\mathcal{O}(N)$ independent samples).

Expression (19) also provides *en passant* the explicit proof that for graphs in which the only structure is that imposed by the degree sequence, namely those generated from (16) corresponding to $Q(k,k')=1$ for all (k,k'), one indeed finds $\Pi(k,k')=1$ for $N\to\infty$. Upon inserting $Q(k,k')=1$ into condition (20) we find that $F^2(k)=1$ for all k, upon which the desired result follows directly from (19).

Asymptotically (i.e. in leading relevant orders in N), the probabilities (14) to find graphs $\mathbf{c}$ with the correct degree statistics, i.e. with degrees drawn randomly from $p(k)$, depends on $\mathbf{c}$ via the degree distribution $p(k|\mathbf{c})$ and the kernel $\Pi(k,k'|\mathbf{c})$ only. To see this we study the following function for large N:

$$\begin{aligned}\Omega(\mathbf{c}|p,Q) &= -N^{-1}\log \text{Prob}(\mathbf{c}|p,Q)\\ &= -N^{-1}\log\sum_{\mathbf{k}}\prod_i[p(k_i)\delta_{k_i,k_i(\mathbf{c})}]\,\mathrm{e}^{-N\Omega(\mathbf{c}|\mathbf{k},Q)}\\ &= -\frac{1}{N}\sum_i \log p(k_i(\mathbf{c}))+\Omega(\mathbf{c}|\mathbf{k}(\mathbf{c}),Q).\end{aligned}\tag{21}$$

The leading order in $\Omega(\mathbf{c}|\mathbf{k},Q)=-N^{-1}\log p(\mathbf{c}|\mathbf{k},Q)$ was studied in [12]. If $\mathbf{k}(\mathbf{c})\neq\mathbf{k}$, one has $\Omega(\mathbf{c}|\mathbf{k},Q)=\infty$ (the degrees are imposed as strict constraints), whereas for $\mathbf{k}=\mathbf{k}(\mathbf{c})$ one has

$$\begin{aligned}\Omega(\mathbf{c}|\mathbf{k},Q) &= \frac{1}{2}\overline{k}\log N+\frac{1}{2}\overline{k}[\log\overline{k}-1]-N^{-1}\sum_i\log k_i!+N^{-1}\sum_i k_i\log F(k_i|Q)\\ &\quad -N^{-1}\sum_{i<j}c_{ij}\log Q(k_i,k_j)+\mathcal{O}(N^{-1})\\ &= \frac{1}{2}\overline{k}\log N+\frac{1}{2}\overline{k}[\log\overline{k}-1]-N^{-1}\sum_i\log k_i!+N^{-1}\sum_i k_i\log F(k_i|Q)\\ &\quad -\frac{1}{2}\sum_{kk'}\log Q(k,k')\frac{1}{N}\sum_{ij}c_{ij}\delta_{k,k_i}\delta_{k',k_j}+\mathcal{O}(N^{-1}),\end{aligned}\tag{22}$$

where $\overline{k}=N^{-1}\sum_i k_i$. We introduce the further short-hand $\tilde{p}(k)=N^{-1}\sum_i\delta_{k,k_i}$, as well as the notation $o(1)$ to denote finite size corrections that obey $\lim_{N\to\infty}o(1)=0$ (to determine the exact scaling with N of these corrections we would have to inspect e.g. the finite size corrections to (19)). We write the leading orders of (22) in terms of the kernel $\Pi(k,k'|\mathbf{c})$, using (9) and (19), and substituting into (19) the present degree distribution $\tilde{p}(k)$, and find

$$\begin{aligned}\Omega(\mathbf{c}|\mathbf{k},Q) &= \frac{1}{2}\overline{k}\log N+\frac{1}{2}\overline{k}[\log\overline{k}-1]-\sum_k\tilde{p}(k)\log k!+\sum_k\tilde{p}(k)k\log F(k|Q)\\ &\quad -\frac{1}{2}\sum_{kk'}\log Q(k,k')N^{-1}\sum_{ij}c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})}+o(1)\\ &= \frac{1}{2}\overline{k}\log N+\frac{1}{2}\overline{k}[\log\overline{k}-1]-\sum_k\tilde{p}(k)\log k!+\sum_k\tilde{p}(k)k\log F(k|Q)\\ &\quad -\sum_{kk'}\frac{\tilde{p}(k)\tilde{p}(k')kk'}{2\overline{k}}\Pi(k,k'|\mathbf{c})\log[\Pi(k,k')F(k|Q)F(k'|Q)]+o(1)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\overline{k}\log N + \frac{1}{2}\overline{k}[\log\overline{k} - 1] - \sum_k \tilde{p}(k)\log k! \\
&\quad + \sum_k \tilde{p}(k)k\log F(k|Q)\Big[1 - \sum_{k'}\frac{\tilde{p}(k')k'}{\overline{k}}\Pi(k,k'|\mathbf{c})\Big] \\
&\quad - \sum_{kk'}\frac{\tilde{p}(k)\tilde{p}(k')kk'}{2\overline{k}}\Pi(k,k'|\mathbf{c})\log\Pi(k,k') + o(1) \\
&= \frac{1}{2}\overline{k}\log N + \frac{1}{2}\overline{k}[\log\overline{k} - 1] - \sum_k \tilde{p}(k)\log k! \\
&\quad - \sum_{kk'}\frac{\tilde{p}(k)\tilde{p}(k')kk'}{2\overline{k}}\Pi(k,k'|\mathbf{c})\log\Pi(k,k') + o(1),
\end{aligned}
\tag{23}
$$

where in the last step we used identities (10). It subsequently follows (21) as

$$
\begin{aligned}
\Omega(\mathbf{c}|p,Q) &= \frac{1}{2}\overline{k}(\mathbf{c})\log N + \frac{1}{2}\overline{k}(\mathbf{c})[\log\overline{k}(\mathbf{c}) - 1] \\
&\quad - \sum_k p(k|\mathbf{c})\log k! - \Omega[p(\mathbf{c}),\Pi(\mathbf{c});p,\Pi] + o(1)
\end{aligned}
\tag{24}
$$

$$
\begin{aligned}
\Omega[p(\mathbf{c}),\Pi(\mathbf{c});p,\Pi] &= \sum_{kk'}\frac{p(k|\mathbf{c})p(k'|\mathbf{c})kk'}{2\overline{k}(\mathbf{c})}\Pi(k,k'|\mathbf{c})\log\Pi(k,k') \\
&\quad + \sum_k p(k|\mathbf{c})\log p(k),
\end{aligned}
\tag{25}
$$

with $\overline{k}(\mathbf{c}) = \sum_k kp(k|\mathbf{c})$, $\Pi(\mathbf{c}) = \{\Pi(k,k'|\mathbf{c})\}$ and $p(\mathbf{c}) = \{p(k|\mathbf{c})\}$. The leading order $\frac{1}{2}\overline{k}(\mathbf{c})\log N$ in $\Omega(\mathbf{c}|Q,p)$ reflects the property that the number of finitely connected graphs grows asymptotically with N as $\exp[\sim N\log N]$. The next order is found to depend only on the macroscopic characterization $\{p(\mathbf{c}),\Pi(\mathbf{c})\}$ of the *specific* graph $\mathbf{c}$, and on the macroscopic characterization $\{p,\Pi\}$ of *typical* graphs from (14), with Π calculated for the kernel Q via (19).

3.3. Existence and uniqueness of tailored ensembles

We will now prove that for each degree distribution $p(k)$ and each relative degree correlation function $\Pi(k,k')$ there exist kernels $Q(k,k')$ such that their associated ensembles (14) will for large N be tailored to the production of random graphs with precisely these statistical features. We identify these kernels and show that they all correspond in leading order in N to the *same* random graph ensemble.

- *Existence of a family of tailored kernels.* For each non-negative function $\phi(k)$ such that $p(k)\phi(k)\Pi(k,k')\phi(k')p(k')$ is nonzero for at least one combination (k,k'), the following kernel satisfies all conditions required to define a random graph ensemble of the family (14) that generates graphs with degree distribution $p(k)$ and relative degree correlation function $\Pi(k,k')$ as $N\to\infty$:

$$
Q(k,k') = \frac{\phi(k)\Pi(k,k')\phi(k')}{Z}, \qquad Z = \sum_{kk'} p(k)\phi(k)\Pi(k,k')\phi(k')p(k').
\tag{26}
$$

$Q(k,k')$ is by construction non-negative, symmetric and correctly normalized. Also we will always find $Z > 0$ due to $\Pi(k,k') \geqslant 0$ in combination with our conditions on $\phi(k)$

and the normalization (12). Recovering the correct degree distribution is built into the ensemble (14) via the degree constraints. To prove that equations (19), (20) are satisfied we define $F(k|Q) = \phi(k)/\sqrt{Z}$, and use the fact that by virtue of (19) the condition (20) reduces to (10), and is therefore guaranteed to hold, provided $\Pi(k,k')$ indeed represents a relative degree correlation function.

What remains is to show that there exist functions $\phi(k)$ that meet the relevant conditions. The simplest candidate is $\phi(k) = k/\langle k\rangle$, for which we find $Z = 1$ via (12) and which is easily confirmed to meet all criteria. It gives what we will call the *canonical kernel*:

$$Q^{\star}(k,k') = \Pi(k,k')kk'/\langle k\rangle^2. \tag{27}$$

- *Completeness of the family of tailored kernels.* The set of kernels defined by (26) is *complete*: if a kernel $Q(k,k')$ generates random graphs with statistics $p(k)$ and $\Pi(k,k')$, then it must be of the form (26).

  The proof is simple. If $Q(k,k')$ generates graphs with relative degree correlation function $\Pi(k,k')$, according to (19) it must be of the form $Q(k,k') = F(k)\Pi(k,k')F(k')$ for some function $F(k)$. Since both $\Pi(k,k')$ and $Q(k,k')$ must be non-negative, the same must be true for $F(k)$. Hence $Q(k,k')$ is also of the form (26), with $\phi(k) = \sqrt{Z}F(k)$ and with the formula for Z in (26) satisfied automatically due to $Q(k,k')$ having to be normalized.

