Interface Focus

Interface Focus (2011) **1**, 836–856
doi:10.1098/rsfs.2011.0050
Published online 5 October 2011

What you see is *not* what you get: how sampling affects macroscopic features of biological networks

A. Annibale[1],* and A. C. C. Coolen[1,2,3]

[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] *London Institute for Mathematical Sciences, 22 South Audley Street, London W1K 2NY, UK*

We use mathematical methods from the theory of tailored random graphs to study systematically the effects of sampling on topological features of large biological signalling networks. Our aim in doing so is to increase our quantitative understanding of the relation between true biological networks and the imperfect and often biased samples of these networks that are reported in public data repositories and used by biomedical scientists. We derive exact explicit formulae for degree distributions and degree correlation kernels of sampled networks, in terms of the degree distributions and degree correlation kernels of the underlying true network, for a broad family of sampling protocols that include random and connectivity-dependent node and/or link undersampling as well as random and connectivity-dependent link oversampling. Our predictions are in excellent agreement with numerical simulations.

Keywords: biological networks; detection bias/sampling; networks ensembles

1. INTRODUCTION

Networks are popular simplified representations of complex biological many-variable systems. The network representation reduces the complexity of the problem by retaining only information on which pairs of dynamical variables in a given system interact, leading to a graph in which the nodes (or vertices) represent the dynamical variables and the links (or edges) represent interacting pairs. If all interactions are symmetric under interchanging the two variables concerned, the resulting network is non-directed (as e.g. in protein–protein interaction networks (PPINs)). If some or all are non-symmetric, the network is directed (as e.g. in gene regulation networks). Present-day biological databases contain PPINs and gene regulation networks of many species, with typically in the order of $N \sim 10^3 - 10^4$ nodes each, measured and post-processed by various different techniques and protocols. However, in biology, the available experimental techniques do not sample the complete system, but only a finite fraction; for the human PPIN this fraction is presently (and inaccurately) estimated to be around 0.5 [1]. Furthermore, the sampling tends to be biased by which experimental method is used [2]. In order to use the available data wisely and reliably, it is vital that we understand in quantitative detail how the topological characteristics of a real network relate to those of a finite (biased or unbiased) random sample of this network. If, for instance, we observe that certain modules appear more often (or less often) than expected in certain cellular signalling networks, we need to be sure that this is not simply a consequence of sampling. The first studies of the effects of false negatives in the detection of links and/or nodes (i.e. bond and/or node undersampling) on network topologies focused on the relation between true and observed degree distributions, either analytically [3,4] or via numerical simulation [5], and found that undersampling changes qualitatively the shape of the degree distribution. Subsequent studies [6,7], based on numerical simulation, revealed the effects of undersampling on topological features other than the degree distribution, such as clustering coefficients, assortativity and the occurrence frequencies of local motifs. More recent publications were devoted to sampling of non-biological networks, such as the Internet [8] and bipartite networks [9]. So far all published studies on the effects of sampling have either been based on numerical simulations, or been restricted to the effects of sampling on a network's degree distribution. Moreover, there are only very few studies that considered connectivity-dependent sampling (e.g. [4]), and none that investigate the effects of false positive (i.e. oversampling). In the present paper, we use statistical mechanical methods from the theory of tailored random graphs to study systematically the effects of sampling on macroscopic topological features of large networks. We extend previous work in several ways. Firstly, we investigate the

*Author for correspondence (alessia.annibale@kcl.ac.uk).

One contribution of 9 to a Theme Issue 'Inference in complex systems'.

Received 30 May 2011
Accepted 30 August 2011

effect of sampling on macroscopic observables beyond the degree distribution, e.g. the joint degree distribution of connected node pairs from which one calculates quantities such as the assortativity. Secondly, we do this for both random and connectivity-dependent sampling of either nodes, links or both. Thirdly, we investigate not only network undersampling, but also the implications of false positives in the detection of links, i.e. bond oversampling. All our results are obtained analytically, and formulated in terms of explicit equations that express degree distributions and degree correlations of observed networks in terms of those of the underlying true networks. We test our analytical predictions against numerical simulations and find excellent agreement.

2. DEFINITIONS

2.1. Networks and sampling protocols

We consider non-directed networks or graphs. Each is defined by a symmetric matrix $\mathbf{c} = \{c_{ij}\}$, with $i,j = 1 \ldots N$ and with $c_{ij} \in \{0,1\}$ for all (i,j). Nodes i and j are connected if and only if $c_{ij} = 1$. We exclude self-interactions, i.e. $c_{ii} = 0$ for all i. The degree $k_i(\mathbf{c})$ of a node i is $k_i(\mathbf{c}) = \sum_j c_{ij}$, the degree distribution of graph $\mathbf{c}$ is $p(k|\mathbf{c}) = N^{-1}\sum_i \delta_{k,k_i}(\mathbf{c})$, and we abbreviate its degree sequence as $\mathbf{k}(\mathbf{c}) = (k_1(\mathbf{c}), \ldots, k_N(\mathbf{c}))$. Sampling stochastically an N-node graph $\mathbf{c}$ will result in observation of an N' node graph $\mathbf{c}'$. The relation between $\mathbf{c}'$ and $\mathbf{c}$ depends on the details of the sampling process. We use random variables $\sigma_i \in \{0,1\}$ to denote whether a true node i is observed, and $\tau_{ij} \in \{0,1\}$ whether a link (i,j) is observed (if nodes i and j are). In studying oversampling $\lambda_{ij} \in \{0,1\}$ will indicate whether an absent link is falsely reported as present. Thus:

$$\left.\begin{array}{rl} \text{node undersampling:} & c'_{ij} = \sigma_i \sigma_j c_{ij} \\ \text{bond undersampling:} & c'_{ij} = \tau_{ij} c_{ij} \\ \text{node and bond undersampling:} & c'_{ij} = \sigma_i \sigma_j \tau_{ij} c_{ij} \\ \text{bond oversampling:} & c'_{ij} = c_{ij} + (1 - c_{ij})\lambda_{ij} \end{array}\right\} \tag{2.1}$$

In a biological context, node oversampling (e.g. detecting a non-existent protein) would be less realistic, so will not be considered in this paper. Note that $N' = \sum_{i \le N} \sigma_i$. We take all sampling variables $\boldsymbol{\sigma} = \{\sigma_i\}$, $\boldsymbol{\tau} = \{\tau_{ij}\}$ and $\boldsymbol{\lambda} = \{\lambda_{ij}\}$ to be stochastically independent, with the *proviso* that $\tau_{ij} = \tau_{ji}$, $\lambda_{ij} = \lambda_{ji}$ and $\lambda_{ii} = 0$ (so sampled networks remain non-directed and without self-interactions). In random sampling their probabilities are functionally independent of the site indices; in connectivity-dependent sampling, the probabilities will depend on the degrees of the nodes involved. We conclude that the different types of sampling under equation (2.1) are all special cases of the following unified process:

$$c'_{ij} = \sigma_i \sigma_j [\tau_{ij} c_{ij} + (1 - c_{ij})\lambda_{ij}] \quad \forall (i < j) \tag{2.2}$$

with

$$\begin{aligned} P(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\lambda} | x, y, z) = & \prod_i [x(k_i)\delta_{\sigma_i,1} + (1 - x(k_i))\delta_{\sigma_i,0}] \\ & \times \prod_{i<j} [y(k_i, k_j)\delta_{\tau_{ij},1} + (1 - y(k_i, k_j))\delta_{\tau_{ij},0}] \\ & \times \prod_{i<j} \left[\frac{z(k_i, k_j)}{N}\delta_{\lambda_{ij},1} \right. \\ & \left. + \left(1 - \frac{z(k_i, k_j)}{N}\right)\delta_{\lambda_{ij},0}\right]. \end{aligned} \tag{2.3}$$

Here $x(k) \in [0,1]$ gives the likelihood that a node with degree k will be detected, $y(k,k') \in [0,1]$ the likelihood that a link connecting nodes with degrees (k,k') will be detected, and $z(k,k')/N \in [0,1]$ the likelihood that an absent bond will be falsely reported as present (the latter scales as N^{-1} to retain finite connectivity for large N). For random sampling, the control parameters in equation (2.3) would all be degree-independent, i.e. $x(k) = x$, $y(k,k') = y$ and $z(k,k') = z$. We note that, since non-existing nodes cannot give false negatives, we may always choose $x(0) = y(0,k) = y(k,0) = 0$ for all k. For connectivity-dependent sampling, plausible choices for the functional dependence of the control parameters on the local degree would be $x(k) = k/k_{\max}$ and/or $y(k,k') = kk'/k^2_{\max}$ and/or $z(k,k') = kk'/k^2_{\max}$, since high-degree nodes and links connecting high-degree nodes are more likely to be reported.

2.2. Macroscopic characterization of network structure

To control analytically the topological properties of the networks to which our sampling protocols (2.1) are applied, we consider the following maximum entropy ensemble, tailored for large N, to the production of graphs with prescribed degrees and prescribed degree correlations:

$$\begin{aligned} p(\mathbf{c}) = \frac{1}{Z_N}\left[\prod_i \delta_{k_i,k_i(\mathbf{c})}\right] \prod_{i<j} & \left[\frac{\bar{k}}{N}\frac{W(k_i, k_j)}{p(k_i)p(k_j)}\delta_{c_{ij},1} \right. \\ & \left. + \left(1 - \frac{\bar{k}}{N}\frac{W(k_i, k_j)}{p(k_i)p(k_j)}\right)\delta_{c_{ij},0}\right] \end{aligned} \tag{2.4}$$

with $p(k) = N^{-1}\sum_i \delta_{k,k_i}$ and $\bar{k} = \sum_k p(k)k$, and with Z_N the appropriate normalization constant. Graphs generated according to ensemble (2.4) will have $\mathbf{k}(\mathbf{c}) = \mathbf{k}$, $p(k|\mathbf{c}) = p(k)$ and $\sum_{\mathbf{c}} p(\mathbf{c})\, W(k,k'|\mathbf{c}) = W(k,k')$, where $W(k, k'|\mathbf{c}) = (N\bar{k})^{-1}\sum_{ij} c_{ij}\delta_{k,k_i}\delta_{k',k_j}$ is the joint degree distribution of connected node pairs. Apart from the information in $\mathbf{k}$ and $W(k,k')$, the ensemble (2.4) is unbiased; see Annibale *et al.* [10] for derivations of its information-theoretic properties, Coolen *et al.* [11,12] for Monte Carlo Markov Chain (MCMC) algorithms via which its graphs can be generated numerically and for a review on the topic. The remainder of this paper is devoted to calculate analytically how in large networks, with given degree sequences and given degree correlations (i.e. as those typically generated via ensemble (2.4)), sampling affects the macroscopic topological characteristics $p(k)$ and $W(k,k')$. To be specific, we calculate the average connectivity, the degree distribution and the degree correlation function, after sampling from large graphs drawn from ensemble (2.4), in terms of the sampling characteristics $\{x(k), y(k,k'), z(k,k')\}$,[1]

$$\bar{k}(x,y,z) = \lim_{N\to\infty} \sum_{\mathbf{c}} p(\mathbf{c}) \left\langle \frac{\sum_{ij} c'_{ij}}{\sum_i \boldsymbol{\sigma}_i} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}, \tag{2.5}$$

$$p(k|x,y,z) = \lim_{N\to\infty} \sum_{\mathbf{c}} p(\mathbf{c}) \left\langle \frac{\sum_i \boldsymbol{\sigma}_i \boldsymbol{\delta}_{k,\sum_j c_{ij'}}}{\sum_i \boldsymbol{\sigma}_i} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} \tag{2.6}$$

and

$$W(k,k'|x,y,z) = \lim_{N\to\infty} \sum_{\mathbf{c}} p(\mathbf{c}) \left\langle \frac{\sum_{ij} c'_{ij} \boldsymbol{\delta}_{k,\sum_\ell c'_{i\ell}} \boldsymbol{\delta}_{k',\sum_\ell c'_{j\ell}}}{\sum_{ij} c'_{ij}} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} \tag{2.7}$$

with c'_{ij} as defined in equation (2.2) and $\langle \cdot \rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}$ denoting averages over the sampling parameters distribution (2.3). The denominators are simplified trivially, using the independence of the sampling variables and the definition of $W(k,k'|\mathbf{c})$, since

$$\begin{aligned} \frac{1}{N}\sum_i \boldsymbol{\sigma}_i &= \frac{1}{N}\sum_i x(k_i) + \mathcal{O}(N^{-1/2}) \\ &= \sum_k p(k)x(k) + \mathcal{O}(N^{-1/2}) \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} \frac{1}{N}\sum_{ij} c'_{ij} &= \frac{1}{N}\sum_{ij} x(k_i)x(k_j)\left[\frac{z(k_i,k_j)}{N}\right. \\ &\quad \left. + c_{ij}\left[y(k_i,k_j) - \frac{z(k_i,k_j)}{N}\right]\right] + \mathcal{O}(N^{-1/2}) \\ &= \sum_{kk'} x(k)x(k')\{p(k)p(k')z(k,k') \\ &\quad + \bar{k}W(k,k')y(k,k')\} + \mathcal{O}(N^{-1/2}). \end{aligned} \tag{2.9}$$

We may therefore write

$$\bar{k}(x,y,z) = \frac{\sum_{qq'} x(q)x(q')[p(q)p(q')z(q,q') + \bar{k}W(q,q')y(q,q')]}{\sum_q p(q)x(q)}, \tag{2.10}$$

$$p(k|x,y,z) = \frac{\lim_{N\to\infty}\sum_{\mathbf{c}} p(\mathbf{c}) \left\langle N^{-1}\sum_i \boldsymbol{\sigma}_i \boldsymbol{\delta}_{k,\sum_j c_{ij'}} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}}{\sum_q p(q)x(q)} \tag{2.11}$$

and

$$W(k,k'|x,y,z) = \frac{\lim_{N\to\infty}\sum_{\mathbf{c}} p(\mathbf{c}) \left\langle N^{-1}\sum_{ij} c'_{ij} \boldsymbol{\delta}_{k,\sum_\ell c'_{i\ell}} \boldsymbol{\delta}_{k',\sum_\ell c'_{j\ell}} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}}{\bar{k}(x,y,z)\sum_q p(q)x(q)}. \tag{2.12}$$

In the following sections, we will calculate analytically the observables (2.10), (2.11) and (2.12) and will test our theoretical results against numerical simulations. To this purpose, we will sample from reasonably large graphs (either synthetically generated or real biological PPINs) and we will measure the degree distribution and the degree correlations after sampling.[2] These will be compared with the analytically calculated post-sampling degree distribution and degree correlations resulting from averages over graph ensembles asymptotically tailored to the production of graphs with the same degree sequence and degree correlations as the graph instances used for numerical simulations. The extent to which theoretical predictions and numerical simulations agree will provide an indication of how well, for reasonably large graphs, the behaviour of degree distribution and degree correlations under sampling is captured by averages of such quantities over the corresponding maximum entropy ensembles.

3. EFFECTS OF SAMPLING ON DEGREE DISTRIBUTIONS

3.1. Connection between observed degree distributions and degree correlations

We note that in the case of connectivity-dependent sampling, the average degree (2.10) in the observed graph will generally depend not only on the degree distribution of the original graph, but also on the latter's degree correlations. Hence, our decision to use the graph ensemble (2.4) for the present study. The observed distributions $p(k|x,y,z)$ and $W(k,k'|x,y,z)$ in expressions (2.11) and (2.12) are connected via a simple identity, as are $p(k)$ and $W(k,k')$ in the original graph $\mathbf{c}$:

$$\begin{aligned} W(k|x,y,z) &= \sum_{k'} W(k,k'|x,y,z) \\ &= \lim_{N\to\infty} \frac{k}{\bar{k}(x,y,z)} \sum_{\mathbf{c}} p(\mathbf{c}) \\ &\quad \times \left\langle \frac{1}{N}\sum_i \boldsymbol{\sigma}_i \boldsymbol{\delta}_{k,\sum_\ell c'_{i\ell}} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} \\ &= \frac{k}{\bar{k}(x,y,z)} p(k|x,y,z). \end{aligned} \tag{3.1}$$

So for large N, we need to calculate, in principle, only $W(k,k'|x,y,z)$, as $p(k|x,y,z)$ follows via identity (3.1). Alternatively, since for random sampling, $p(k|x,y,z)$ can be found analytically with little effort, the identity (3.1) can be used for verifying the result of our calculation of expression (2.12).

[1] One should expect that macroscopic physical observables such as $p(k|\mathbf{c})$ and $W(k,k'|\mathbf{c})$ are self-averaging, and can therefore be calculated, to leading order in N, in terms of their expectation values (2.5), (2.6) and (2.7) over the ensemble (2.4).

[2] The software used in this paper for generating and sampling from networks is available from the authors upon request (in standard C).

3.2. Degree distribution for random sampling

Calculating $p(k|x,y,z)$ is only straightforward for random sampling, irrespective of whether the source graph is generated according to ensemble (2.4), since, in that case, expression (2.11) can be made to factorize over the sampling variables by writing the Kronecker-δ in integral form. In order to appreciate the roles played by the different ingredients of expression (2.11), we first write it in the form $p(k|x,y,z) = \lim_{N\to\infty} \sum_{\mathbf{c}} p(\mathbf{c}) p_N(k|x,y,z;\ \mathbf{c})$, with

$$\begin{aligned}
p_N(k|x,y,z;\mathbf{c}) &= \frac{1}{\sum_q p(q)x(q) + \mathcal{O}(N^{-1/2})} \frac{1}{N} \sum_i \left\langle \sigma_i \delta_{k,\sum_j c_{ij'}} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} \\
&= \frac{1}{\sum_q p(q)x(q)} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}k\omega} \left\langle \sigma_i \mathrm{e}^{-\mathrm{i}\omega \sum_j \sigma_i \sigma_j [\tau_{ij} c_{ij} + (1-c_{ij})\lambda_{ij}]} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} + \mathcal{O}(N^{-1/2}) \\
&= \frac{1}{\sum_q p(q)x(q)} \frac{1}{N} \sum_i x(k_i) \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}k\omega} \prod_{j\neq i} \left\{ 1 + x(k_j) \left[\left\langle \mathrm{e}^{-\mathrm{i}\omega[\tau_{ij}c_{ij} + (1-c_{ij})\lambda_{ij}]} \right\rangle_{\boldsymbol{\tau},\boldsymbol{\lambda}} - 1 \right] \right\} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \\
&= \frac{1}{\sum_q p(q)x(q)} \frac{1}{N} \sum_i x(k_i) \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}k\omega} \prod_{j\neq i} \left\{ 1 + x(k_j)(\mathrm{e}^{-\mathrm{i}\omega} - 1) \left[\frac{z(k_i,k_j)}{N} + c_{ij} y(k_i,k_j) \right] \right\} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \\
&= \frac{1}{\sum_q p(q)x(q)} \sum_{q'} x(q') \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}k\omega + (\mathrm{e}^{-\mathrm{i}\omega}-1)\sum_{k'} p(k')x(k')z(q',k')} \\
&\quad \times \frac{1}{N} \sum_i \delta_{q',k_i} \exp\left(\sum_{k'} \log\{1 + x(k')y(q',k')(\mathrm{e}^{-\mathrm{i}\omega} - 1)\} \sum_j \delta_{k',k_j} c_{ij} \right) + \mathcal{O}(N^{-1/2}).
\end{aligned} \tag{3.2}$$

For random sampling protocols, where $x(k) = x$, $y(k,k') = y$ and $z(k,k') = z$, this expression immediately simplifies to the transparent result

$$\begin{aligned}
p_N(k|x,y,z;\mathbf{c}) &= \sum_{k'} p(k'|\mathbf{c}) \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}k\omega + xz(\mathrm{e}^{-\mathrm{i}\omega}-1)} \\
&\quad \times \{1 + xy(\mathrm{e}^{-\mathrm{i}\omega} - 1)\}^{k'} + \mathcal{O}(N^{-1/2}) \\
&= \sum_{k'} p(k'|\mathbf{c}) \sum_{n\geq 0} \frac{z^n}{n!} \sum_{m=0}^{k'} \binom{k'}{m} x^{n+m} y^m \\
&\quad \times \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}k\omega} (\mathrm{e}^{-\mathrm{i}\omega} - 1)^{n+m} + \mathcal{O}(N^{-1/2}) \\
&= \sum_{k'} p(k'|\mathbf{c}) \sum_{n\geq 0} \frac{z^n}{n!} \sum_{m=0}^{k'} \binom{k'}{m} \binom{m+n}{k} \\
&\quad \times x^{n+m} y^m (-1)^{n+m-k} I(k \leq n+m) \\
&\quad + \mathcal{O}(N^{-1/2}) \\
&= x^k \sum_{k'} p(k'|\mathbf{c}) \sum_{n\geq 0} \frac{z^n}{n!} \sum_{m=0}^{k'} \binom{k'}{m} \binom{m+n}{k} \\
&\quad \times x^{n+m-k} y^m (-1)^{n+m-k} I(k \leq n+m) \\
&\quad + \mathcal{O}(N^{-1/2}),
\end{aligned} \tag{3.3}$$

in which $I(S)$ is the indicator function (i.e. $I(S) = 1$ if S is true, otherwise $I(S) = 0$). The observed average degree (2.10) for random sampling is, as expected,

$$\bar{k}(x,y,z) = x(z + y\bar{k}). \tag{3.4}$$

Formula (3.3) simplifies further for various special cases. For instance:

— *Random bond and/or node undersampling*, i.e. $z = 0$:

$$\begin{aligned}
p(k|x,y,0) &= (xy)^k \sum_{k'\geq k} p(k') \binom{k'}{k'-k} (1 - xy)^{k'-k} \\
&= \frac{(xy)^k}{k!} \sum_{\ell\geq 0} p(k+\ell) \frac{(k+\ell)!}{\ell!} (1 - xy)^{\ell}.
\end{aligned} \tag{3.5}$$

This implies that if we sample from a graph with Poissonian degree distribution, i.e. $p(k) = \bar{k}^k\, \mathrm{e}^{-k}/k!$, then the degree distribution of the sampled graph will be

$$\begin{aligned}
p(k|x,y,0) &= \frac{(xy)^k}{k!} \sum_{k'\geq k} \frac{\bar{k}^{k'}\, \mathrm{e}^{-k}}{(k'-k)!} (1 - xy)^{k'-k} \\
&= \frac{(\bar{k}xy)^k \mathrm{e}^{-\bar{k}xy}}{k!},
\end{aligned} \tag{3.6}$$

i.e. again a Poissonian distribution, but with a reduced average degree $\bar{k}(x,y,0) = xy\bar{k}$. This recovers earlier results of Stumpf and co-workers [3,4]. We note also that equation (3.5) is invariant under exchanging x and y, so sampling all nodes and a fraction $x = \xi$ of the bonds is equivalent to sampling all bonds and a fraction $y = \xi$ of the nodes. We show in §4.1 that this equivalence between bonds and nodes under random undersampling also holds for the degree correlations. In figure 1, we show the predicted degree distributions (3.5) together with the corresponding results of

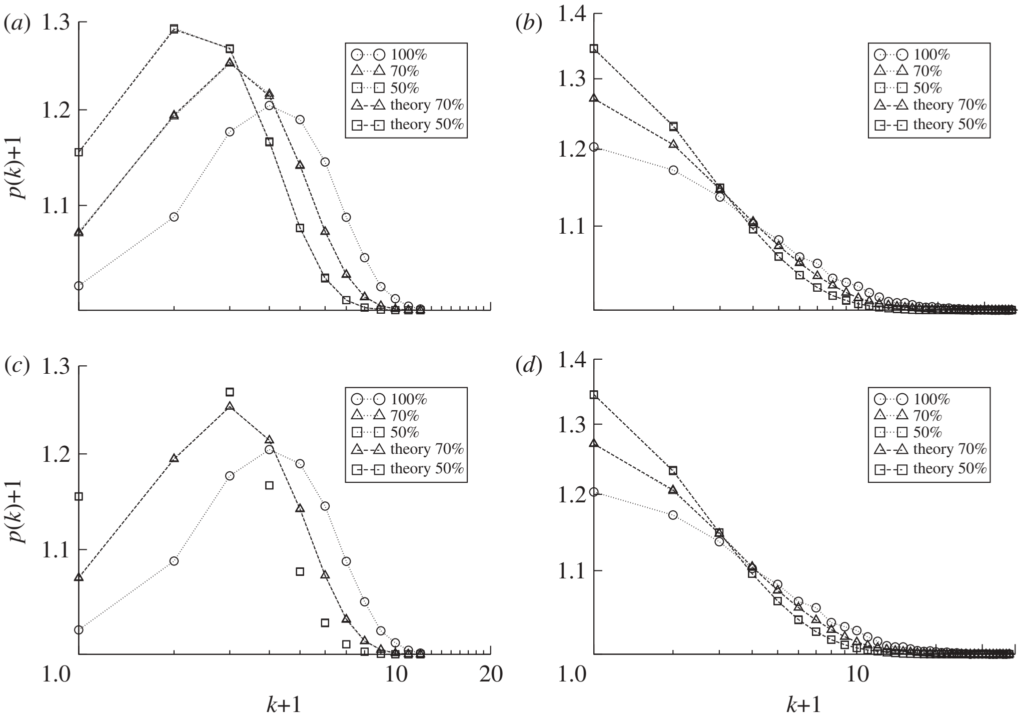

Figure 1. Effect of random node undersampling (a,b) and bond undersampling (c,d) on the degree distribution of synthetically generated networks with size $N = 3512$, average connectivity $\bar{k} = 3.72$ and Poissonian degree distribution (a,c) or power-law distributed degrees (b,d). Different symbols correspond to different fractions of sampled nodes (0.5, 0.7 and 1 as shown in the legend) and predicted values (symbols connected by dotted lines) lay on the top of data points from simulations (symbols connected by dashed lines), obtained by averaging over 50 samples.

numerical simulation of the sampling process, for synthetically generated networks with size $N = 3512$ and average connectivity $\bar{k} = 3.72$ (as in the biological PPIN of *Caenorhabditis elegans* [13]) and Poissonian and power-law degree distributions. The agreement between theory and experiment is perfect.