  A further corollary is that all kernels tailored to the generation of graphs with statistics $p(k)$ and $\Pi(k,k')$ are related to the canonical kernel (27) via separable transformations, with suitably normalized non-negative functions $G(k)$:

$$Q(k,k') = G(k)Q^{\star}(k,k')G(k'). \tag{28}$$

- *Asymptotic uniqueness of the canonical ensemble.* The random graph ensembles of all kernels of the family (26), tailored to generating random graphs with statistical properties $p(k)$ and $\Pi(k,k')$, are asymptotically (i.e. for large enough N) identical; if all $\{k_i\}$ are drawn randomly from $p(k)$, and $Q(k,k')$ belongs to the family (26) with canonical member $Q^{\star}(k,k')$ defined in (27), then

$$[\mathrm{Prob}(\mathbf{c}|p,Q)]^{1/N} = [\mathrm{Prob}(\mathbf{c}|p,Q^{\star})]^{1/N}\,\mathrm{e}^{o(1)}. \tag{29}$$

  This follows from (24), which tells us that in the two leading orders in N the probabilities of graphs generated from (26) depend on the kernel $Q(k,k')$ of the ensemble only via its associated function $\Pi(k,k')$, so that $N^{-1}\log\mathrm{Prob}(\mathbf{c}|p,Q) - N^{-1}\log\mathrm{Prob}(\mathbf{c}|p,Q^{\star}) = o(1)$.

The above results imply that we may regard the random graph ensemble (14), equipped with the kernel (27), as the natural ensemble for generating large random graphs with topologies controlled strictly by a prescribed degree distribution $p(k)$ and prescribed relative degree correlations $\Pi(k,k')$. We will call $p(\mathbf{c}|p,Q^{\star})$, with $Q^{\star}(k,k') = \Pi(k,k')kk'/\langle k\rangle^2$, the *canonical ensemble* for graphs with $p(k)$ and $\Pi(k,k')$. Note that for $\Pi(k,k') = 1$ one has $Q^{\star}(k,k') = kk'/\langle k\rangle^2$, which is indeed equivalent to the trivial choice $Q(k,k') = 1$ (as it is related to the latter by a separable transformation).

We can now also define what we mean by 'null models'. Given the hypothesis that a network $\mathbf{c}$ has no structure beyond that imposed by its degree statistics, the appropriate null model is a random graph generated by the canonical ensemble with degree distribution $p(k) = p(k|\mathbf{c})$ and relative degree correlations $\Pi(k,k') = 1$ (giving the trivial kernel $Q(k,k') = 1$; these are usually referred to as 'simple graphs'). Similarly, given the hypothesis that a network has no structure beyond that imposed by its degree statistics and its degree–degree correlations, the appropriate null model is a random graph generated by the canonical ensemble with degree distribution $p(k) = p(k|\mathbf{c})$ and relative degree correlations $\Pi(k, k') = \Pi(k, k'|\mathbf{c})$.

Finally, self-consistency demands that $p(k)$ and the canonical kernel (or a member of its equivalent family, related by separable transformations) are also the most probable pair $\{p, Q\}$ in a Bayesian sense. The probability $\text{Prob}(p, Q|\mathbf{c})$ that a pair $\{p, Q\}$ was the 'generator' of $\mathbf{c}$ via (14) can be expressed, via standard Bayesian relations, in terms of the probability $\text{Prob}(\mathbf{c}|p, Q)$ of drawing $\mathbf{c}$ at random from (14):

$$\text{Prob}(p, Q|\mathbf{c}) = \frac{\text{Prob}(\mathbf{c}, p, Q)}{\text{Prob}(\mathbf{c})} = \frac{\text{Prob}(\mathbf{c}|p, Q)\text{Prob}(p, Q)}{\sum_{Q'}\sum_{p'}\text{Prob}(\mathbf{c}|p', Q',)\text{Prob}(p', Q')}. \tag{30}$$

The most probable pair $\{p, Q\}$ is the one that maximizes $\log\text{Prob}(p, Q|\mathbf{c}) = \log\text{Prob}(p, Q) + \log\text{Prob}(\mathbf{c}|p, Q)$ (modulo terms independent of $\{p, Q\}$), so in the absence of any prior bias, i.e. if $\text{Prob}(p, Q)$ is independent of $\{p, Q\}$, it is the kernel that maximizes $\text{Prob}(\mathbf{c}|p, Q)$. Since $\sum_{\mathbf{c}}\text{Prob}(\mathbf{c}|p, Q) = 1$ for any $\{p, Q\}$, finding the most probable $\{p, Q\}$ for a graph $\mathbf{c}$ boils down to finding the *smallest* ensemble of graphs compatible with the structure of $\mathbf{c}$. Intuitively this makes sense; a more detailed characterization of the topology of an observed graph allows for more information being carried over from the graph to the ensemble, reducing the number of potential graphs allowed for by the ensemble. The smaller the number of graphs in the ensemble, the more accurate will these graphs be when used as proxies for the observed one.

Maximizing $\text{Prob}(\mathbf{c}|p, Q)$ over $\{p, Q\}$ means minimizing $\Omega(\mathbf{c}|p, Q)$ in (21), of which the leading orders in N are given in (24). Demonstrating Bayesian self-consistency of our canonical graph ensemble for large N hence boils down to proving that the maximum of (25) over $\{p, \Pi\}$ (subject to the relevant constraints) is obtained for $\{p, \Pi\} = \{p(\mathbf{c}), \Pi(\mathbf{c})\}$. The constraints include the set (10). There are clearly more, e.g. $\Pi(k, k') \geqslant 0$ for all (k, k'); however, we show below that maximizing (25) over $\{p, \Pi\}$ subject only to (10) and $\sum_k p(k) = 1$ already generates the desired result: $\{p, \Pi\} = \{p(\mathbf{c}), \Pi(\mathbf{c})\}$. Extremizing (25) with the Lagrange formalism, leads to the following equations, which are to be solved in combination with (10) and $\sum_k p(k) = 1$:

$$\forall(k, k') : \quad \frac{\partial}{\partial\Pi(k, k')}\Omega[p(\mathbf{c}), \Pi(\mathbf{c}); p, \Pi] = \sum_{\ell\geqslant 0}\lambda(\ell)\frac{\partial}{\partial\Pi(k, k')}\left(\frac{1}{\langle k\rangle}\sum_{\ell'}\ell' p(\ell')\Pi(\ell, \ell') - 1\right) \tag{31}$$

$$\forall k : \quad \frac{\partial}{\partial p(k)}\Omega[p(\mathbf{c}), \Pi(\mathbf{c}); p, \Pi] = \sum_{\ell\geqslant 0}\lambda(\ell)\frac{\partial}{\partial p(k)}\left(\frac{1}{\langle k\rangle}\sum_{\ell'}\ell' p(\ell')\Pi(\ell, \ell') - 1\right) + \mu\frac{\partial}{\partial p(k)}\left(\sum_{k'}p(k') - 1\right), \tag{32}$$

where $\{\lambda(\ell)\}$ and μ are Lagrange multipliers. Working out (31) gives

$$\forall(k, k') : \quad \frac{p(k)\Pi(k, k')p(k')}{\langle k\rangle} = \frac{p(k|\mathbf{c})\Pi(k, k'|\mathbf{c})p(k'|\mathbf{c})}{\bar{k}(\mathbf{c})}\frac{p(k)k}{2\lambda(k)}. \tag{33}$$

Since both $\Pi(k, k')$ and $\Pi(k, k'|\mathbf{c})$ must satisfy the constraints (10), with degree distributions $p(k)$ and $p(k|\mathbf{c})$, respectively, it follows from (33) that

$$\lambda(k) = \tfrac{1}{2}p(k)k. \tag{34}$$

With this expression we eliminate $\lambda(k)$ from (33) to find

$$\forall (k,k'): \quad \frac{p(k)\Pi(k,k')p(k')}{\langle k\rangle} = \frac{p(k|\mathbf{c})\Pi(k,k'|\mathbf{c})p(k'|\mathbf{c})}{\bar{k}(\mathbf{c})}. \tag{35}$$

Next we work out (32) and substitute (34) into the result. This gives, using symmetry of Π,

$$\forall k: \quad p(k|\mathbf{c}) = \mu p(k) + \frac{p(k)}{2\langle k\rangle}\sum_{k'} p(k')k'\Pi(k',k) = \left(\mu + \frac{1}{2}\right)p(k). \tag{36}$$

The normalization conditions $\sum_k p(k|\mathbf{c}) = \sum_k p(k) = 1$ then tell us that $\mu = \frac{1}{2}$, so $p(k) = p(k|\mathbf{c})$ for all k, and finally also (via (35)):

$$\forall (k,k'): \quad \Pi(k,k') = \Pi(k,k'|\mathbf{c}). \tag{37}$$

Hence, the choice $\{p,\Pi\} = \{p(\mathbf{c}),\Pi(\mathbf{c})\}$ indeed extremizes the leading two orders in N of (24), subject to (10) and to normalization of $p(k)$. The above extremum must be a maximum, since by making pathological choices for $\{p,\Pi\}$ (namely choices inconsistent with the structure of $\mathbf{c}$) we can make prob$(\mathbf{c}|p,Q)$ arbitrary small, and hence $\Omega[p(\mathbf{c}),\Pi(\mathbf{c});p,\Pi]$ arbitrarily small. Hence our canonical ensembles are indeed self-consistent in a Bayesian sense, as expected.

3.4. The random graphs ensemble as a conditioned maximum entropy ensemble

In this section we show that our canonical ensemble gives the maximum entropy within the subspace of graphs with prescribed degrees and upon imposing as a constraint the average values $\Pi(k,k') = \langle \Pi(k,k'|\mathbf{c})\rangle$ of the relative degree correlations. First we define our constraining observables, i.e. the degree sequence and the re-scaled degree correlation:

$$k_i(\mathbf{c}) = k_i (\forall i) \tag{38}$$

$$q(k,k'|\mathbf{c}) = N^{-1}\sum_{ij} c_{ij}\delta_{k,k_i(\mathbf{c})}\delta_{k',k_j(\mathbf{c})} \qquad (k,k'>0). \tag{39}$$

Note that if $N^{-1}\sum_i \delta_{k,k_i(\mathbf{c})} = p(k)$ and $\langle k\rangle = \sum_k p(k)$, then

$$\Pi(k,k'|\mathbf{c}) = \frac{\langle k\rangle}{p(k)p(k')kk'}q(k,k'|\mathbf{c}) \qquad \text{for} \quad N\to\infty. \tag{40}$$

We are interested in the maximum entropy random graph ensemble $p(\mathbf{c})$ (limited to symmetric graphs without self-interactions) such that $q(k,k') = \sum_{\mathbf{c}} p(\mathbf{c})q(k,k'|\mathbf{c})$ for all (k,k') and $k_i = k_i(\mathbf{c})$ for all i. This is given by the ensemble $p(\mathbf{c})$ for which the Shannon entropy

$$S[\mathbf{k},q] = \sum_{\mathbf{c}} p(\mathbf{c}|\mathbf{k},q)\log p(\mathbf{c}|\mathbf{k},q) \tag{41}$$

is maximal subject to our constraints. Extremization of (41) with Lagrange multipliers gives, without enforcing $p(\mathbf{c}) \geqslant 0$ explicitly,

$$\forall \mathbf{c}: \quad \frac{\partial}{\partial p(\mathbf{c})}\left\{\sum_{\mathbf{c}'} p(\mathbf{c}')\left[\log p(\mathbf{c}') + \Lambda_0 + \sum_{kk'}\lambda(k,k')q(k,k'|\mathbf{c}')\right]\right\} = 0 \tag{42}$$

$$\forall \mathbf{c}: \quad \log p(\mathbf{c}) + \Lambda_0 + \sum_{kk'}\lambda(k,k')q(k,k'|\mathbf{c}) + 1 = 0 \tag{43}$$

$$\forall \mathbf{c}: \quad p(\mathbf{c}) = \frac{1}{\mathcal{Z}}\,\mathrm{e}^{-\sum_{kk'}\lambda(k,k')q(k,k'|\mathbf{c})} \tag{44}$$

$$\forall \mathbf{c}: \quad p(\mathbf{c}) = \frac{1}{\mathcal{Z}} \mathrm{e}^{-N^{-1}\sum_{ij} c_{ij}\lambda(k_i(\mathbf{c}),k_j(\mathbf{c}))}, \tag{45}$$

with $\mathcal{Z}$ such that $\sum_{\mathbf{c}} p(\mathbf{c}) = 1$. As expected for an ensemble of random graphs with maximum entropy, where a set of averages of obervables are constrained to assume prescribed values, the result of the extremization gives an exponential family, where the parameters $\{\lambda(k_i, k_j)\}$ are to be calculated from the equations for the constraints. What is left is to show that the exponential family can be reduced to the micro-canonical ensemble (15), where degrees are prescribed by a simple redefinition of the Lagrange multipliers. Let us first rewrite (45) as

$$\forall \mathbf{c}: \quad p(\mathbf{c}) = \frac{1}{\mathcal{Z}} \left(\prod_{i<j} \mathrm{e}^{-N^{-1}c_{ij}[\lambda(k_i,k_j)+\lambda(k_j,k_i)]}\right)\left(\prod_i \delta_{k_i,k_i(\mathbf{c})}\right) \tag{46}$$

$$\forall \mathbf{c}: \quad p(\mathbf{c}) = \frac{1}{\mathcal{Z}} \left(\prod_{i<j} [\mathrm{e}^{-N^{-1}[\lambda(k_i,k_j)+\lambda(k_j,k_i)]}\delta_{c_{ij},1} + \delta_{c_{ij},0}]\right)\left(\prod_i \delta_{k_i,k_i(\mathbf{c})}\right). \tag{47}$$

We can then redefine our Langrange multipliers in terms of the function $Q(k, k')$ via

$$\frac{\langle k\rangle}{N} Q(k,k') = \frac{\mathrm{e}^{-N^{-1}[\lambda(k_i,k_j)+\lambda(k_j,k_i)]}}{1+\mathrm{e}^{-N^{-1}[\lambda(k_i,k_j)+\lambda(k_j,k_i)]}}.$$

This results in

$$p(\mathbf{c}) = \frac{1}{\mathcal{Z}} \prod_{i<j}\left(1 - \frac{\langle k\rangle Q(k_i,k_j)}{N}\right)^{-1} \times \prod_{i<j}\left[\frac{\langle k\rangle}{N}Q(k_i,k_j)\delta_{c_{ij},1} + \left(1-\frac{\langle k\rangle}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right]\cdot \prod_i \delta_{k_i,k_i(\mathbf{c})}. \tag{48}$$

The first product in (48) only depends on the constrained degrees $\{k_i\}$ (in fact, to leading order this dependence is only via their average $\langle k\rangle$, since $\prod_{i<j}(1 - \langle k\rangle Q(k_i, k_j)/N)^{-1} = \mathrm{e}^{N\langle k\rangle/2+\mathcal{O}(1)}$), so it drops out of the measure, and hence (48) can be rewritten as

$$p(\mathbf{c}) = \frac{1}{\mathcal{Z}(\mathbf{k},Q)} \prod_{i<j}\left[\frac{\langle k\rangle}{N}Q(k_i,k_j)\delta_{c_{ij},1} + \left(1-\frac{\langle k\rangle}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right]\prod_i \delta_{k_i,k_i(\mathbf{c})} \tag{49}$$

$$\mathcal{Z}(\mathbf{k},Q) = \sum_{\mathbf{c}}\prod_{i<j}\left[\frac{\langle k\rangle}{N}Q(k_i,k_j)\delta_{c_{ij},1} + \left(1-\frac{\langle k\rangle}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right]\prod_i \delta_{k_i,k_i(\mathbf{c})} \tag{50}$$

which indeed reduces to (15), as claimed.

3.5. Shannon entropy

The (rescaled) Shannon entropy of the canonical ensemble $\mathrm{Prob}(\mathbf{c}|Q^\star, p)$, as defined by $Q^\star(k, k') = \Pi(k, k')kk'/\langle k\rangle^2$ in combination with (14), is an important quantity as it allows us to define and calculate the effective number of graphs $\mathcal{N}[p, \Pi]$ in the ensemble:

$$S[p,\Pi] = -\frac{1}{N}\sum_{\mathbf{c}} \mathrm{Prob}(\mathbf{c}|p,Q^\star)\log \mathrm{Prob}(\mathbf{c}|p,Q^\star) \tag{51}$$

$$\mathcal{N}[p,\Pi] = \mathrm{e}^{NS[p,\Pi]}. \tag{52}$$

In (51) one defines as always $0\log 0=\lim_{\epsilon\downarrow 0}\epsilon\log\epsilon=0$. For large N we can use our earlier results (21), (24), (25) to find the leading orders of the entropy since

$$\begin{aligned}
S[p,\Pi] &= \sum_{\mathbf{c}}\mathrm{Prob}(\mathbf{c}|p,Q^\star)\Omega(\mathbf{c}|p,Q^\star)\\
&= \frac{1}{2}\langle k\rangle[\log[N\langle k\rangle]-1]-\sum_k p(k)\log k!\\
&\quad -\sum_{\mathbf{c}}\mathrm{Prob}(\mathbf{c}|p,Q^\star)\Omega[\Pi(\mathbf{c}),p(\mathbf{c});\Pi,p]+o(1)\\
&= \frac{1}{2}\langle k\rangle[\log[N\langle k\rangle]-1]-\sum_k p(k)\log k!-\sum_k p(k)\log p(k)\\
&\quad -\sum_{kk'}\frac{p(k)p(k')kk'}{2\langle k\rangle}\Pi(k,k')\log\Pi(k,k')+o(1)\\
&= \frac{1}{2}\langle k\rangle[\log[N/\langle k\rangle]+1]-\sum_k p(k)\log[p(k)/\pi(k)]\\
&\quad -\frac{1}{2\langle k\rangle}\sum_{kk'}p(k)p(k')kk'\Pi(k,k')\log\Pi(k,k')+o(1),
\end{aligned}\tag{53}$$

where $\pi(k)$ denotes the Poissonian degree distribution with average degree $\langle k\rangle$, namely $\pi(k)=\mathrm{e}^{-\langle k\rangle}\langle k\rangle^k/k!$. To prove various properties of the above expression for the entropy it will be convenient to introduce a new (symmetric) quantity $W(k,k')$, defined as the probability that a randomly drawn link in a graph that has $\Pi(k,k'|\mathbf{c})=\Pi(k,k')$ connects two nodes with degrees k and k'. It can be shown to be related to $\Pi(k,k')$ via

$$W(k,k')=p(k)p(k')kk'\Pi(k,k')/\langle k\rangle^2. \tag{54}$$

Irrespective of its exact meaning, the crucial mathematical advantage here of working with $W(k,k'|p,\Pi)$ is that it represents a probability distribution: $W(k,k')\geqslant 0$ and $\sum_{kk'}W(k,k')=1$ (normalization follows from (10)). One also verifies explicitly that $W(k)=\sum_{k'}W(k,k')=p(k)k/\langle k\rangle$ for all k. If we use (54) to eliminate $\Pi(k,k')$ from (53) in favour of $W(k,k')$ we get

$$\begin{aligned}
S[p,\Pi] &= \frac{1}{2}\langle k\rangle[\log[N/\langle k\rangle]+1]-\sum_k p(k)\log[p(k)/\pi(k)]\\
&\quad -\frac{1}{2}\langle k\rangle\sum_{kk'}W(k,k')\log[W(k,k')/W(k)W(k')]+o(1).
\end{aligned}\tag{55}$$

The term in (55) with $W(k,k')$ is seen to be proportional to minus the mutual information between two connected sites, and is therefore non-positive, vanishing if and only if $\Pi(k,k')=1$ for all (k,k'). Furthermore, the term in (55) involving $\pi(k)$ is minus a KL-divergence, and therefore also non-positive, vanishing if and only if $p(k)=\pi(k)$ for all k. Our result (55) therefore has a clear and elegant interpretation.