— *Random bond oversampling*, i.e. $x = y = 1$:

$$
\begin{aligned}
p(k|1,1,z) &= \sum_{k'} p(k') \sum_{n\geq 0} \frac{z^n}{n!} \sum_{m=0}^{k'} \binom{k'}{m} \\
&\qquad \times \binom{m+n}{k} (-1)^{n+m-k} I(k \leq n+m) \\
&= \sum_{k'} p(k') \sum_{n\geq 0} \frac{(-z)^n}{n!} \\
&\qquad \times \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^{\ell} \delta_{\ell,k-k'} \\
&= \sum_{k'\leq k} p(k') \sum_{s\geq 0} \frac{z^{k-k'+s}(-1)^s}{s!(k-k')!} \\
&= \sum_{\ell=0}^{k} \frac{p(k-\ell)\mathrm{e}^{-z} z^{\ell}}{\ell!}. \qquad (3.7)
\end{aligned}
$$

As with random undersampling, we observe that sampling from a graph with Poissonian degree distribution, i.e. $p(k) = \bar{k}^k\,\mathrm{e}^{-\bar{k}}/k!$ leads to a sampled graph that is again Poissonian, but now with average degree $\bar{k}(1,1,z) = z + \bar{k}$:

$$
\begin{aligned}
p(k|1,1,z) &= \mathrm{e}^{-z} \sum_{q\leq k} \frac{\bar{k}^q\,\mathrm{e}^{-\bar{k}}}{q!} \frac{z^{k-q}}{(k-q)!} = \frac{z^k \mathrm{e}^{-(\bar{k}+z)}}{k!} \\
&\qquad \times \sum_{q\leq k} \frac{k!}{q!(k-q)!} \left(\frac{\bar{k}}{z}\right)^q \\
&= \frac{z^k \mathrm{e}^{-(\bar{k}+z)}}{k!} \left(1 + \frac{\bar{k}}{z}\right)^k = \frac{\mathrm{e}^{-(\bar{k}+z)}(\bar{k}+z)^k}{k!}.
\end{aligned} \qquad (3.8)
$$

Results from numerical simulations applied to Poissonian and preferential attachment networks are shown in figure 2 together with the corresponding theoretical predictions. Again the agreement between theory and experiment is perfect.

3.3. *Degree distribution for connectivity-dependent sampling*

In the case of connectivity-dependent sampling, where $x(k)$, $y(k,k')$ and $z(k,k')$ are no longer all degree-independent, one can no longer evaluate (3.9) without knowledge of the degree–degree correlations in the sources graph **c**. However, the average (3.9) over the graph ensemble with controlled degree correlations is still feasible. In appendix A, we calculate the marginal (A 24) of the expected kernel $W(k,k'|x,y,z)$ for the

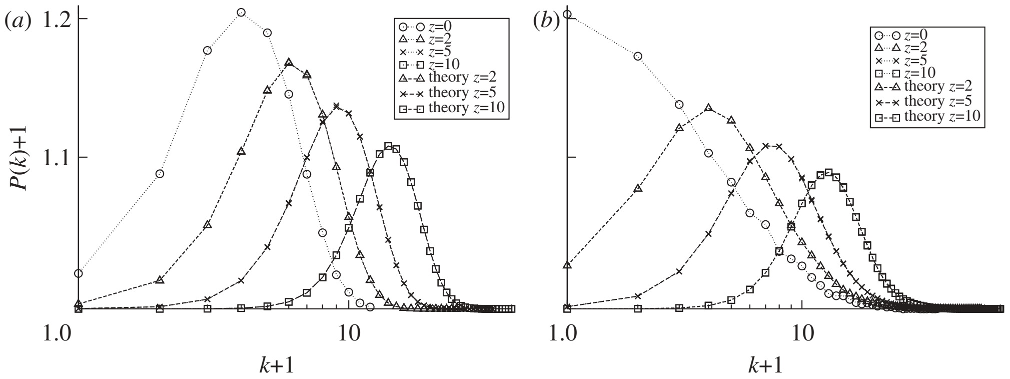

Figure 2. Effect of random bond oversampling on (*a*) the degree distribution of synthetic Poissonian graphs and (*b*) synthetic power-law graphs, both with size $N = 3512$ and average connectivity $\bar{k} = 3.72$. Different symbols correspond to different fractions z/N of 'false positive' bonds, with $z = 0, 2, 5, 10$ as shown in the legend. The theoretically predicted values (symbols connected by dotted lines) are found to lay perfectly on top of the data points from simulations (symbols connected by dashed lines), obtained by averaging over 100 samples.

sampled network, from which we obtain $p(k|x,y,z)$ via the connection (3.1). One always has $p(0|x,y,z) = 0$, whereas for $k > 0$:

$$p(k|x,y,z) = \frac{\sum_q x(q)p(q)\{a(q)\mathcal{J}(k|q) + qb(q)\mathcal{L}(k|q)\}}{k\sum_q p(q)x(q)} \tag{3.9}$$

with

$$\mathcal{J}(k|q) = \mathrm{e}^{-a(q)} \sum_{n=0}^{\min\{k-1,q\}} \binom{q}{n} \frac{a^{k-1-n}(q)}{(k-1-n)!} b^n(q) \times (1-b(q))^{q-n}, \tag{3.10}$$

$$\mathcal{L}(k|q) = \mathrm{e}^{-a(q)} \sum_{n=0}^{\min\{k-1,q-1\}} \binom{q-1}{n} \times \frac{a^{k-1-n}(q)}{(k-1-n)!} b^n(q)(1-b(q))^{q-1-n} \tag{3.11}$$

and

$$a(q) = \sum_{q'\geq 0} p(q')x(q')z(q,q'),$$
$$b(q) = \frac{\bar{k}}{qp(q)} \sum_{q'\geq 0} x(q')y(q,q')\,W(q,q'). \tag{3.12}$$

The average connectivity $\bar{k}(x,y,z)$, as given in observables (2.10), is easily obtained from equation (3.9) using normalization of the conditional probabilities $\mathcal{J}(k|q)$ and $\mathcal{L}(k|q)$

$$\bar{k}(x,y,z) = \sum_k kp(k|x,y,z) = \frac{\sum_q x(q)p(q)[a(q)+qb(q)]}{\sum_q p(q)x(q)}. \tag{3.13}$$

Let us now work out these results for the 'natural' types of connectivity-dependent samplings, where the likelihood of observing nodes or links is proportional to the degrees of the nodes involved, with $\alpha \in [0,1]$:

— *Connectivity-dependent node undersampling*, i.e. $x(k) = \alpha k/k_{\max}$, $y(k,k') = 1$, $z(k,k') = 0$:
Here, we have

$$a(q) = 0,$$
$$qb(q)\mathcal{L}(k|q) = k\binom{q}{k} b^k(q)(1-b(q))^{q-k} I(q \geq k) \tag{3.14}$$

and

$$\sum_q p(q)x(q) = \frac{\alpha\bar{k}}{k_{\max}},$$
$$b(q) = \frac{\alpha\bar{k}}{qp(q)k_{\max}} \sum_{q'>0} q'\,W(q,q'). \tag{3.15}$$

This leads to

$$p(k|\alpha,1,0) = \sum_{q\geq k} \frac{qp(q)}{\bar{k}} \binom{q}{k} \times \left(\frac{\alpha\bar{k}}{qp(q)k_{\max}} \sum_{q'>0} q'\,W(q,q')\right)^k \times \left(\frac{\alpha\bar{k}}{qp(q)k_{\max}} \sum_{q'>0} q'\,W(q,q')\right)^{q-k} \tag{3.16}$$

and

$$\bar{k}(\alpha,1,0) = \frac{\alpha}{k_{\max}} \sum_{qq'>0} qq'\,W(q,q') = \frac{\alpha}{k_{\max}} \frac{\overline{k^{(3)}}}{\bar{k}}, \tag{3.17}$$

where $\overline{k^{(3)}} = N^{-1}\sum_{ijk\ell} c_{ij}c_{jk}c_{k\ell}$ is the average number of paths of length 3.

— *Connectivity-dependent bond undersampling*, i.e. $x(k) = 1$, $y(k,k') = \alpha\, kk'/k_{\max}^2$, $z(k,k') = 0$:
This choice leads again to equation (3.14), while

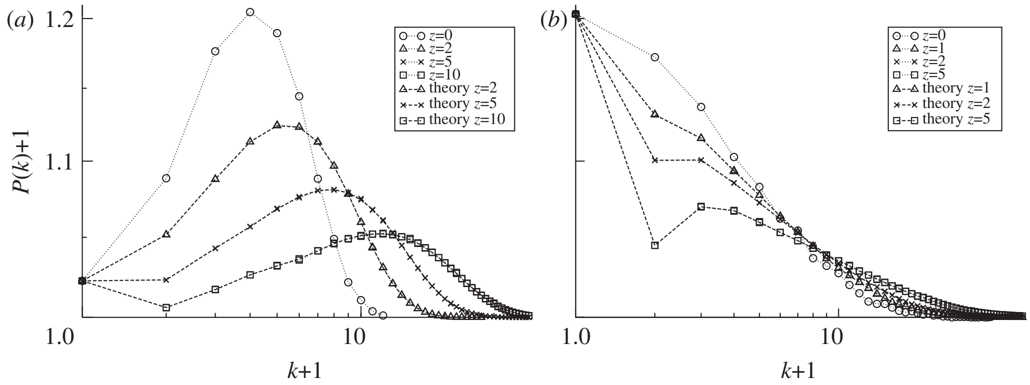

Figure 3. Effect of connectivity-dependent bond oversampling (i.e. $x(k)=1$, $y(k,k')=1$, $z(k,k')=\alpha kk'/k_{\max}^2$) on (*a*) the degree distribution of synthetic Poissonian graphs and (*b*) synthetic power-law graphs, both with size $N=3512$ and average connectivity $\bar{k}=3.72$. Different symbols correspond to different values of $z=\alpha\,\bar{k}^2/k_{\max}^2=0,2,5,10$, as shown in the legend. Theoretically predicted values (symbols connected by dotted lines) are found to lay perfectly on top of the data points from simulations (symbols connected by dashed lines), obtained by averaging over 100 samples.

equations (3.15) are now replaced by

$$\sum_q p(q)x(q)=1,$$

$$b(q)=\frac{\alpha\bar{k}}{p(q)k_{\max}^2}\sum_{q'>0} q'\,W(q,q'). \tag{3.18}$$

Hence, one gets

$$p(k|1,\alpha,0)=\sum_{q\geq k} p(q)\binom{q}{k}\left(\frac{\alpha\bar{k}}{p(q)k_{\max}^2}\sum_{q'>0} q'\,W(q,q')\right)^k \times\left(\frac{\alpha\bar{k}}{p(q)k_{\max}^2}\sum_{q'>0} q'\,W(q,q')\right)^{q-k} \tag{3.19}$$

and

$$\bar{k}(1,\alpha,0)=\frac{\alpha\bar{k}}{k_{\max}^2}\sum_{qq'>0} qq'\,W(q,q')=\frac{\alpha}{k_{\max}^2}\overline{k^{(3)}}. \tag{3.20}$$

— *Connectivity-dependent bond oversampling*, i.e. $x(k)=y(k,k')=1$, $z(k,k')=\alpha\;kk'/k_{\max}^2$:
Here, we have

$$a(q)=\sum_{q'\geq 0} p(q')z(q,q')=\frac{\alpha\bar{k}}{k_{\max}^2}q,\quad b(q)=1,$$

$$\sum_q p(q)x(q)=1, \tag{3.21}$$

$$\mathcal{L}(k|q)=\mathrm{e}^{-a(q)}\frac{a^{k-q}(q)}{(k-q)!}I(q\leq k) \tag{3.22}$$

and

$$\mathcal{J}(k|q)a(q)=\mathrm{e}^{-a(q)}\frac{a^{k-q}(q)}{(k-1-q)!}I(q\leq k-1) \equiv (k-q)\mathcal{L}(k|q). \tag{3.23}$$

Substituting into equations (3.9) and (3.13) yields

$$p(k|1,1,\alpha)=\sum_q p(q)\mathcal{L}(k|q) =\sum_q p(q)\mathrm{e}^{-q\alpha\bar{k}/k_{\max}^2}\frac{(q\alpha\bar{k}/k_{\max}^2)^{k-q}}{(k-q)!} \tag{3.24}$$

and

$$\bar{k}(1,1,\alpha)=\bar{k}+\frac{\alpha\,\bar{k}^2}{k_{\max}^2}. \tag{3.25}$$

In figure 3, we show the predicted degree distribution (3.24) together with the corresponding results from numerical simulations of the connectivity-dependent bond oversampling process.