- For the simplest graphs of the Erdös–Rényi type, where only the average degree $\langle k\rangle$ is imposed, one has $p(k)=\pi(k)$ for all k and $\Pi(k,k')=1$ for all (k,k'). This gives $W(k,k')=W(k)W(k')$, and the entropy takes its maximal value:

$$S[p,\Pi]=\tfrac{1}{2}\langle k\rangle[\log[N/\langle k\rangle]+1]. \tag{56}$$

- For graphs where the degree distribution $p(k)$ is imposed, but without further structure (i.e. still $\Pi(k,k')=1$ for all (k,k')), the entropy decreases by an amount $\sum_k p(k)\log[p(k)/\pi(k)]$ which is the KL-distance between the imposed $p(k)$ and the

Poissonian degree distribution with the same average connectivity:

$$S[p,\Pi] = \frac{1}{2}\langle k\rangle[\log[N/\langle k\rangle]+1] - \sum_k p(k)\log[p(k)/\pi(k)]. \tag{57}$$

- For the more sophisticated graphs where both a degree distribution $p(k)$ and nontrivial degree correlations defined via $\Pi(k,k')$ are imposed, one no longer has $W(k,k') = W(k)W(k')$ and the entropy decreases further by an amount $\frac{1}{2}\langle k\rangle\sum_{kk'}W(k,k')\log[W(k,k')/W(k)W(k')]$, which is proportional to the mutual information regarding degrees of connected nodes:

$$S[p,\Pi] = \frac{1}{2}\langle k\rangle[\log[N/\langle k\rangle]+1] - \sum_k p(k)\log[p(k)/\pi(k)] \\ - \frac{1}{2}\langle k\rangle\sum_{kk'}W(k,k')\log[W(k,k')/W(k)W(k')]. \tag{58}$$

4. Quantitative tools for networks

In the previous sections we have shown that ensemble (14) is tailored, for large N, to the production of graphs with degree distribution $p(k)$ and degree correlation $\Pi(k,k')$ given by (19), (20). Conversely, for each desired function $\Pi(k,k')$, one may always choose the canonical kernel $Q^{\star}(k,k') = \Pi(k,k')kk'/\langle k\rangle^2$ in (14) to tailor the ensemble to the production of graphs with the desired degree correlation.

The availability for any given/observed network **c** of a well-defined canonical random graph ensemble, that produces random graphs with microscopic topologies controlled solely by the observed degree statistics and degree correlations of the given **c**, allows us to develop practical quantitative tools with which to analyse and compare (structure in) real networks. Here we focus on three such tools.

4.1. Quantifying structural network complexity

The natural definition of the complexity of a given network **c** is based on the number $\mathcal{N}[p,\Pi]$ of graphs in its canonical ensemble $\{p,\Pi\}$, and hence on the entropy per node $S[p,\Pi]$ given in (55). It makes sense to write this entropy for large N as $S[p,\Pi] = S_0 - \mathcal{C}[p,\Pi] + o(1)$, with a first (positive) contribution $S_0 = \frac{1}{2}\langle k\rangle\big[\log[N/\langle k\rangle+1\big]$ that originates simply from counting the total number of bonds and would also be found for structureless Erdös–Rényi graphs (where only the average degree is prescribed), minus a second term $\mathcal{C}[p,\Pi]$ which acts to *reduce* the entropy as soon as there is a structure in the graph beyond a prescribed average degree. The latter quantity $\mathcal{C}[p,\Pi]$ can be identified as the complexity of graphs in the canonical ensemble associated with **c**, and hence as the complexity of **c**:

$$\mathcal{C}[p,\Pi] = \sum_k p(k)\log[p(k)/\pi(k)] + \frac{1}{2\langle k\rangle}\sum_{kk'}p(k)p(k')kk'\Pi(k,k')\log\Pi(k,k'), \tag{59}$$

where $\pi(k)$ is the Poissonian distribution with average degree $\langle k\rangle$:

$$\pi(k) = \mathrm{e}^{-\langle k\rangle}\langle k\rangle^k/k! \tag{60}$$

The larger the $\mathcal{C}[p,\Pi]$, the more 'rare' or 'special' are graphs with characteristics $\{p,\Pi\}$. For every N, the complexity is bounded from above by (56); at this value the network undergoes an entropy 'crisis', as (58) vanishes and the degree distribution ceases to be graphical, i.e. no network can be found with this degree distribution (see [22] for the notion of graphicality). Note, however, that our results were obtained in the limit $\langle k\rangle \ll N$; they no longer apply

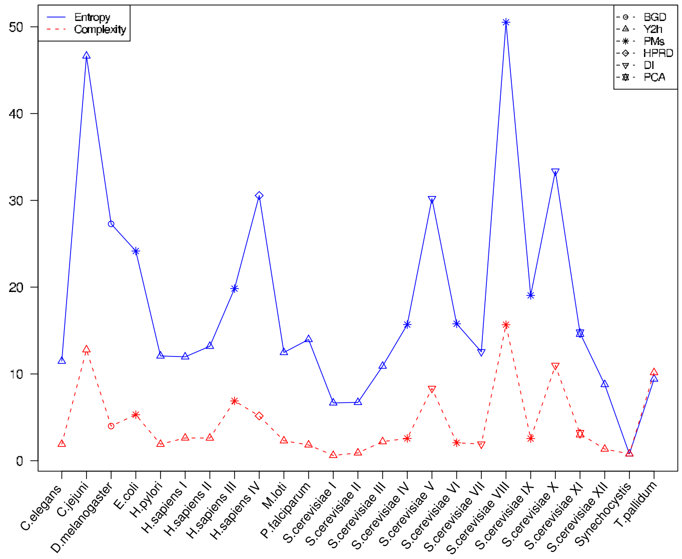

Figure 1. Shannon entropy per node $S[p, \Pi]$ (markers connected by solid lines) and complexity $C[p, \Pi]$ (markers connected by dashed lines) of the canonical ensembles tailored to the production of random graphs with microscopic topologies controlled solely by the degree sequence and degree correlation of experimentally determined PPINs. The methods/sources for the experimental data sets are the following: BGD, BioGrid database; Y2h, yeast two-hybrid screen; PMs, purification mass spectrometry; HPRD, human protein reference database; DI, data integration (database with combined experimental data); PCA, protein fragment complementation assay. The studied organisms are listed in alphabetical order on the x-axis. Data sets properties and references are summarized in table 1.

(This figure is in colour only in the electronic version)

for degree distributions with an average connectivity of the order of the system size. For example, for fully connected graphs, where the complexity is maximal, the entropy should vanish, whereas (58) indeed yields an incorrect $\mathcal{O}(N)$ result. As an illustration one may check how close to the entropy crisis are PPIN of different species (PPIN typically meet the requirements $\langle k\rangle \ll N$ for our theory to apply). For this purpose we have computed (58) for protein interaction networks of different species and show the results in figure 1. A more systematic and extensive application of our tools to PPIN will be published in [16].

4.2. *Quantifying structural distance between networks*

In the same spirit we can now also use our tailored graph ensembles to define an information-theoretic distance D_{AB} between any two networks $\mathbf{c}_A$ and $\mathbf{c}_B$, based solely on the macroscopic structure statistics as captured by their associated (observed) structure function pairs $\{p_A, \Pi_A\}$ and $\{p_B, \Pi_B\}$. The natural definition would be in terms of the Jeffreys divergence (i.e. the symmetrized KL-distance) between the two associated canonical ensembles, which is non-negative and equals zero if and only if $\{p_A, \Pi_A\} = \{p_B, \Pi_B\}$, i.e. if the graphs $\mathbf{c}_A$ and $\mathbf{c}_B$ belong to the same canonical ensemble:

$$D_{AB} = \frac{1}{2N}\sum_{\mathbf{c}} \text{Prob}(\mathbf{c}|p_A, Q_A) \log\left[\frac{\text{Prob}(\mathbf{c}|p_A, Q_A)}{\text{Prob}(\mathbf{c}|p_B, Q_B)}\right] + \frac{1}{2N}\sum_{\mathbf{c}} \text{Prob}(\mathbf{c}|p_B, Q_B) \log\left[\frac{\text{Prob}(\mathbf{c}|p_B, Q_B)}{\text{Prob}(\mathbf{c}|p_A, Q_A)}\right]. \tag{61}$$

Table 1. Maximum degree $k_{\max}$, detection method/source and reference for the biological network data sets. The detection methods/sources are abbreviated as in figure 1.

Species	$k_{\max}$	Method	Reference
C. elegans	99	Y2h	[35]
C. jejuni	207	Y2h	[36]
D. melanogaster	176	BGD	[37]
E. coli	641	PMs	[32]
H. pylori	55	Y2h	[38]
H. sapiens I	125	Y2h	[39]
H. sapiens II	95	Y2h	[40]
H. sapiens III	314	PMs	[41]
H.sapiens IV	247	HPRD	[34]
M. loti	401	Y2h	[42]
P. falciparum	51	Y2h	[43]
S. cerevisiae I	24	Y2h	[44]
S. cerevisiae II	55	Y2h	[45]
S. cerevisiae III	279	Y2h	[45]
S. cerevisiae IV	62	PMs	[46]
S. cerevisiae V	118	DI	[47]
S. cerevisiae VI	53	PMs	[48]
S. cerevisiae VII	32	DI	[49]
S.cerevisiae VIII	955	PMs	[50]
S. cerevisiae IX	141	PMs	[51]
S. cerevisiae X	127	DI	[52]
S. cerevisiae XI	58	PCA	[53]
*S. cerevisiae*XII	86	Y2h-PCA	[54]
Synechocystis	51	Y2h	[55]
T. pallidum	285	Y2h	[56]

Working out this formula, using (24) and (55), gives for large N

$$\begin{aligned}
D_{AB} &= \frac{1}{2}\sum_{\mathbf{c}} \mathrm{Prob}(\mathbf{c}|p_A, Q_A)\Omega(\mathbf{c}|p_B, Q_B) - \frac{1}{2}S[p_A, \Pi_A] \\
&\quad + \frac{1}{2}\sum_{\mathbf{c}} \mathrm{Prob}(\mathbf{c}|p_B, Q_B)\Omega(\mathbf{c}|p_A, Q_A) - \frac{1}{2}S[p_B, \Pi_B] \\
&= \frac{1}{2}\sum_{k} p_A(k)\log\left[\frac{p_A(k)}{p_B(k)}\right] + \sum_{kk'} \frac{p_A(k)p_A(k')kk'}{4\langle k\rangle_A}\Pi_A(k,k')\log\left[\frac{\Pi_A(k,k')}{\Pi_B(k,k')}\right] \\
&\quad + \frac{1}{2}\sum_{k} p_B(k)\log\left[\frac{p_B(k)}{p_A(k)}\right] \\
&\quad + \sum_{kk'} \frac{p_B(k)p_B(k')kk'}{4\langle k\rangle_B}\Pi_B(k,k')\log\left[\frac{\Pi_B(k,k')}{\Pi_A(k,k')}\right] + o(1).
\end{aligned} \qquad (62)$$

This quantity is used in [16] for comparing and clustering PPIN data sets, even if these differ in size, solely on the basis of their degree sequence and degree correlations. The combination of its information-theoretic origin and explicit nature (so that it involves almost no computational cost) makes (62) an efficient practical tool in bio-informatics.

4.3. Numerical generation of canonical 'null models'

We have shown that for any given network **c** it is possible to define a tailored ensemble of graphs, that share with **c** those structural aspects that follow directly from its degree distribution and degree correlations, and used it to define and calculate complexities and structural distances. Our next aim is to use the ensemble for *generating* random graphs with structure $\{p, \Pi\}$ identical to that of a given network. The problems associated with generating complex random graphs with controlled properties are well known [23–30]. In [17] a general method was proposed for generating random graphs with built-in constraints and specific statistic weights, such as described by the invariant measure (14), in the form of a Monte Carlo process that is guaranteed to evolve from any initial graph $\mathbf{c}_0$ that meets the relevant constraints towards the prescribed invariant measure (14). The initial graph $\mathbf{c}_0$ can be constructed by hand, for any choice of $p(k)$, such that for sufficiently large N it will have the required degree statistics (see e.g. [31]). With the general and exact algorithm [17] we can generate graphs according to the measure (14), with the kernel $Q(k, k')$ of (27) chosen such as to impose any desired degree correlations $\Pi(k, k')$. These graphs can then serve as 'null models', allowing us, for instance, to determine to what extent specific small motifs in biological networks (such as short loops) can be regarded as mere consequences of the overall structure dictated by their degree statistics and degree correlations, or whether they reflect deeper biological principles. See [16] for the results of such tests.

Here we generate, as an illustration, a synthetic network which is to have the same degree sequence and the same degree correlations as the protein interaction network of *Escherichia coli*, as given in [32] (i.e. we produce a member of the tailored graph ensemble of this particular PPIN) where $N = 2457$ and $\langle k\rangle = 7.05$. The degree correlations of the resulting graph after 67 147 accepted moves of the Markov chain algorithm of [17] are shown in figure 2(*b*), and are seen to be in very good agreement with the degree correlations of the PPIN that are being targeted, displayed in figure 2(*a*) (note that there is no need to compare degree distributions since all degrees are guaranteed to be conserved by the graph dynamics [17]). To rule out the possibility that the observed similarity in degree correlations between the synthetic graph and the original PPIN could have arisen from poor sampling of the microscopic configurations (and just reflect direct similarities in the connectivity matrices), we also calculated the Hamming distance between the connectivity matrices **c** and **c**′ of the original PPIN and the synthetic graph,

$$\rho(\mathbf{c}, \mathbf{c}') = \frac{1}{2N\langle k\rangle} \sum_{ij} |c_{ij} - c'_{ij}| \tag{63}$$

(the prefactor is chosen such that when the two matrices differ in all the $2N\langle k\rangle$ entries which could be different, then $\rho(\mathbf{c}, \mathbf{c}') = 1$). The Hamming distance vanishes if the two matrices are identical. In the present case we find $\rho = 0.90$, which implies that although our two graphs have similar macroscopic structure, their microscopic realizations are indeed very different.

For comparison, we also show in figure 2(*c*) the degree correlations of a synthetic graph obtained via the Markov chain dynamics of [17], starting from the same initial graph, but now targeting degree correlations described by $\Pi(k, k') = 1$ for all (k, k'). All residual deviations in the bottom plot of figure 2 from the objective $\Pi(k, k') = 1 \forall (k, k')$ are due to finite size effects. Again we also calculate the Hamming distance between the original and the synthetic matrix, giving $\rho = 0.93$. This value is similar to the one found previously, but now the macroscopic structure of the synthetic graph in terms of the degree correlations is considerably different from the underlying PPIN.

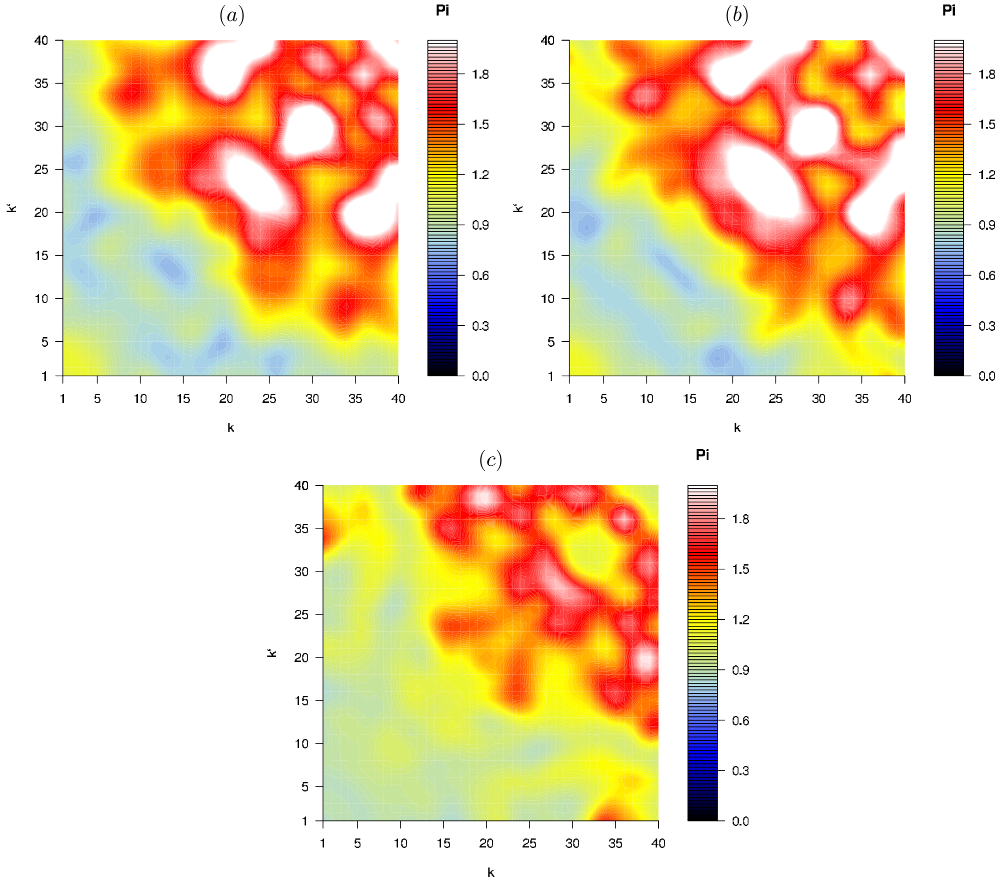

Figure 2. Results of Markov chain graph dynamics proposed in [17] tailored to generating equilibrium random graph ensembles with specific degree sequences *and* specific degree correlations. (*a*) Colour plot of the relative degree correlations $\Pi(k, k'|\mathbf{c})$ as measured in the *Escherichia coli* PPIN (here $N = 2457$ and $\langle k\rangle = 7.05$). (*b*) Colour plot of $\Pi(k, k'|\mathbf{c}')$ in the synthetic graph $\mathbf{c}'$ generated with Markov chain dynamics targeting the measured degree correlation if the PPIN, after 67, 147 accepted moves. (*c*) Colour plot of $\Pi(k, k'|\mathbf{c}')$ in the final graph generated with dynamics targeting $\Pi(k, k') = 1\forall(k, k')$, after 1 968 000 accepted moves.