3.4. Summary

We have seen that the degree distributions of large sampled networks can be calculated and written explicitly in terms of the topological characteristics of the true network, for random and connectivity-dependent under- and oversampling. From the resulting equations, we can draw the following conclusions:

— Sampling generally affects the shape of the degree distribution of a network, with the exception of a Poissonian distribution (as for Erdos–Renyi graphs), where the sampled network will only have a rescaled average degree compared with the original. This result is consistent with findings in Stumpf and co-workers [3,4].

— The degree distribution observed after *random* node undersampling of a network is identical to that following *random* bond undersampling, for any large graph, if the two (node- or bond-) sampling probabilities are identical.

— In contrast, *connectivity-dependent* node undersampling (where the probability of observing a node is proportional to its degree) generally leads to a network with a degree distribution that is different from the one that would result from *connectivity-dependent* bond undersampling (where the probability of observing a bond is proportional to the degrees of the two attached nodes).

4. EFFECTS OF SAMPLING ON DEGREE CORRELATION FUNCTION

In appendix A, we calculate the degree correlation function $W(k,k'|x,y,z)$ of large networks that are sampled according to the general protocol (2.2), from graphs generated from ensemble (2.4). The resulting, expressed in terms of the topological properties $p(k)$ and $W(k,k')$ of the true network, is

$$W(k,k'|x,y,z) = \frac{\sum_{q,q'>0} x(q)x(q')\{p(q)p(q')z(q,q')\mathcal{J}(k|q)\mathcal{J}(k'|q') + \bar{k}\,W(q,q')y(q,q')\mathcal{L}(k|q)\mathcal{L}(k'|q')\}}{\bar{k}(x,y,z)\sum_q p(q)x(q)} \tag{4.1}$$

with $\bar{k}(x,y,z)$ as given in observables (2.10), two conditional distributions $\mathcal{J}(k|q)$ and $\mathcal{L}(k|q)$ defined in equations (3.10) and (3.11), and with the short-hands $a(q)$ and $b(q)$ defined in equation (3.12). We will now work out this general result for the most common types of sampling, *viz.* node undersampling, bond undersampling and bond oversampling, including both random- and connectivity-dependent protocols.

4.1. Degree correlations for random sampling

For random sampling protocols where $x(q)=x$, $y(q,q')=y$ and $z(q,q')=z$, one has $a(q)=xz$, $b(q)=xy$ and $\mathcal{L}(k|q)=\mathcal{J}(k|q-1)$, so (4.1) simplifies immediately to

$$W(k,k'|x,y,z) = \frac{\sum_{q,q'>0}\{zp(q)p(q')\mathcal{J}(k|q)\mathcal{J}(k'|q') + y\bar{k}\,W(q,q')\mathcal{J}(k|q-1)\mathcal{J}(k'|q'-1)\}}{z+y\bar{k}} \tag{4.2}$$

with

$$\mathcal{J}(k|q) = \mathrm{e}^{-xz}x^{k-1}\sum_{n=0}^{\min\{k-1,q\}}\binom{q}{n} \times \frac{z^{k-1-n}(q)}{(k-1-n)!}\,y^n(1-xy)^{q-n}. \tag{4.3}$$

Formula (4.2) simplifies further for various special cases:

— *Random node and/or bond undersampling*, i.e. $z=0$. Here, we obtain

$$\mathcal{J}(k|q) = \binom{q}{k-1}(xy)^{k-1}(1-xy)^{q-k+1}I(q\geq k-1) \tag{4.4}$$

so equation (4.2) reduces to

$$W(k,k'|x,y,0) = \sum_{q\geq k}\sum_{q'\geq k'} W(q,q')\binom{q-1}{k-1}\binom{q'-1}{k'-1} \times (xy)^{k+k'-2}(1-xy)^{q+q'-k-k'}. \tag{4.5}$$

We note that $W(x,y,0)$, like equation (3.5) previously, is symmetric under exchanging x and y, i.e. node and bond random undersampling lead to the same degree correlations. Therefore, the equivalence between the two samplings is now fully established for large graphs drawn from ensemble (2.4).

Equation (4.5) clearly shows that sampling from graphs in which degree correlations are present will generally affect those correlations, even in Poissonian networks, in spite of the fact that there the degree distribution is only changed via a reduction of the average degree. Conversely, if we sample from graphs without degree correlations, i.e. for which $W(k,k') = W(k)\,W(k') = p(k)p(k')kk'/\bar{k}^2$, equation (4.5) reveals that the degree correlation function in the sampled graph factorizes in the product of its marginals as well, i.e. $W(k,k'|x,y,0) = W(k|x,y,0)\,W(k'|x,y,0)$. This means that random bond and/or node undersampling from graphs without degree correlations does not generate any degree correlations.

In order to observe how sampling protocols affect degree correlations, we will monitor, instead of $W(k,k')$ itself, the normalized kernel $\Pi(k,k') = W(k,k')/W(k)\,W(k')$, which will by definition equal unity in the absence of degree correlations. Any deviation from $\Pi(k,k')=1$ will thus signal the presence of degree correlations. We show the predicted degree correlations in the case of random bond undersampling, together with the corresponding results of numerical simulations, for Poissonian and power-law graphs, in figures 4 and 5, respectively. In figure 6, we show numerical results and theoretical predictions for random node undersampling from Poissonian and power-law graphs. Results for random bond undersampling from the real, biological PPIN of *C. elegans*

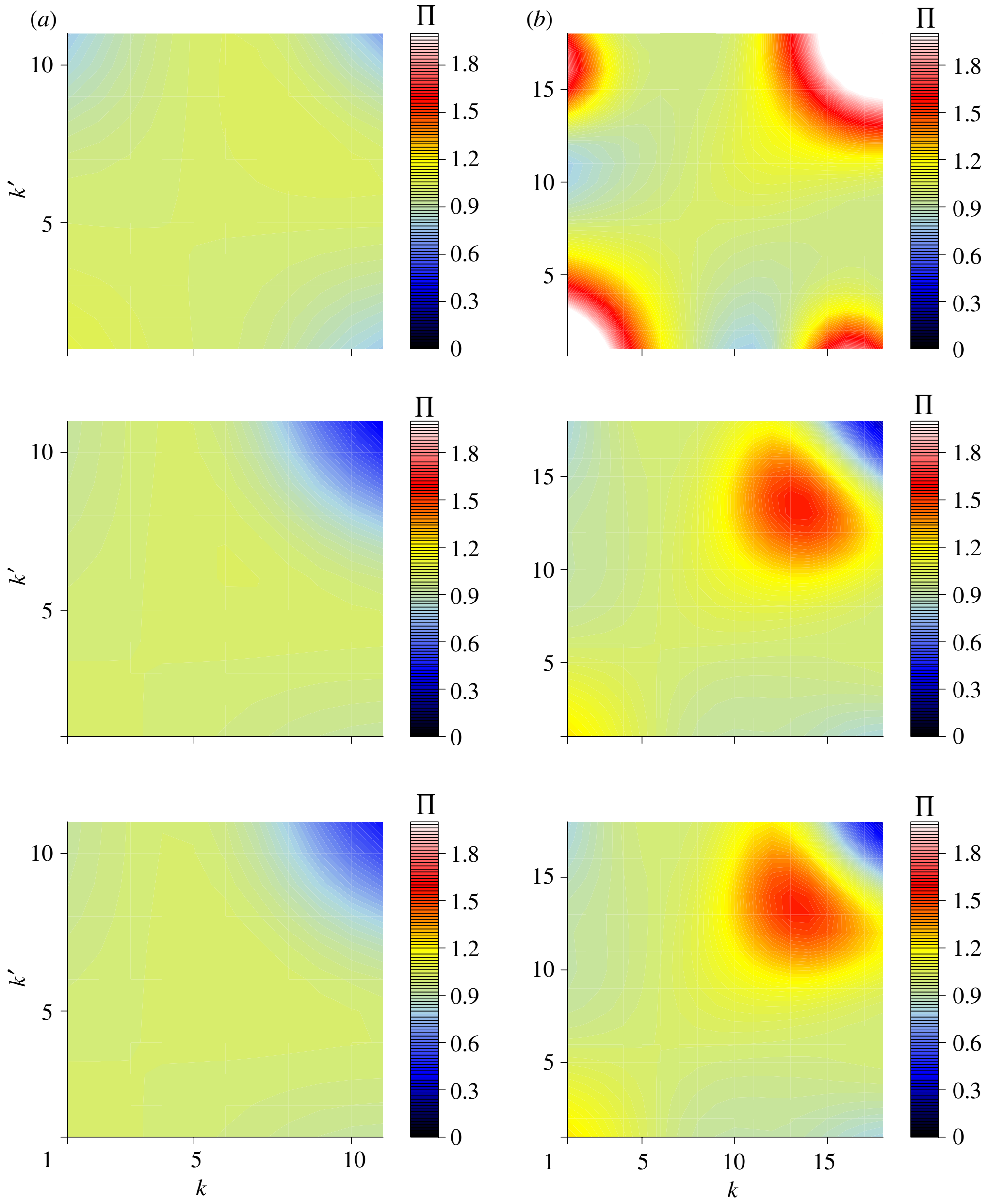

Figure 4. Normalized degree correlation function $\Pi(k,k') = W(k,k')/W(k)\,W(k')$ of two synthetically generated Poissonian graphs with $N = 3512$ and $\bar{k} = 3.72$ and different degree correlations (as shown in panels (a,b) respectively) before (top panels) and after (middle panels) sampling, a fraction $y = 0.7$ of the bonds of the original graphs (data result from averaging over 10^4 samples) and their respective theoretical predictions (bottom panels).

are shown in figure 7 (left panels). The agreement between theory and experiment is very satisfactory; deviations are small and consistent with finite size effects. As explained earlier, this confirms *a posteriori* that the performance of the biological network to be sampled (here *C. elegans*) is similar to the average behaviour of the maximum entropy ensemble (2.4), where $p(k)$ and $W(k,k')$ are the degree distribution and the degree correlation function of the biological PPIN, respectively. As a consequence, the biological network can be realistically approximated by a member of such ensemble. As an additional test, we generate synthetically a member of the maximum entropy ensemble asymptotically tailored to the production of graphs with the same degree sequence and degree correlations as the PPIN of *C. elegans* by using the MCMC algorithm proposed in Coolen *et al.* [11]. The degree correlations of the resulting graph are shown in the top right panel of figure 7 and are in good agreement with the degree correlations of the PPIN that are being targeted (top left panel). Note that the Hamming distance between the biological PPIN and the synthetically generated graph is 0.93, so similarity in degree correlations is not consequence of similarity in the connectivity matrices.[3] Theoretical

[3] The Hamming distance between two graphs $\mathbf{c}$ and $\mathbf{c}'$ of size N and average degree $\bar{k}$ is defined as $\rho(\mathbf{c},\mathbf{c}') = (2N\bar{k})^{-1}\sum_{ij}|c_{ij} - c_{ij}'|$ and takes values between 0 ($c_{ij} = c_{ij}'\ \forall i,j$) and 1 ($c_{ij} \neq c_{ij}'\ \forall i,j$).

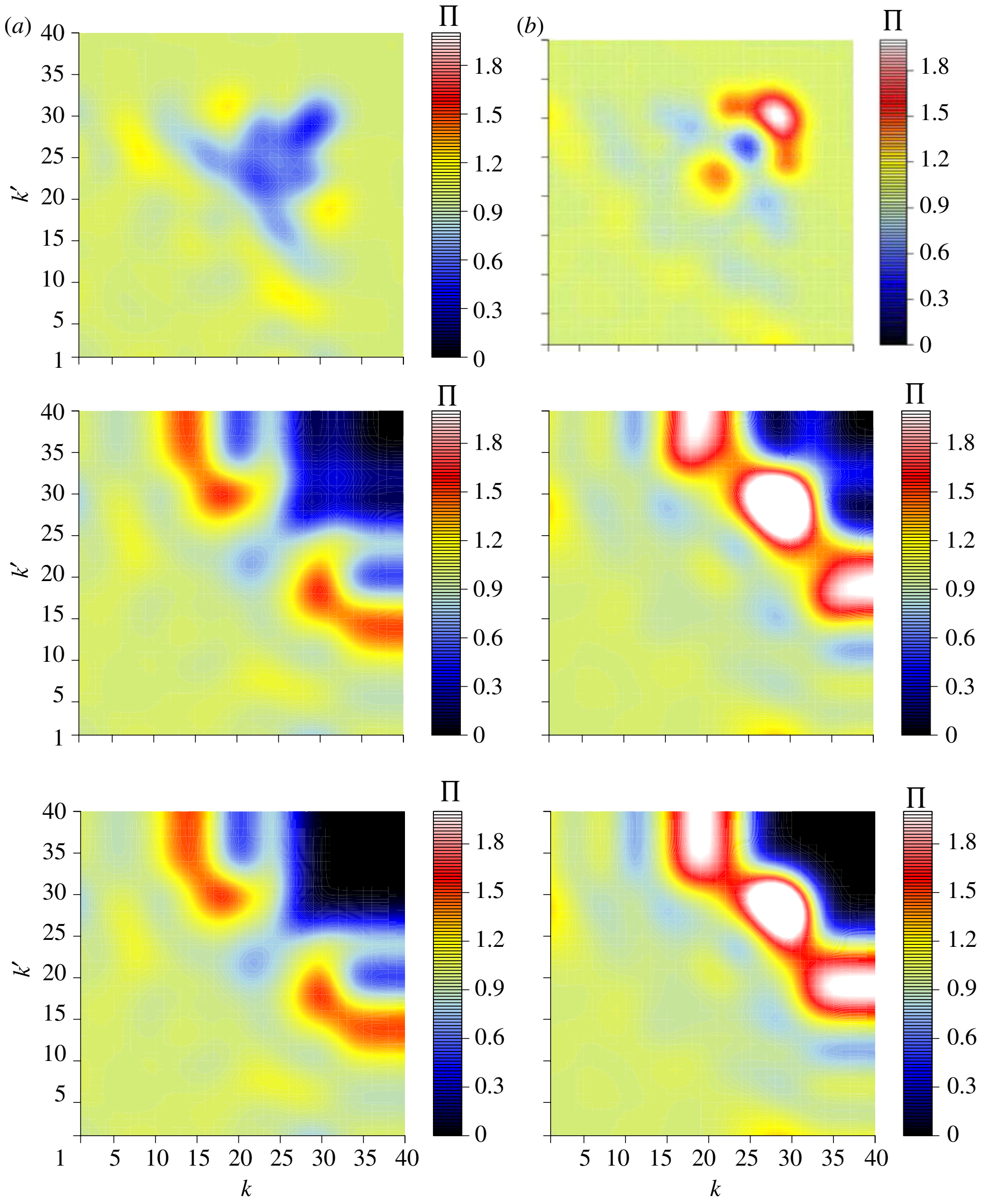

Figure 5. Normalized degree correlation function $\Pi(k,k')$ of two synthetically generated power-law graphs with $N = 3512$ and $\bar{k} = 3.72$ and different degree correlations (as shown in panels (*a*,*b*) respectively) before (top panels) and after (middle panels) sampling a fraction $y = 0.9$ of the bonds of the original graph (data result from averaging over 10^4 samples) and their theoretical prediction (bottom panels).

and numerical results for random bond undersampling from such randomized counterpart of *C. elegans* are shown in figure 7 (middle and bottom right panels).

— *Random bond oversampling*, i.e. $x = y = 1$.

Here, we have

$$\mathcal{J}(k|q) = \mathrm{e}^{-z}\frac{z^{k-1-q}}{(k-1-q)!}I(k \geq q+1), \qquad (4.6)$$

so using our earlier result from equation (3.7)

$$p(k|1,1,z) = \mathrm{e}^{-z}\sum_{q=0}^{k} p(q)\frac{\mathrm{e}^{k-q}}{(k-q)!}, \qquad (4.7)$$

we may write

$$\sum_{q\geq 0} p(q)\mathcal{J}(k|q) = p(k-1|1,1,z), \qquad (4.8)$$

which leads to the transparent expression

$$W(k,k'|1,1,z) = \frac{z}{\bar{k}+z}\left[p(k-1|1,1,z)p(k'-1|1,1,z) + \mathrm{e}^{-2z}\frac{\bar{k}}{z}\sum_{q,q'=1}^{k,k'} W(q,q')\frac{z^{k-q}}{(k-q)!}\frac{z^{k'-q'}}{(k'-q')!}\right]. \qquad (4.9)$$

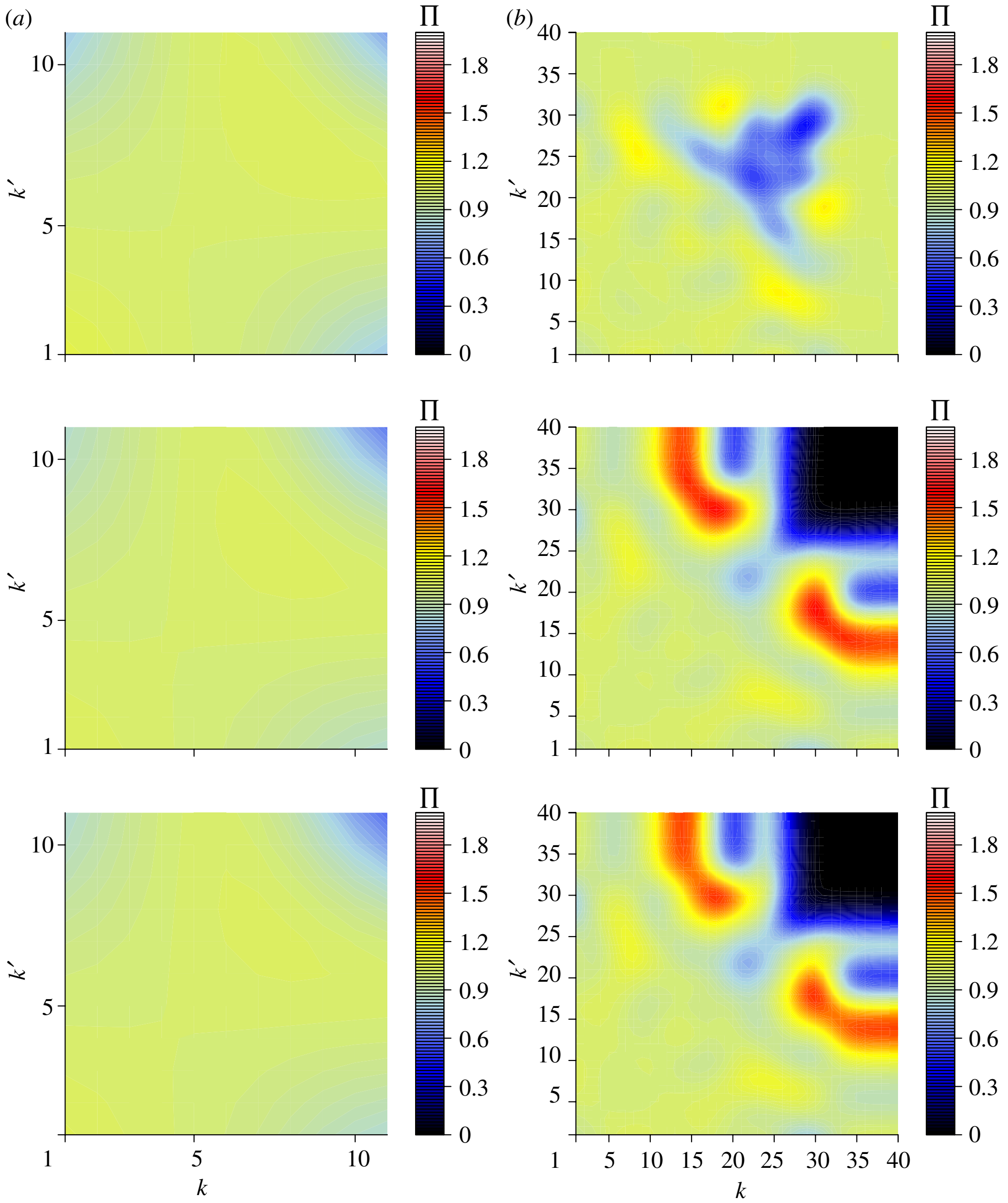

Figure 6. Normalized degree correlation function $\Pi(k,k')$ of (*a*) synthetically generated Poissonian and (*b*) power-law graphs with $N = 3512$ and $\bar{k} = 3.72$ before (top panels) and after (middle panels) sampling a fraction $x = 0.9$ of the nodes of the original graph (data result from averaging over 10^4 samples) and their theoretical prediction (bottom panels).

We note for later that substituting equation (4.8) into equation (3.9) and bearing in mind that $\mathcal{L}(k|q) = \mathcal{J}(k|q-1)$, $a(q) = z$ and $b(q) = 1$, we have

$$p(k|1,1,z) = \frac{1}{k}\left[zp(k-1|1,1,z) + \sum_{q\geq 1} p(q)q\mathcal{J}(k|q-1)\right], \qquad (4.10)$$

which yields

$$p(k-1|1,1,z) = \frac{\bar{k}}{z}p(k|1,1,z) - \frac{\bar{k}}{z}\sum_{q\geq 1} W(q)\mathcal{J}(k|q-1), \qquad (4.11)$$

where $W(k) = kp(k)/\bar{k}$.

We now study the effects of oversampling on graphs without degree correlations. Denoting

$$S_z(k) = \sum_{q\geq 1} W(q)\mathcal{J}(k|q-1), \qquad (4.12)$$

which is z-dependent via the function $\mathcal{J}$, we may rewrite (4.9) as

$$\begin{aligned} W(k,k'|1,1,z) = \frac{z}{\bar{k}+z}\Bigg[&\left(\frac{k}{z}p(k|1,1,z) - \frac{\bar{k}}{z}S_z(k)\right) \\ &\times\left(\frac{k'}{z}p(k'|1,1,z) - \frac{\bar{k}}{z}S_z(k')\right) \\ &+ \frac{\bar{k}}{z}\sum_{q,q'\geq 1} W(q,q')\mathcal{J}(k|q-1)\mathcal{J}(k'|q'-1)\Bigg]. \end{aligned} \qquad (4.13)$$

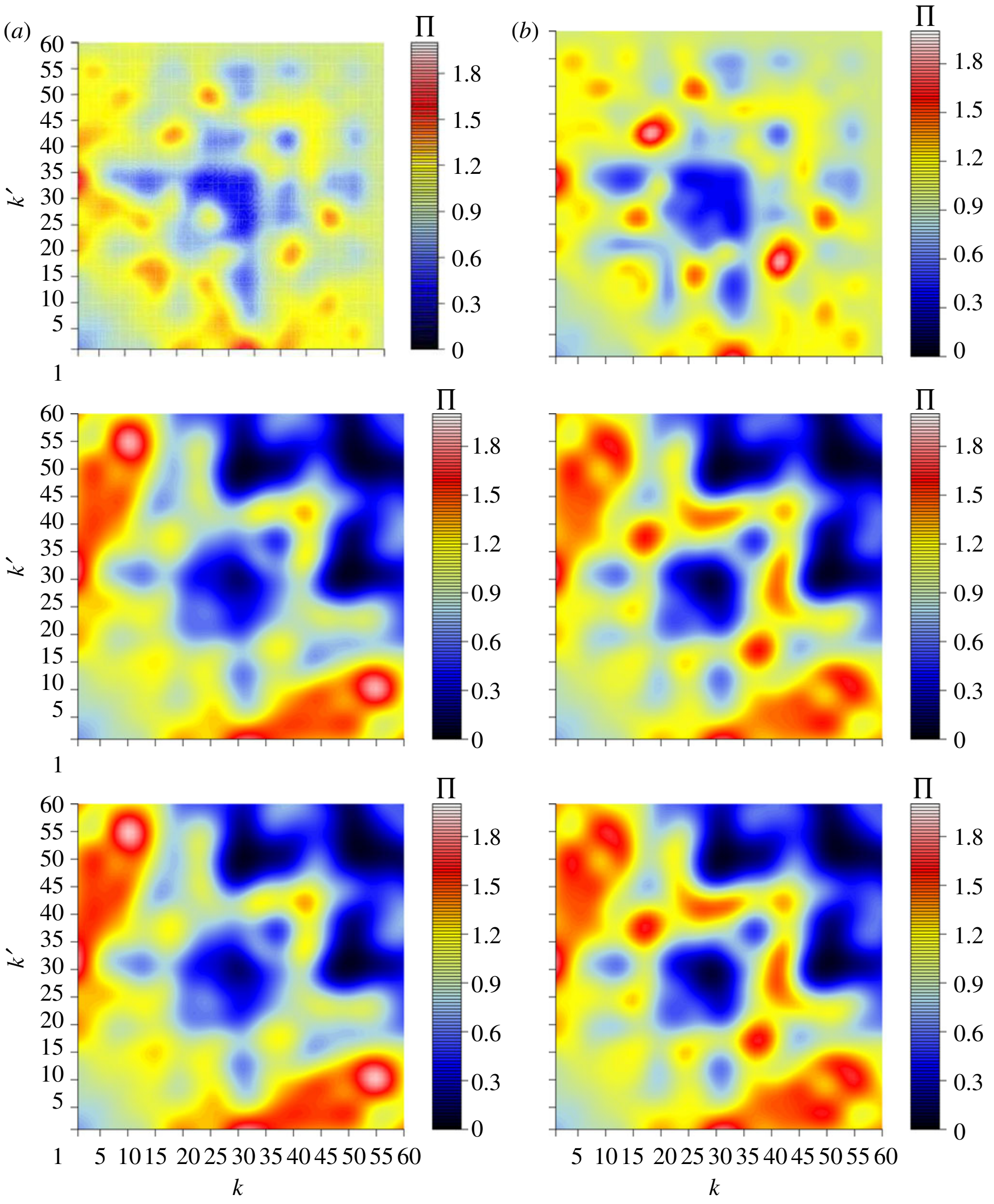

Figure 7. Normalized degree correlation function $\Pi(k,k')$ for (*a*) the biological PPIN of *C. elegans* and (*b*) one synthetically generated member of its corresponding tailored graph ensemble, before (top panels) and after (middle panels) sampling a fraction $x = 0.9$ of the bonds of the original graph and their theoretical prediction (bottom panels). For both networks, $N = 3512$ and $\bar{k} = 3.72$ and data resulted from averaging over 10^4 samples.