It has been noted by several authors that most PPINs are disassortative, i.e. nodes with high degrees tend to connect with nodes with low degrees [4]. Measures of degree assortativity have been proposed in [3, 4, 33]. A conventional measure of assortativity is the correlation coefficient $(\langle kk'\rangle - \langle k\rangle\langle k'\rangle)/(\langle k^2\rangle - \langle k\rangle^2)$, calculated over the joint distribution $W(k, k')$ in (54). Degree assortativity has been shown to have important consequences on both the topology of a network and the process which it supports. In particular, it was shown that assortative networks are more resistant to random attacks, i.e. random vertex removal, whereas disassortative networks are less resistant [4]. It may be useful from a practical point of view to generate networks with a prescribed assortative character. This can again be achieved by using the measure (14), where the kernels $Q(k, k')$ are now chosen to produce assortative or disassortative graphs. In [17] it was shown that the kernel

$$Q(k, k') = \frac{|k - k'|^2}{2(\langle k^2\rangle - \langle k\rangle^2)} \tag{64}$$

tailors the ensemble (14) to the production, for large N, of graphs with degree correlations

$$\Pi(k,k') = \frac{\langle k\rangle(k-k')^2}{[\alpha_3 - 2\alpha_2 k + \alpha_1 k^2][\alpha_3 - 2\alpha_2 k' + \alpha_1 k'^2]}, \tag{65}$$

where the three coefficients α_ℓ are to be solved numerically from

$$\alpha_\ell = \sum_k \frac{k^\ell p(k)}{\alpha_3 - 2\alpha_2 k + \alpha_1 k^2}. \tag{66}$$

This degree correlation has a disassortative character. In fact, any kernel of the form

$$Q(k,k') = C^{-1}|k-k'|^n, \qquad n = 1, 2, \ldots, \tag{67}$$

with C such that $\sum_{k,k'} p(k)p(k')Q(k,k') = 1$ will tailor the ensemble (14), for large N, to the production of graphs with increasingly negative assortative coefficients as n increases. A prototype of an assortative kernel would be

$$Q(k,k') = \frac{1}{C}\frac{1}{1+|k-k'|^n} \tag{68}$$

where $C = \sum_{k,k'} p(k)p(k')[1 + |k-k'|^n]$. For sufficiently large N, the predicted values for $\Pi(k,k')$ follow from (19), where $F(k)$ is to be solved numerically from (20). As an example we generated two synthetic graphs, both with the same degree sequence as the PPIN of *Homo sapiens* (the experimental data used were taken from the human protein reference database (HPRD) [34]). In the first graph we enforced an assortative connectivity using (68) with $n = 1$, and in the second one a disassortative connectivity using (67) with $n = 1$. Both graphs were generated with the algorithm of [17], starting from the actual *Homo sapiens* PPIN. In figure 3(a) we show the colour plot of the relative degree correlations $\Pi(k,k')$ as measured in the *Homo sapiens* PPIN, and in figures 3(c) and (e) we show the same quantity in the two synthethic graphs generated. For comparison we also show (figures 3(b) and (d), respectively) the functional $\Pi(k,k')$ that are being targeted, via the kernels in (68) and (67).

5. Discussion

In this paper we have studied the tailoring of a particular structured random graph ensemble to real-world networks. We have first derived several mathematical properties of this ensemble, including information-theoretic properties, its Shannon entropy and the relation between its control parameters and the statistics and correlations of the degrees in the network to which the ensemble is tailored. We were then able to use the mathematical results in order to derive explicit and transparent mathematical tools with which to quantify structure in large real networks, define rational distance measures for comparing networks and for generating controlled null models as benchmark graphs. These tools are precise and based on information-theoretic principles, yet they take the form of fully explicit formulae (as opposed to implicit equations that require equilibration of extensive graph simulations). We therefore hope and anticipate that they will be particularly useful in bio-informatics; indeed a subsequent paper will be fully devoted to their application to a broad range of protein–protein interaction networks, involving multiple organisms and multiple experimental protocols [16].

Let us turn to the limitations of this study. Our work so far has focused on characterizing network structure macroscopically at the level of degree distributions and degree–degree correlations, and was limited to undirected networks and graphs. We therefore envisage two main directions in which the present theory could and should be developed further. The first, and relatively straightforward, one is generalization of the analysis to tailored *directed* random graph ensembles. Here one does not envisage insurmountable obstacles, and it would

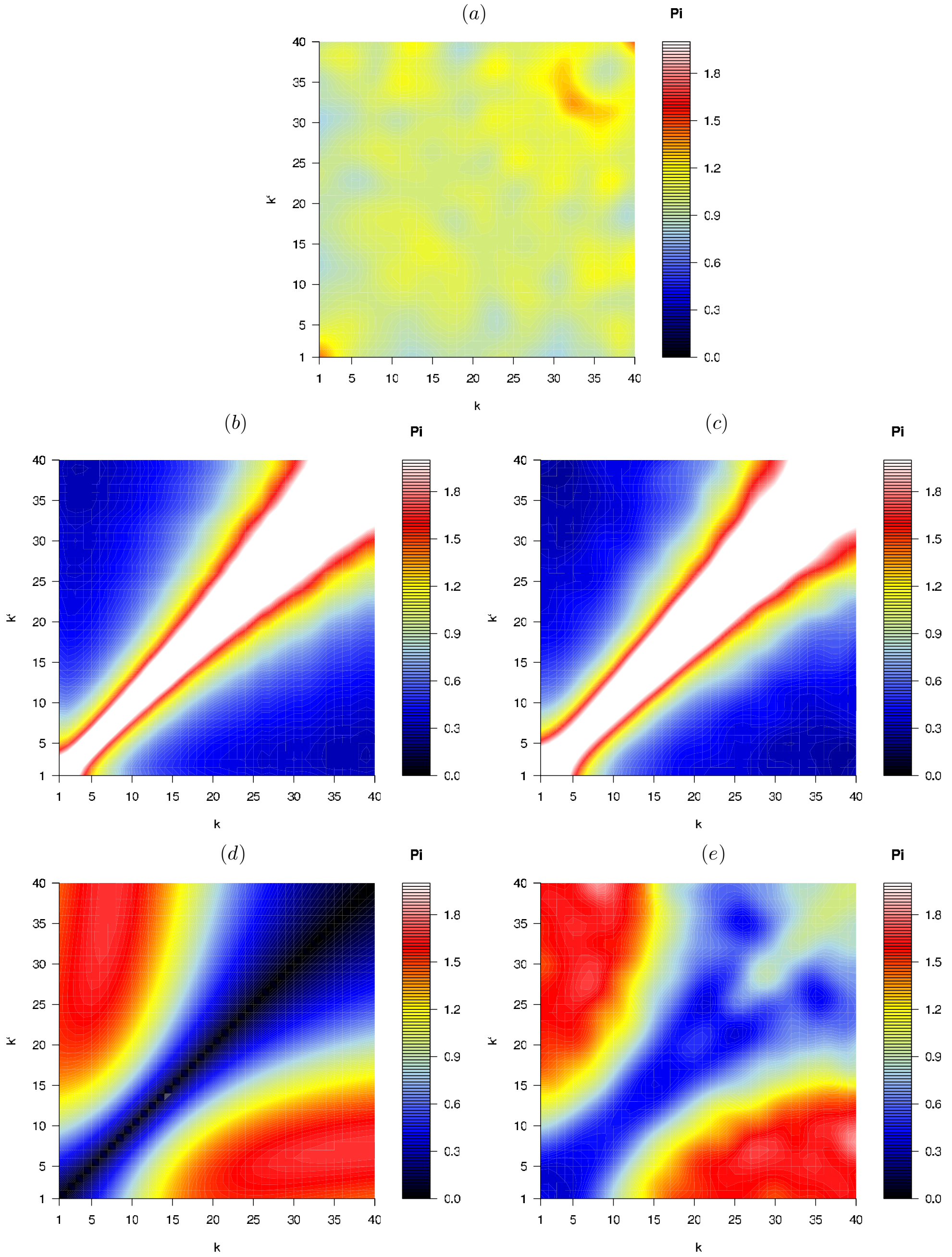

Figure 3. Colour plots of the relative degree correlations $\Pi(k, k')$ of networks which all have the degree sequence of the *Homo sapiens* (from the HPRD database) PPIN (with $N = 9463$ and $\langle k\rangle = 7.4$). (*a*) $\Pi(k, k'|\mathbf{c})$ as measured for the *Homo sapiens* PPIN, (*b*) the target assortative function $\Pi(k, k')$ given in (68), (*c*) the actual function $\Pi(k, k'|\mathbf{c}')$ measured after 203 441 accepted moves of the Markov chain in [17], (*d*) the target disassortative function $\Pi(k, k')$ given in (67), (*e*) the actual function $\Pi(k, k'|\mathbf{c}')$ measured after 266 763 accepted moves. These results confirm the efficiency of our canonical graph ensemble and its associated Markov chain algorithm, in generating controlled null models.

in bio-informatics open up the possibility of application to e.g. gene regulation networks. The second direction is towards the inclusion of measures of macroscopic structure that take account of loops, such as the distribution of length-three loops in which individual network nodes participate. Here the mathematical task is much more challenging since in entropy calculations it is no longer clear whether and how one can achieve factorization over nodes.

Acknowledgments

ACCC would like to thank Conrad Pérez-Vicente for stimulating discussions and the Engineering and Physical Sciences Research Council (UK) for support in the form of a Springboard Fellowship.

Appendix. Degree correlations in the random graph ensemble

In this appendix we prove the validity of the crucial relation (19), with $\Pi(k,k')$ as defined in (18) for the random graph ensemble (14):

$$\Pi(k,k') = \frac{\langle k\rangle}{p(k)p(k')kk'}\lim_{N\to\infty}\sum_{rs}\sum_{\mathbf{k}}\left[\prod_{\ell}p(k_\ell)\right]\delta_{k,k_r}\delta_{k',k_s}\frac{1}{N}\sum_{\mathbf{c}}\text{Prob}(\mathbf{c}|\mathbf{k},Q)c_{rs}. \tag{A.1}$$

Let us work out the sum over the graphs $\mathbf{c}$ in (A.1), using the integral representation $\delta_{nm} = (2\pi)^{-1}\int_0^{2\pi}\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega(n-m)}$ to deal with the N degree constraints $\delta_{k_i,k_i(\mathbf{c})}$. This introduces an N-fold integration over $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_N)\in[0,2\pi]^N$. With a modest amount of foresight we introduce the two abbreviations $\boldsymbol{\omega}\cdot\mathbf{k} = \sum_i\omega_i k_i$ and

$$W(\boldsymbol{\omega},\mathbf{k}) = \prod_{i<j}\left\{1+\frac{\overline{k}}{N}Q(k_i,k_j)[\mathrm{e}^{-\mathrm{i}(\omega_i+\omega_j)}-1]\right\} \tag{A.2}$$