If the original graph has no degree correlation, i.e.

$$W(q,q') = W(q)\,W(q') = p(q)p(q')\,\frac{qq'}{\bar{k}^2}, \tag{4.14}$$

the sampled graph will have degree correlation

$$\begin{aligned}
W(k,k'|1,1,z) &= \frac{z}{\bar{k}+z}\left[\left(\frac{k}{z}p(k|1,1,z) - \frac{\bar{k}}{z}S_z(k)\right)\left(\frac{k'}{z}p(k'|1,1,z) - \frac{\bar{k}}{z}S_z(k')\right) + \frac{\bar{k}}{z}S_z(k)S_z(k')\right]\\
&= \frac{\bar{k}}{z}\left[\frac{\bar{k}+z}{\bar{k}}\,W(k|1,1,z)\,W(k'|1,1,z) + S_z(k)S_z(k') - W(k|1,1,z)S_z(k') - W(k'|1,1,z)S_z(k)\right]\\
&= \frac{\bar{k}}{z}\left[(W(k|1,1,z) - S_z(k))\;(W(k'|1,1,z) - S_z(k')) + \frac{z}{\bar{k}}\,W(k|1,1,z)\,W(k'|1,1,z)\right]\\
&= W(k|1,1,z)\,W(k'|1,1,z) + \frac{\bar{k}}{z}(W(k|1,1,z) - S_z(k))(W(k'|1,1,z) - S_z(k')),
\end{aligned} \tag{4.15}$$

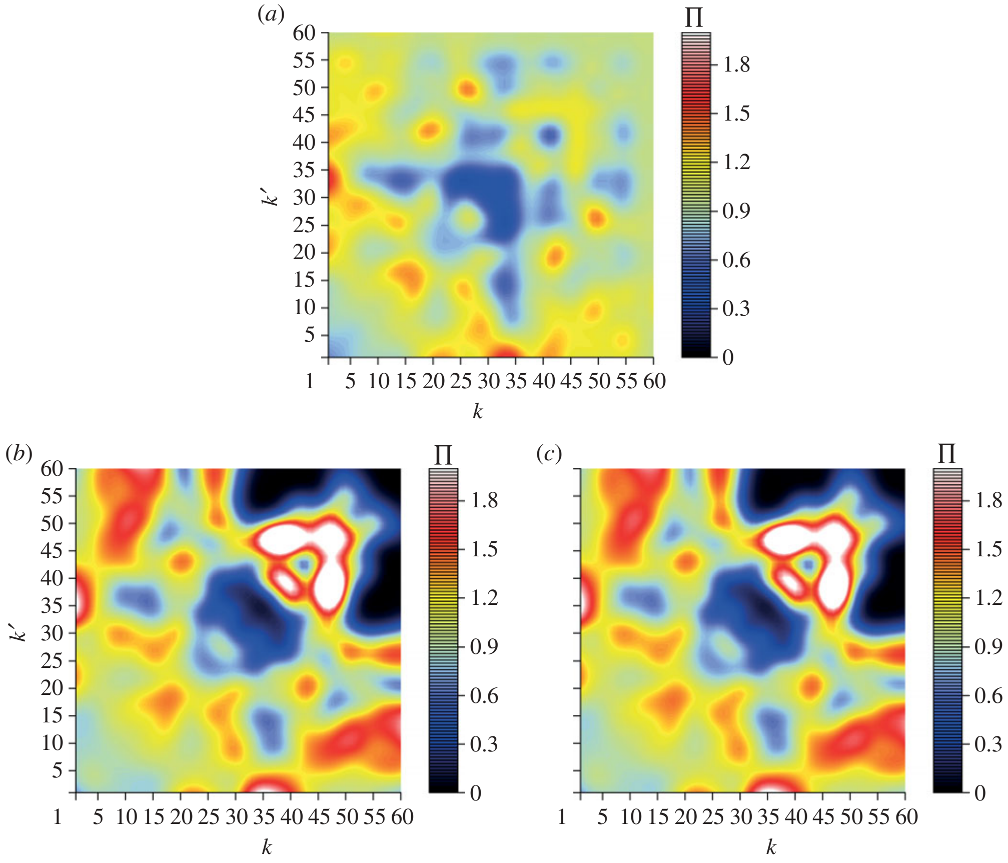

Figure 8. Normalized degree correlation function $\Pi(k,k')$ of biological protein interaction network of *C. elegans* ($N=3512$ and $\bar{k}=3.72$) (*a*) before and (*b*) after adding a fraction z/N of bonds, with $z=1$, and (*c*) its theoretical prediction. Data obtained by averaging over 10^4 samples.

where we have used $W(k,k'|1,1,z)=kp(k|1,1,z)/(\bar{k}+z)$, in accordance with identity (3.1) and equation (3.4). For $z=0$, $\mathcal{J}(k|q)=\delta_{k,q+1}$ and $W(k|0)=S_0(k)=W(k)$, so the second term in equation (4.15) vanishes; however, for $z\neq 0$, this will be generally different from zero: *crucially*, but not unexpectedly, oversampling from a graph without degree correlations automatically introduces degree correlations. Numerical results and theoretical predictions for random bond oversampling are shown in figures 8 and 9 for the biological PPIN of *C. Elegans* and synthetically generated Poissonian and power-law counterparts, respectively.

4.2. Degree correlations for connectivity dependent sampling

Let us now work out equation (4.1) for the types of biased sampling considered above.

— *Connectivity-dependent node undersampling*, i.e. $x(k)=\alpha k/k_{\max}$, $y(k,k')=1$, $z(k,k')=0$

Here, we have

$$b(q)=\frac{\alpha\bar{k}}{k_{\max}\,qp(q)}\sum_{q'}W(q,q')q',\quad a(q)=0 \quad (4.16)$$

and

$$\mathcal{L}(k|q)=\binom{q-1}{k-1}\left(\frac{\alpha\bar{k}}{k_{\max}\,qp(q)}\sum_{q''}q''\,W(q,q'')\right)^{k-1}\times\left(1-\frac{\alpha\bar{k}}{k_{\max}\,qp(q)}\sum_{q''}q''\,W(q,q'')\right)^{q-k}, \quad (4.17)$$

so our equations reduce to

$$W(k,k'|x)=\frac{\alpha\bar{k}^2}{k_{\max}\overline{k^{(3)}}}\sum_{qq'}W(q,q')qq'\binom{q-1}{k-1}\left(\frac{\alpha\bar{k}}{k_{\max}\,qp(q)}\sum_{q''}q''\,W(q,q'')\right)^{k-1}\left(1-\frac{\alpha\bar{k}}{k_{\max}\,qp(q)}\sum_{q''}q''\,W(q,q'')\right)^{q-k}$$
$$\times\binom{q'-1}{k'-1}\left(\frac{\alpha\bar{k}}{k_{\max}\,qp(q)}\sum_{q''}q''\,W(q',q'')\right)^{k'-1}\left(1-\frac{\alpha\bar{k}}{k_{\max}\,q'p(q')}\sum_{q''}q''\,W(q',q'')\right)^{q'-k'}. \quad (4.18)$$

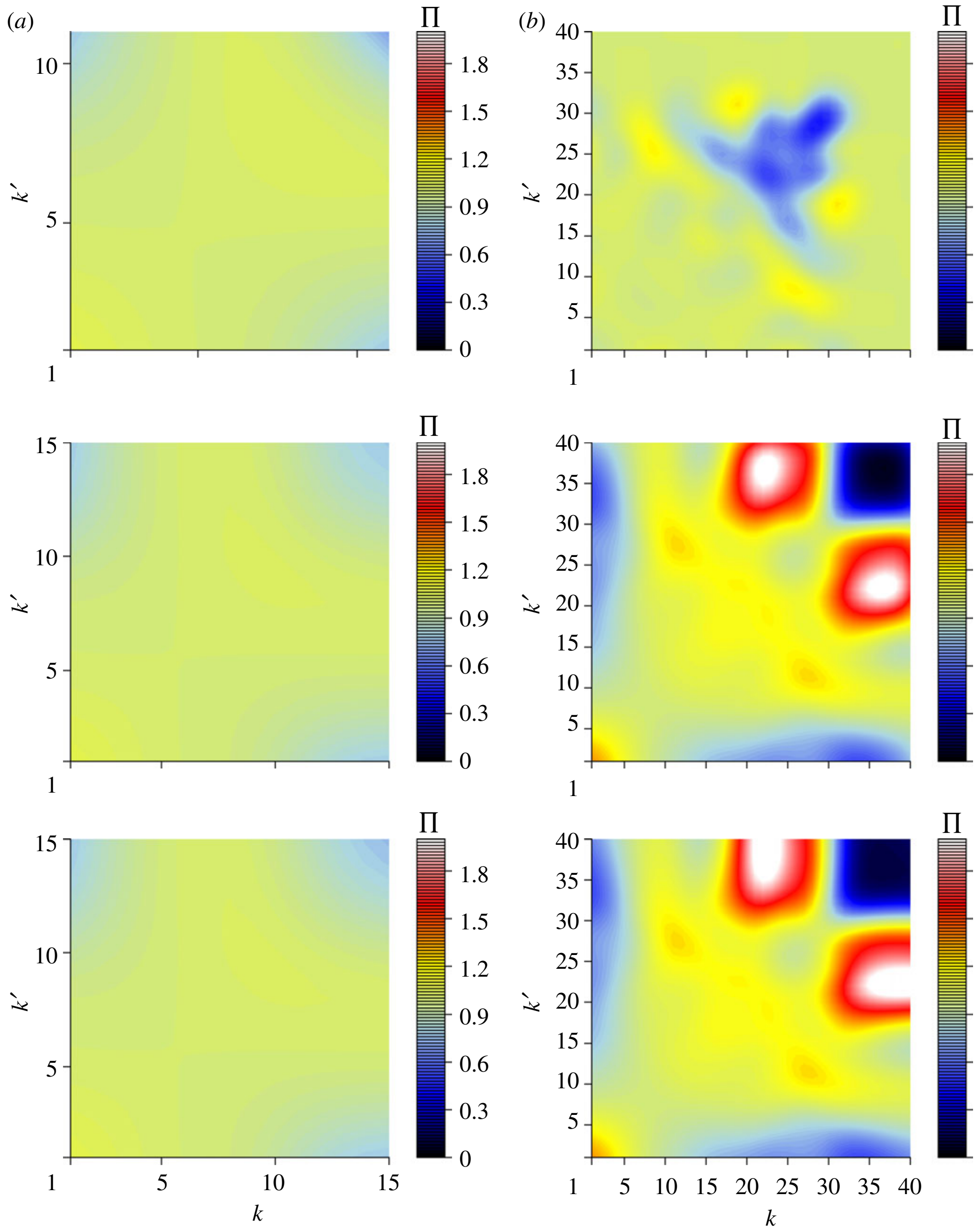

Figure 9. Normalized degree correlation function $\Pi(k,k')$ of (*a*) synthetically generated Poissonian and (*b*) power-law graphs with $N = 3512$ and $\bar{k} = 3.72$ before (top panels) and after (middle panels) adding a fraction z/N of bonds, with (*a*) $z = 1$ (left) and (*b*) $z = 2$, and their respective theoretical predictions (bottom panels). Data obtained by averaging over 10^4 samples.

— *Connectivity-dependent bond undersampling*, i.e. $x(k) = 1$, $y(k,k') = \alpha\ kk'/k^2_{\max}$, $z(k,k') = 0$

For this choice, we obtain

$$b(q) = \frac{\alpha\bar{k}}{k^2_{\max}p(q)}\sum_{q'} W(q,q')q', \quad a(q) = 0 \qquad (4.19)$$

and

$$\mathcal{L}(k|q) = \binom{q-1}{k-1}\left(\frac{\alpha\bar{k}}{k^2_{\max}p(q)}\sum_{q''} q''\, W(q,q'')\right)^{k-1} \times \left(1 - \frac{\alpha\bar{k}}{k^2_{\max}p(q)}\sum_{q''} q''\, W(q,q'')\right)^{q-k} \qquad (4.20)$$

which leads to

$$W(k,k'|x) = \frac{\alpha\,\bar{k}^2}{k_{\max}\overline{k^{(3)}}}\sum_{qq'} W(q,q')qq'\binom{q-1}{k-1}\left(\frac{\alpha\bar{k}}{k^2_{\max}p(q)}\sum_{q''} q''\, W(q,q'')\right)^{k-1}\left(1 - \frac{\alpha\bar{k}}{k^2_{\max}p(q)}\sum_{q''} q''\, W(q,q'')\right)^{q-k}$$
$$\times\binom{q'-1}{k'-1}\left(\frac{\alpha\bar{k}}{k^2_{\max}p(q)}\sum_{q''} q''\, W(q',q'')\right)^{k'-1}\left(1 - \frac{\alpha\bar{k}}{k^2_{\max}p(q')}\sum_{q''} q''\, W(q',q'')\right)^{q'-k'}. \qquad (4.21)$$

— *Connectivity-dependent bond oversampling*, i.e. $x(k)=1$, $y(k,k')=1$, $z(k,k')=\alpha kk'/k_{\max}^2$
Here, we get

$$a(q)=\frac{\alpha}{k_{\max}^2}q\bar{k},\quad b(q)=1, \tag{4.22}$$

$$\mathcal{J}(k|q)=\frac{(\alpha\bar{k}q/k_{\max}^2)^{k-q}}{(k-q)!}\mathrm{e}^{-\alpha\bar{k}q/k_{\max}^2}I(k\geq q) \tag{4.23}$$

and

$$\mathcal{L}(k|q)=\frac{(\alpha\bar{k}q/k_{\max}^2)^{k-1-q}}{(k-1-q)!}\mathrm{e}^{-\alpha\bar{k}q/k_{\max}^2}\times I(k\geq q+1). \tag{4.24}$$

Hence, we obtain

$$W(k,k'|x)=\frac{1}{\bar{k}+\alpha\bar{k}^2/k_{\max}^2}\sum_{qq'}\mathrm{e}^{-(\alpha\bar{k}q/k_{\max}^2)-(\alpha\bar{k}q'/k_{\max}^2)}\times\left[\frac{\bar{k}W(qq')}{(k-q)!(k'-q')!}\left(\frac{\alpha\bar{k}q}{k_{\max}^2}\right)^{k-q}\times\left(\frac{\alpha\bar{k}q'}{k_{\max}^2}\right)^{k'-q'}+\frac{\alpha}{k_{\max}^2}\frac{p(q)p(q')qq'}{(k-q-1)!(k'-q'-1)!}\times\left(\frac{\alpha\bar{k}q}{k_{\max}^2}\right)^{k-q}\left(\frac{\alpha\bar{k}q'}{k_{\max}^2}\right)^{k'-q'}\right]. \tag{4.25}$$

Numerical results and theoretical predictions for connectivity-dependent bond oversampling are shown in figure 10 for synthetically generated Poissonian and power-law graphs.

4.3. Summary

As was the case for the degree distribution, also the degree correlations can for a broad class of sampling protocols be calculated exactly and in terms of fully explicit relations. In contrast to the degree distribution, for which the sampling problem had already been studied partly by other authors, we are not aware of any analytical results for degree correlations. Our equations revealed that:

— Sampling will always affect the degree correlations of networks, even in the random (i.e. connectivity independent) case, if the original networks had such degree correlations.

— Uncorrelated networks will remain uncorrelated after sampling only for random node and/or bond undersampling. Bond oversampling will in general introduce degree correlations, even in the connectivity independent case.

— Random node and bond undersampling both modify the degree correlations (and the degree distribution) in the same way, so they are equivalent for any graph with prescribed topological features $p(k)$ and $W(k,k')$, as generated from ensemble (2.4).

— Node and bond undersampling cannot be mapped onto each other in the case of connectivity-dependent sampling; their effects are qualitatively different.

5. DISCUSSION

It is well known that the presently available data on cellular signalling networks are incomplete, and often suffer from serious experimental bias, reflecting the highly non-trivial nature of the experimental methods available for their collection. Yet, a significant number of research papers continue to be written in which such data are used to infer statements on the possible biological relevance of local network modules or motifs. In addition, the signalling network data are increasingly used for preprocessing gene expression data in order to derive more robust disease-specific prognostic signatures [14–16], and will very soon impact on actual treatment decisions in medicine (e.g. will be used to suggest which cancer patients are likely to benefit from which chemotherapy). Given this situation, it is vital that we understand quantitatively the data imperfections, i.e. the relation between the true biological signalling networks probed and the imperfect network samples of these networks that are reported in public data repositories and presently used by biomedical scientists. To do this, we need mathematical tools; the relevant networks are too large to rely on numerical simulation alone. Moreover, unlike simulations, analytical results can be used in reverse to infer the most probable true networks from the imperfect observed samples.

Ensembles of tailored random graphs with controlled topological properties are a natural and rigorous language for describing biological networks. They suggest precise definitions of structural features, they allow us to classify networks and obtain precise (dis)-similarity measures, they provide precise ‘null models’ for hypothesis testing, and they can serve as efficient proxies for real networks in process modelling. In this paper, we have shown how they can also be used to study analytically the effects of sampling on macroscopic topological properties of large biological networks, under a much wider range of conditions than those considered in previous analytical studies (the latter are recovered as special simple cases). We have obtained explicit expressions for both degree distributions and degree correlation kernels of sampled networks, and have been able to do this for sampling protocols that involve node and/or link undersampling as well as for link oversampling. Our predictions are in excellent agreement with numerical simulations.

As could have been expected, the most dangerous types of sampling are the connectivity-dependent ones, where the probability to observe bonds or links depends on the degrees of the nodes concerned. Unfortunately, present experimental protocols are quite likely to involve precisely such sampling. We therefore hope that our new analytical tools, which take the form of explicit and transparent equations that connect the topological structure functions $p(k)$ and $W(k,k')$ of the sampled and the true networks, can prove useful in explaining and decontaminating signalling network data.

The authors are grateful to L. Fernandes, F. Fraternali, J. Kleinjung and E.S. Roberts for stimulating discussions. A.C.C.C. would like to thank the Engineering and Physical Sciences Research Council (UK) and the Biotechnology and Biological Sciences Research Council (UK) for support.

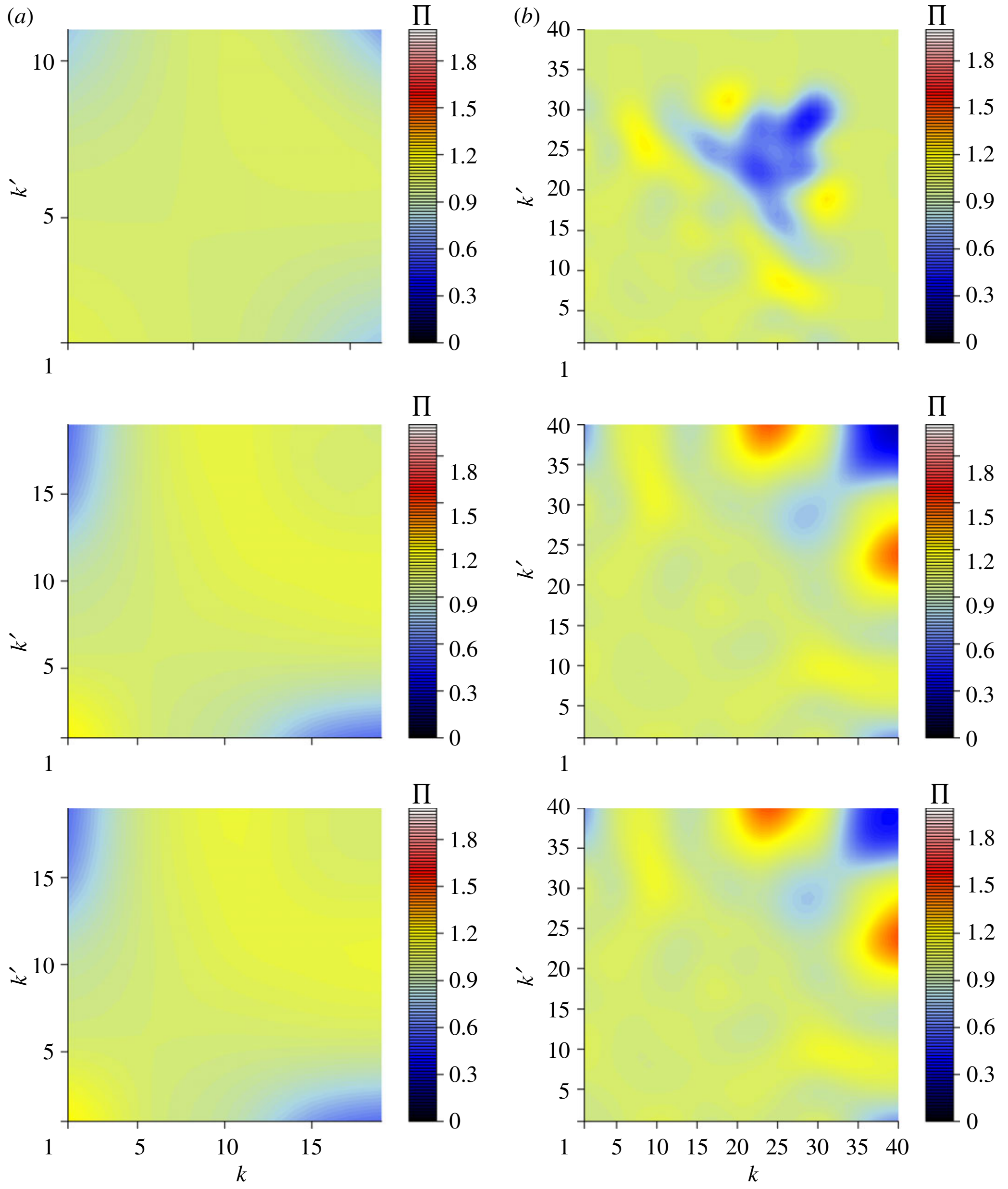

Figure 10. Effects of connectivity-dependent bond oversampling on the normalized degree correlation $\Pi(k,k')$ of the synthetically generated Poissonian and power-law graphs ($N = 3512$, $\bar{k} = 3.72$) shown in the panels (a,b), respectively. Middle panels show the result of simulations for (a) $\alpha\,\bar{k}^2/k^2_{\max} = 1$ and (b) $\alpha\,\bar{k}^2/k^2_{\max} = 0.7$ (right) and bottom panels show the corresponding theoretical predictions. Numerical data result from averaging over 10^4 samples.

APPENDIX A. JOINT DEGREE DISTRIBUTION OF CONNECTED NODES

A.1. Path integral representation of $W(k,k'|x,y,z)$

Here, we calculate the joint degree distribution of connected nodes (2.12) that will be observed in large networks that are sampled, according to protocol (2.2), from typical graphs with prescribed macroscopic topological features $p(k)$ and $W(k,k')$, as generated from ensemble (2.4). With the short-hands $\tilde{W}(k,k'|x,y,z) = W(k,k'|x,y,z)\bar{k}(x,y,z)\sum_q p(q)x(q)$, $k = (k_1,\ldots,k_N)$, and $\boldsymbol{\Omega} = (\Omega_1,\ldots,\Omega_N) \in [-\boldsymbol{\pi},\boldsymbol{\pi}]^N$ we may write

$$\tilde{W}(k,k'|x,y,z) = \lim_{N\to\infty}\sum_{\mathbf{c}} p(\mathbf{c})\left\langle \frac{1}{N}\sum_{ij} c'_{ij}\delta_{k,\sum_\ell c'_{i\ell}}\delta_{k',\sum_\ell c'_{j\ell}}\right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} = \int_{\boldsymbol{\pi}}^{\pi}\frac{\mathrm{d}\omega\mathrm{d}\omega'}{4\boldsymbol{\pi}^2}\mathrm{e}^{\mathrm{i}(\omega\mathrm{k}+\omega'\mathrm{k}')}\lim_{N\to\infty}\frac{1}{Z_{\mathrm{N}}}\int\frac{\mathrm{d}\boldsymbol{\Omega}}{(2\boldsymbol{\pi})^N}\mathrm{e}^{\mathrm{i}\boldsymbol{\Omega}\cdot\mathbf{k}}\sum_{\mathbf{c}}\prod_{r<s}$$

$$\left[\delta_{c_{rs},1}\frac{\bar{k}}{N}\frac{W(k_r,k_s)}{p(k_r)p(k_s)}\mathrm{e}^{-i(\Omega_r+\Omega_s)} + \delta_{c_{rs},0}\left(1-\frac{\bar{k}}{N}\frac{W(k_r,k_s)}{p(k_r)p(k_s)}\right)\right]\frac{1}{N}\sum_{ij}\left\langle c'_{ij}\mathrm{e}^{-i(\omega+\omega')-i\omega\sum_{\ell\neq i,j}c'_{i\ell}-i\omega'\sum_{\ell\neq i,j}c'_{j\ell}}\right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}}$$