These allow us to write

$$\begin{aligned}
\sum_{\mathbf{c}}\text{Prob}(\mathbf{c}|\mathbf{k},Q)c_{rs} &= \frac{\sum_{\mathbf{c}}c_{rs}\prod_{i<j}\left[\frac{\overline{k}}{N}Q(k_i,k_j)\delta_{c_{ij},1}+\left(1-\frac{\overline{k}}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right].\prod_i\delta_{k_i,k_i(\mathbf{c})}}{\sum_{\mathbf{c}}\prod_{i<j}\left[\frac{\overline{k}}{N}Q(k_i,k_j)\delta_{c_{ij},1}+\left(1-\frac{\overline{k}}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right].\prod_i\delta_{k_i,k_i(\mathbf{c})}}\\
&= \frac{\int\mathrm{d}\boldsymbol{\omega}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}}\sum_{\mathbf{c}}c_{rs}\prod_{i<j}\left\{\left[\frac{\overline{k}}{N}Q(k_i,k_j)\delta_{c_{ij},1}+\left(1-\frac{\overline{k}}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right]\mathrm{e}^{-\mathrm{i}c_{ij}(\omega_i+\omega_j)}\right\}}{\int\mathrm{d}\boldsymbol{\omega}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}}\sum_{\mathbf{c}}\prod_{i<j}\left\{\left[\frac{\overline{k}}{N}Q(k_i,k_j)\delta_{c_{ij},1}+\left(1-\frac{\overline{k}}{N}Q(k_i,k_j)\right)\delta_{c_{ij},0}\right]\mathrm{e}^{-\mathrm{i}c_{ij}(\omega_i+\omega_j)}\right\}}\\
&= \frac{\int\mathrm{d}\boldsymbol{\omega}\,W(\boldsymbol{\omega},\mathbf{k})\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}}\left[\frac{\frac{\overline{k}}{N}Q(k_r,k_s)\,\mathrm{e}^{-\mathrm{i}(\omega_r+\omega_s)}}{1+\frac{\overline{k}}{N}Q(k_r,k_s)[\mathrm{e}^{-\mathrm{i}(\omega_r+\omega_s)}-1]}\right]}{\int\mathrm{d}\boldsymbol{\omega}\,W(\boldsymbol{\omega},\mathbf{k})\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}}}\\
&= \frac{\overline{k}}{N}Q(k_r,k_s)[1+\mathcal{O}(N^{-1})]\frac{\int\mathrm{d}\boldsymbol{\omega}\,W(\boldsymbol{\omega},\mathbf{k})\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}-\mathrm{i}(\omega_r+\omega_s)}}{\int\mathrm{d}\boldsymbol{\omega}\,W(\boldsymbol{\omega},\mathbf{k})\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}}}.
\end{aligned} \tag{A.3}$$

We next expand the function $W(\boldsymbol{\omega},\mathbf{k})$, as defined in (A.2), in leading orders for large N, using the abbreviation $P(q,\omega|\boldsymbol{\omega},\mathbf{k}) = N^{-1}\sum_i\delta_{q,k_i}\delta(\omega-\omega_i)$:

$$\begin{aligned}
W(\boldsymbol{\omega},\mathbf{k}) &= \prod_{i<j}\exp\left\{\frac{\overline{k}}{N}Q(k_i,k_j)[\mathrm{e}^{-\mathrm{i}(\omega_i+\omega_j)}-1]+\mathcal{O}(N^{-2})\right\}\\
&= \exp\left\{\frac{\overline{k}}{2N}\sum_{ij}Q(k_i,k_j)[\mathrm{e}^{-\mathrm{i}(\omega_i+\omega_j)}-1]+\mathcal{O}(1)\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{\frac{1}{2}\bar{k}N\sum_{qq'}\int \mathrm{d}\omega\,\mathrm{d}\omega' P(q,\omega|\boldsymbol{\omega},\mathbf{k})P(q',\omega'|\boldsymbol{\omega},\mathbf{k})Q(q,q')\right.\\
&\quad\left.\times\ [\mathrm{e}^{-\mathrm{i}(\omega+\omega')}-1]+\mathcal{O}(1)\right\}. \qquad \text{(A.4)}
\end{aligned}
$$

We now insert the following representation of unity, for each combination of (q,ω):

$$
\begin{aligned}
1 &= \int \mathrm{d}P(q,\omega)\delta[P(q,\omega)-P(q,\omega|\boldsymbol{\omega},\mathbf{k})]\\
&= \int \frac{\mathrm{d}P(q,\omega)\,\mathrm{d}\hat{P}(q,\omega)}{2\pi/N}\,\mathrm{e}^{\mathrm{i}N\hat{P}(q,\omega)[P(q,\omega)-P(q,\omega|\boldsymbol{\omega},\mathbf{k})]}, \qquad \text{(A.5)}
\end{aligned}
$$

and convert the previous expression for $W(\boldsymbol{\omega},\mathbf{k})$ into the form of a functional integral, with a path integral measure $\{\mathrm{d}P\} = \prod_{q,\omega}[\mathrm{d}P(q,\omega)\Delta\omega/\sqrt{2\pi}]$ (where the values of $\omega \in [0,2\pi]$ are first discretized, with the discretization spacing $\Delta\omega$ sent to zero as soon as this is possible):

$$
\begin{aligned}
W(\boldsymbol{\omega},\mathbf{k}) &= \int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{\mathrm{i}N\sum_q\int\mathrm{d}\omega\hat{P}(q,\omega)P(q,\omega)-\mathbf{i}\sum_i\hat{P}(k_i,\omega_i)+\mathcal{O}(1)}\\
&\quad\times\,\mathrm{e}^{\frac{1}{2}\bar{k}N\sum_{qq'}\int\mathrm{d}\omega\,\mathrm{d}\omega' P(q,\omega)P(q',\omega')Q(q,q')[\mathrm{e}^{-\mathrm{i}(\omega+\omega')}-1]}\\
&= \int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}])-\mathbf{i}\sum_i\hat{P}(k_i,\omega_i)+\mathcal{O}(1)} \qquad \text{(A.6)}
\end{aligned}
$$

with

$$
\Psi[\{P,\hat{P}\}] = \mathrm{i}\sum_q\int_0^{2\pi}\mathrm{d}\omega\,\hat{P}(q,\omega)P(q,\omega) \qquad \text{(A.7)}
$$

$$
\Phi[\{P\}] = \frac{1}{2}\bar{k}\sum_{qq'}\int_0^{2\pi}\mathrm{d}\omega\,\mathrm{d}\omega' P(q,\omega)P(q',\omega')Q(q,q')[\mathrm{e}^{-\mathrm{i}(\omega+\omega')}-1]. \qquad \text{(A.8)}
$$

We can now integrate over the N-fold angles $\boldsymbol{\omega} \in [0,2\pi]^N$, and obtain

$$
\begin{aligned}
\int\mathrm{d}\boldsymbol{\omega}\,W(\boldsymbol{\omega},\mathbf{k})\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}'} &= \int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}])+\mathcal{O}(1)}\int\mathrm{d}\boldsymbol{\omega}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}'-\mathbf{i}\sum_i\hat{P}(k_i,\omega_i)}\\
&= \int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}])+\mathcal{O}(1)}\prod_i\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k_i'-\hat{P}(k_i,\omega)]} \qquad \text{(A.9)}
\end{aligned}
$$

and write the ratio of integrals in (A.3) as

$$
\begin{aligned}
&\frac{\int\mathrm{d}\boldsymbol{\omega}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}-\mathbf{i}(\omega_r+\omega_s)}W(\boldsymbol{\omega},\mathbf{k})}{\int\mathrm{d}\boldsymbol{\omega}\,\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\mathbf{k}}W(\boldsymbol{\omega},\mathbf{k})}\\
&= \frac{\int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}|\mathbf{k}])+\mathcal{O}(1)}\left\{\frac{[\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega(k_r-1)-\mathbf{i}\hat{P}(k_r,\omega)}][\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega(k_s-1)-\mathbf{i}\hat{P}(k_s,\omega)}]}{[\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega k_r-\mathbf{i}\hat{P}(k_r,\omega)}][\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega k_s-\mathbf{i}\hat{P}(k_s,\omega)}]}\right\}}{\int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}|\mathbf{k}])+\mathcal{O}(1)}} \qquad \text{(A.10)}
\end{aligned}
$$

with

$$
\Omega[\{\hat{P}\}|\mathbf{k}] = \frac{1}{N}\sum_i\log\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k_i-\hat{P}(k_i,\omega)]}. \qquad \text{(A.11)}
$$

Therefore we find upon combining the previous intermediate results that the quantity of interest (A.1) can be written in the following form:

$$
\begin{aligned}
\Pi(k,k') &= \frac{\langle k\rangle^2 Q(k,k')}{p(k)p(k')kk'}\lim_{N\to\infty}\sum_{\mathbf{k}}\left[\prod_{\ell}p(k_\ell)\right]\left[\frac{1}{N}\sum_r \delta_{k,k_r}\right]\left[\frac{1}{N}\sum_s \delta_{k',k_s}\right]\\
&\times\frac{\int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}|\mathbf{k}])+\mathcal{O}(1)}\left\{\frac{[\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega(k-1)-\hat{P}(k,\omega)]}][\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega(k'-1)-\hat{P}(k',\omega)]}]}{[\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k-\hat{P}(k,\omega)]}][\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k'-\hat{P}(k',\omega)]}]}\right\}}{\int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}|\mathbf{k}])+\mathcal{O}(1)}}\\
&= \frac{\langle k\rangle^2 Q(k,k')}{kk'}\\
&\times\lim_{N\to\infty}\frac{\int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}])+\mathcal{O}(1)}\left\{\frac{[\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega(k-1)-\hat{P}(k,\omega)]}][\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega(k'-1)-\hat{P}(k',\omega)]}]}{[\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k-\hat{P}(k,\omega)]}][\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k'-\hat{P}(k',\omega)]}]}\right\}}{\int\{\mathrm{d}P\,\mathrm{d}\hat{P}\}\,\mathrm{e}^{N(\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}])+\mathcal{O}(1)}}
\end{aligned}
\qquad\text{(A.12)}
$$

where

$$\Omega[\{\hat{P}\}] = \sum_{k''}p(k'')\log\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}[\omega k''-\hat{P}(k'',\omega)]}. \qquad\text{(A.13)}$$

We conclude from (A.12), in which the functional integrals can be done by steepest descent in the limit $N\to\infty$, that $\Pi(k,k')$ takes the form

$$\Pi(k,k') = Q(k,k')/F(k|Q)F(k'|Q) \qquad\text{(A.14)}$$

with

$$\frac{1}{F(k|Q)} = \frac{\langle k\rangle}{k}\frac{\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega(k-1)-\mathrm{i}\hat{P}(k,\omega)}}{\int\mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega k-\mathrm{i}\hat{P}(k,\omega)}}, \qquad\text{(A.15)}$$

and where the functions $P(k,\omega)$ and $\hat{P}(k,\omega)$ are to be solved from extremization of $\Psi[\{P,\hat{P}\}]+\Phi[\{P\}]+\Omega[\{\hat{P}\}]$, with the three functions given in (A.7), (A.8), (A.13), leading to the two coupled functional saddle-point equations $\delta[\Psi+\Phi]/\delta P=0$ and $\delta[\Psi+\Omega]/\delta\hat{P}=0$.

The last step in this appendix is to derive from the saddle-point equations an equation for the function $F(k|Q)$ in (A.14). Upon transforming $\exp[-\mathrm{i}\hat{P}(k,\omega)] = R(k,\omega)$, our saddle-point equations simplify to

$$R(k,\omega) = \exp\left\{\langle k\rangle\sum_{k'}\int\mathrm{d}\omega' P(k',\omega')Q(k,k')[\mathrm{e}^{-\mathrm{i}(\omega+\omega')}-1]\right\} \qquad\text{(A.16)}$$

$$P(k,\omega) = p(k)\frac{R(k,\omega)\mathrm{e}^{\mathrm{i}\omega k}}{\int\mathrm{d}\omega' R(k,\omega')\,\mathrm{e}^{\mathrm{i}\omega' k}}. \qquad\text{(A.17)}$$

Elimination of $P(k,\omega)$ from this set gives, using the identity $\int\mathrm{d}\omega P(k,\omega) = p(k)$,

$$
\begin{aligned}
R(k,\omega) &= \exp\left\{\langle k\rangle\sum_{k'}p(k')Q(k,k')\left[\mathrm{e}^{-\mathrm{i}\omega}\frac{\int\mathrm{d}\omega' R(k',\omega')\,\mathrm{e}^{\mathrm{i}\omega'(k'-1)}}{\int\mathrm{d}\omega' R(k',\omega')\,\mathrm{e}^{\mathrm{i}\omega' k'}}-1\right]\right\}\\
&= \exp\left\{\sum_{k'}p(k')Q(k,k')\,\mathrm{e}^{-\mathrm{i}\omega}k'/F(k'|Q)-G(k|Q)\right\},
\end{aligned}
\qquad\text{(A.18)}
$$

in which $F(k|Q)$ is defined in (A.15) and $G(k|Q) = \langle k\rangle\sum_{k'}p(k')Q(k,k')$. Insertion of our expression for $R(k,\omega)$ into (A.15), using $\exp[-\mathrm{i}\hat{P}(k,\omega)] = R(k,\omega)$, leaves us with an equation for $F(k|Q)$ only, from which the object $G(k|Q)$ simply drops out since it gives an identical pre-factor $\exp[-G(k|Q)]$ in both the numerator and the denominator of the formula for $F(k|Q)$:

$$\begin{aligned}\frac{1}{F(k|Q)} &= \frac{\langle k\rangle}{k}\frac{\int \mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega(k-1)}R(k,\omega)}{\int \mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega k}R(k,\omega)}\\ &= \frac{\langle k\rangle}{k}\frac{\sum_{m\geqslant 0}\frac{1}{m!}\left[\sum_{k'}p(k')Q(k,k')k'/F(k'|Q)\right]^m\int \mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega(k-1)-\mathrm{i}m\omega}}{\sum_{m\geqslant 0}\frac{1}{m!}\left[\sum_{k'}p(k')Q(k,k')k'F(k'|Q)\right]^m\int \mathrm{d}\omega\,\mathrm{e}^{\mathrm{i}\omega k-\mathrm{i}m\omega}}\\ &= \frac{\langle k\rangle}{k}\frac{\sum_{m\geqslant 0}\frac{1}{m!}\delta_{m,k-1}\left[\sum_{k'}p(k')Q(k,k')k'F(k'|Q)\right]^m}{\sum_{m\geqslant 0}\frac{1}{m!}\delta_{mk}\left[\sum_{k'}p(k')Q(k,k')k'/F(k'|Q)\right]^m}\\ &= \frac{\langle k\rangle}{\sum_{k'}p(k')Q(k,k')k'/F(k'|Q)}.\end{aligned}\tag{A.19}$$

Equivalently,

$$F(k|Q) = \langle k\rangle^{-1}\sum_{k'}p(k')k'Q(k,k')F^{-1}(k'|Q).\tag{A.20}$$

Note that the present derivation of the combined result (A.14), (A.20) also serves as the explicit proof of the validity of corrigendum [21].

References

[1] Albert R and Barabasi A L 2002 *Rev. Mod. Phys.* **74** 47–97
[2] Barabasi A L and Albert R 1999 *Science* **286** 509
[3] Pastor-Satorras R, Vazquez A and Vespignani A 2001 *Phys. Rev. Lett.* **87** 258701
[4] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701
[5] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[6] Newman M E J and Leicht E A 2007 *Proc. Natl Acad. Sci. USA* **104** 9564
[7] Maslov S and Sneppen K 2002 *Science* **296** 910
[8] Maslov S, Sneppen K and Zaliznyak A 2004 *Physica* A **333** 529–40
[9] Shen-Orr S S, Milo R, Mangan S and Alon U 2002 *Nature Genetics* **31** 64–8
[10] Junker B H and Schreiber F 2008 *Analysis of Biological Networks* (Hoboken, NJ: Wiley Series on Bioinformatics)
[11] Artzy-Randrup Y, Fleishman S J, Ben-Tal N and Stone L 2004 *Science* **305** 1107
[12] Pérez-Vicente C J and Coolen A C C 2008 *J. Phys. A: Math. Theor.* **41** 255003
[13] Bianconi G 2008 *Europhys. Lett.* **81** 28005
[14] Bianconi G, Coolen A C C and Vicente C J P 2008 *Phys. Rev.* E **78** 016114
[15] Bianconi G 2009 *Phys. Rev.* E **79** 036114
[16] Fernandes L P, Annibale A, Kleinjung J, Coolen A C C and Fraternali F in preparation
[17] Coolen A C C, De Martino A and Annibale A 2009 *J. Stat. Phys.* **136** 1035
[18] Dorogovtsev S N and Mendes J F 2003 *Evolution of Networks* (Oxford: Oxford University Press)
[19] Ivanic J, Wallqvist A and Reifman J 2008 *BMC Syst. Biol.* **2** 11
[20] Ivanic J, Wallqvist A and Reifman J 2008 *PLoS Comput. Biol.* **4** e1000114
[21] Pérez-Vicente C J and Coolen A C C 2009 *J. Phys. A: Math. Theor.* **42** 169801
[22] Erdos P and Gallai T 1960 *Mat. Lapok* **11** 264–74
[23] Rao A R, Jana R and Bandyopadhyay S 1996 *Indian J. Stat.* **58** 225
[24] Gkantsidis C, Mihail M and Zegura E 2003 *Proc. 5th Workshop on Algorithm Engineering and Experiments (ALENEX) (Siam)*
[25] Viger F and Latapy M 2005 *COCOON 2005, the 11th International Computing and Combinatorics Conference (Lecture Notes in Computer Science)* pp 440–449
[26] Chen Y, Diaconis P, Holmes S and Liu J S 2005 *J. Am. Stat. Assoc.* **100** 109
[27] Catanzaro M, Boguña M and Pastor-Satorras R 2005 *Phys. Rev.* E **71** 027103
[28] Serrano M A and Boguña M 2005 *Phys. Rev.* E **72** 036133
[29] Foster J G, Foster D V, Grassberger P and Paczuski M 2007 *Phys. Rev.* E **76** 046112
[30] Verhelst N D 2008 *Psychometrika* **73** 705
[31] Newman M E J, Strogatz S H and Watts D J 2001 *Phys. Rev.* E **64** 026118

[32] Arifuzzaman M *et al* 2006 *Genome Res* **16** 686–91
[33] Newman M E J 2003 *Phys. Rev. Lett.* **67** 026126
[34] Keshava Prasad T S *et al* 2009 *Nucleic Acids Res.* **37** D767 (Database issue)
[35] Simonis N *et al* 2009 *Nat. Methods* **6** 47–54
[36] Parrish J R *et al* 2007 *Genome Biol.* **8** R130
[37] Stark C *et al* 2006 *Nucleic Acids Res* **34** D535 (Database issue)
[38] Rain J C *et al* 2001 *Nature* **409** 211–5
[39] Rual J-F F *et al* 2005 7062 *Nature* **437** 1173–8
[40] Stelzl U *et al* 2005 *Cell* **122** 957–68
[41] Ewing R M *et al* 2007 *Mol. Syst. Biol.* **3** 89
[42] Shimoda Y *et al* 2008 *DNA Res* **15** 13–23
[43] Lacount D J *et al* 2005 *Nature* **438** 103–7
[44] Uetz P *et al* 2000 *Nature* **403** 623–7
[45] Ito T *et al* 2001 *Proc. Natl Acad. Sci. US* A **98** 4569–74
[46] Ho Y *et al* 2002 *Nature* **415** 180–3
[47] Mering von *et al* 2002 *Nature* **417** 399–403
[48] Gavin A C *et al* 2002 *Nature* **415** 141147
[49] Han J D J *et al* 2004 *Nature* **430** 88–93
[50] Gavin A C C *et al* 2002 *Nature* **415** 141–7
[51] Krogan N J *et al* 2006 *Nature* **440** 637–43
[52] Collins S R *et al* 2007 *Mol. Cell Proteomics* **6** 439–50
[53] Tarassov K *et al* 2008 *Science* **320** 1465–70
[54] Yu H *et al* 2008 *Science* **322** 104–10
[55] Sato S *et al* 2007 *DNA Res* **14** 207–16
[56] Titz B *et al* 2008 *PLoS ONE* **3** e2292