$$
\begin{aligned}
&= \int_{\pi}^{\pi} \frac{\mathrm{d}\omega \mathrm{d}\omega'}{4\pi^2} \mathrm{e}^{\mathrm{i}(\omega(\mathrm{k}-1)+\omega'(\mathrm{k}'-1))} \lim_{N\to\infty} \frac{1}{Z_\mathrm{N}} \int \frac{\mathrm{d}\boldsymbol{\Omega}}{(2\pi)^N} \mathrm{e}^{\mathrm{i}\boldsymbol{\Omega}\cdot\mathbf{k}} \frac{1}{N} \sum_{ij} x(k_i)x(k_j) \\
&\times \sum_{\mathbf{c}} \prod_{r<s} \left[\delta_{c_{rs},1} \frac{\overline{k}}{N} \frac{W(k_r,k_s)}{p(k_r)p(k_s)} \mathrm{e}^{-i(\Omega_r+\Omega_s)} + \delta_{c_{rs},0} \left(1 - \frac{\overline{k}}{N} \frac{W(k_r,k_s)}{p(k_r)p(k_s)} \right) \right] \\
&\times \left\langle [\tau_{ij}c_{ij} + (1-c_{ij})\lambda_{ij}] \mathrm{e}^{-i\omega \sum_{\ell\neq i,j} \sigma_\ell(\tau_{i\ell}c_{i\ell}+\lambda_{i\ell}(1-c_{i\ell})) - i\omega' \sum_{\ell \neq i,j} \sigma_\ell(\tau_{j\ell}c_{j\ell}+\lambda_{j\ell}(1-c_{j\ell}))} \right\rangle_{\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{\lambda}} \qquad \text{(A1)} \\
&= \int_{\pi}^{\pi} \frac{\mathrm{d}\omega \mathrm{d}\omega'}{4\pi^2} \mathrm{e}^{\mathrm{i}(\omega(\mathrm{k}-1)+\omega'(\mathrm{k}'-1))} \lim_{N\to\infty} \frac{1}{Z_\mathrm{N}} \mathrm{e}^{\sum_{r<s} \log[1+(\overline{k}/N)(W(k_r,k_s)/(p(k_r)p(k_s)))(\mathrm{e}^{-i(\Omega_r+\Omega_s)}-1)]} \int \frac{\mathrm{d}\boldsymbol{\Omega}}{(2\pi)^N} \mathrm{e}^{\mathrm{i}\boldsymbol{\Omega}\cdot\mathbf{k}} \frac{1}{N^2} \sum_{ij} x(k_i)x(k_j) \\
&\times \left[z(k_i,k_j) + \overline{k} y(k_i,k_j) \frac{W(k_i,k_j)}{p(k_i)p(k_j)} \mathrm{e}^{-i(\Omega_i+\Omega_j)} \right] \mathrm{e}^{(\mathrm{e}^{-i\omega}-1)(1/N)\sum_\ell x(k_\ell)[z(k_i,k_\ell)+\overline{k}y(k_i,k_\ell)(W(k_i,k_\ell)/(p(k_i)p(k_\ell)))\mathrm{e}^{-i(\Omega_i+\Omega_\ell)}]} \\
&\times \mathrm{e}^{(\mathrm{e}^{-i\omega'}-1)(1/N)\sum_\ell x(k_\ell)[z(k_j,k_\ell)+\overline{k}y(k_j,k_\ell)(W(k_j,k_\ell)/(p(k_j)p(k_\ell)))\mathrm{e}^{-i(\Omega_j+\Omega_\ell)}]}.
\end{aligned}
$$

We next introduce the following order parameters

$$
P(q,\Omega|\mathbf{k},\boldsymbol{\Omega}) = \frac{1}{N} \sum_r \delta_{q,k_r} \delta(\Omega - \Omega_r) \qquad \text{(A 2)}
$$

and insert into equation (A 1) for each (q,Ω) the following integral:

$$
1 = \int \mathrm{d}P(q,\Omega) \delta[P(q,\Omega) - P(q,\Omega|\mathbf{k},\boldsymbol{\Omega})] = \left(\frac{N}{2\pi}\right) \int \mathrm{d}P(q,\Omega) \mathrm{d}\hat{P}(q,\Omega) \mathrm{e}^{\mathrm{i}N\hat{P}(q,\Omega)P(q,\Omega) - \mathrm{i}\sum_r \delta_{q,k_r}\delta(\Omega-\Omega_\mathrm{r})}. \qquad \text{(A3)}
$$

This converts equation (A 1) into the following path integral, with the short-hand $\{\mathrm{d}P\mathrm{d}\hat{P}\} = \prod_{q,\Omega} [dP(q,\Omega)\mathrm{d}\hat{P}(q,\Omega)/2\pi]$ and with $Z_N{}'$ a new constant that apart from containing Z_N absorbs various factors N and constants that are generated when transforming sums over Ω into integrals:

$$
\begin{aligned}
\tilde{W}(k,k'|x,y,z) = \lim_{N\to\infty} \int \frac{\{\mathrm{d}P\mathrm{d}\hat{P}\}}{Z_\mathrm{N}'} \mathrm{e}^{N\Psi[P,\hat{P}]+\Phi[P,\hat{P}]+\mathcal{O}(N^{-1})} \sum_{q,q'} \int \mathrm{d}\Omega \mathrm{d}\Omega' P(q,\Omega)P(q',\Omega')x(q)x(q') \\
\times\, \Xi(q,\Omega;q',\Omega') \left(\int_{\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega(k-1)+(\mathrm{e}^{-\mathrm{i}\omega}-1)Q(q,\Omega)} \right) \left(\int_{\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega(k'-1)+(\mathrm{e}^{-\mathrm{i}\omega}-1)Q(q',\Omega')} \right)
\end{aligned}
\qquad \text{(A 4)}
$$

in which $\Phi[P,\hat{P}]$ will eventually drop out of our formulae (via normalization) and

$$
\begin{aligned}
\Psi[P,\hat{P}] = i\sum_q \int \mathrm{d}\Omega \hat{P}(q,\Omega)P(q,\Omega) + \sum_k p(k) \log \int \frac{\mathrm{d}\Omega}{2\pi} \mathrm{e}^{\mathrm{i}\Omega k - \mathrm{i}\hat{P}(k,\Omega)} \\
+ \frac{1}{2}\overline{k} \sum_{qq'} \int \mathrm{d}\Omega \mathrm{d}\Omega' P(q,\Omega)P(q',\Omega') \frac{W(q,q')}{p(q)p(q')} (\mathrm{e}^{-i(\Omega+\Omega')} - 1)
\end{aligned}
\qquad \text{(A 5)}
$$

$$
\Xi(q,\omega;q',\omega') = z(q,q') + y(q,q')\overline{k} \frac{W(q,q')}{p(q)p(q')} \mathrm{e}^{-i(\omega+\omega')} \qquad \text{(A 6)}
$$

and

$$
Q(q,\Omega) = \sum_{q''} \int \mathrm{d}\Omega'' P(q'',\Omega'')x(q'')\Xi(q,\Omega;q'',\Omega''). \qquad \text{(A 7)}
$$

The relevant ω-integrals are of the familiar form

$$
\int_{\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega\ell+(\mathrm{e}^{-\mathrm{i}\omega}-1)Q} = \mathrm{e}^{-Q} \sum_{n\geq 0} \frac{Q^n}{n!} \int_{\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega(\ell-\mathrm{n})} = \frac{\mathrm{e}^{-Q}Q^\ell}{\ell!} \qquad \text{(A 8)}
$$

(unless $\ell < 0$, in which case the integral is zero). We also note that by definition we always have the normalization identity $\sum_{k,k'\geq 0} W(k,k'|x,y,z) = 1$. So we arrive at:

$$W(k,k'|x,y,z) = \bar{\delta}_{k,0}\,\bar{\delta}_{k',0}\;\frac{\sum_{q,q'} x(q)x(q')\int \mathrm{d}\Omega\mathrm{d}\Omega' P(q,\Omega)P(q',\Omega')\Xi(q,\Omega;q',\Omega')I(k|q,\Omega)I(k'|q',\Omega')}{\sum_{q,q'} x(q)x(q')\int \mathrm{d}\Omega\mathrm{d}\Omega' P(q,\Omega)P(q',\Omega')\Xi(q,\Omega;q',\Omega')} \tag{A 9}$$

with

$$I(k|q,\Omega) = \frac{\mathrm{e}^{-Q(q,\Omega)}\,Q^{k-1}(q,\Omega)}{(k-1)!}, \tag{A 10}$$

and in which, via the steepest descent argument, the order parameters $\{P,\hat{P}\}$ are the functions that extremize the kernel (A 5).

A.2. Functional saddle-point equations

Functional variation of equation (A 5) gives the following saddle-point equations for $\{P,\hat{P}\}$:

$$\mathrm{i}\hat{P}(q,\Omega) = -\bar{k}\sum_{q'}\int \mathrm{d}\Omega' P(q',\Omega')\frac{W(q,q')}{p(q)p(q')}(\mathrm{e}^{-i(\Omega+\Omega')}-1) \tag{A 11}$$

and

$$P(q,\Omega) = p(q)\frac{\mathrm{e}^{\mathrm{i}\Omega q-\mathrm{i}\hat{P}(q,\Omega)}}{\int \mathrm{d}\Omega' \mathrm{e}^{\mathrm{i}\Omega' q-\mathrm{i}\hat{P}(q,\Omega')}}. \tag{A 12}$$

Equivalently:

$$\mathrm{i}\hat{P}(q,\Omega) = \bar{k}\lambda(q) - \bar{k}\mathrm{e}^{-i\Omega}\phi(q), \quad P(q,\Omega) = p(q)\frac{\mathrm{e}^{\mathrm{i}\Omega q+\bar{k}\mathrm{e}^{-\mathrm{i}\Omega}\phi(q)}}{\int \mathrm{d}\Omega' \mathrm{e}^{\mathrm{i}\Omega' q+\bar{k}\mathrm{e}^{-\mathrm{i}\Omega'}\phi(q)}} \tag{A 13}$$

with

$$\phi(q) = \sum_{q'}\frac{W(q,q')}{p(q)p(q')}\int \mathrm{d}\Omega P(q',\Omega)\mathrm{e}^{-i\Omega} \tag{A 14}$$

and

$$\lambda(q) = \sum_{q'}\frac{W(q,q')}{p(q)p(q')}\int \mathrm{d}\Omega P(q',\Omega). \tag{A 15}$$

The integrals over Ω in equations (A 14) and (A 15) are again of the type (A 8), from which we derive $\int \mathrm{d}\Omega P(q,\Omega) = p(q)$ and $\int \mathrm{d}\Omega P(q,\Omega)\mathrm{e}^{-i\Omega} = p(q)q/\bar{k}\phi(q)$. This then converts equations (A 14) and (A 15) into

$$\phi(q)p(q) = \frac{\sum_{q'} W(q,q')q'}{\bar{k}\phi(q')} \quad \text{and} \quad \lambda(q) = \frac{1}{p(q)}\sum_{q'} W(q,q'). \tag{A 16}$$

Since we know the marginal of the distribution $W(k,k')$ to be $\sum_{k'} W(k,k') = kp(k)/\bar{k}$ (which follows directly from its definition), we can immediately read off the solution of equation (A 16):

$$\phi(q) = \lambda(q) = \frac{q}{\bar{k}}. \tag{A 17}$$

Insertion into equation (A 13) and using equation (A 8) gives the solution of equations (A 11) and (A 12) in explicit form:

$$\mathrm{i}\hat{P}(q,\Omega) = q - q,\mathrm{e}^{-i\Omega} \quad \text{and} \quad P(q,\Omega) = p(q)\frac{\mathrm{e}^{\mathrm{i}\Omega q+q\,\mathrm{e}^{-\mathrm{i}\Omega}}}{2\pi q^q/q!}. \tag{A 18}$$

A.3. Final result for the distribution W(k,k′/x,y,z)

We can now evaluate the various ingredients of equation (A 9). The function $Q(q,\Omega)$ becomes

$$Q(q,\Omega) = \sum_{q'\geq 0} p(q')x(q')z(q,q') + \bar{k}\mathrm{e}^{-i\Omega}\frac{1}{p(q)}\sum_{q'\geq 0} x(q')y(q,q')\,W(q,q'). \tag{A 19}$$

Hence

$$
\begin{aligned}
&\int \mathrm{d}\Omega \mathrm{d}\Omega' P(q,\Omega)P(q',\Omega')\Xi(q,\Omega;q',\Omega')I(k|q,\Omega)I(k'|q',\Omega')\\
&\quad = \frac{p(q)\,q!}{q^q(k-1)!}\int \frac{\mathrm{d}\Omega}{2\pi}\mathrm{e}^{\mathrm{i}\Omega q + q\mathrm{e}^{-\mathrm{i}\Omega} - Q(q,\Omega)}\,Q^{k-1}(q,\Omega)\\
&\qquad \times \frac{p(q')(q')!}{(q')^{q'}(k'-1)!}\int \frac{\mathrm{d}\Omega'}{2\pi}\mathrm{e}^{\mathrm{i}\Omega' q' + q'\mathrm{e}^{-\mathrm{i}\Omega'} - Q(q',\Omega')}\,Q^{k'-1}(q',\Omega')\\
&\qquad \times \left[z(q,q') + y(q,q')\overline{k}\frac{W(q,q')}{p(q)p(q')}\mathrm{e}^{-i(\Omega+\Omega')}\right]\\
&\quad = p(q)p(q')z(q,q')\mathcal{J}(k|q)\mathcal{J}(k'|q') + \overline{k}\,W(q,q')y(q,q')\mathcal{L}(k|q)\mathcal{L}(k'|q')
\end{aligned}
\tag{A 20}
$$

in which

$$
\mathcal{J}(k|q) = \overline{\delta}_{k,0}\frac{q!}{q^q(k-1)!}\int_{-\pi}^{\pi}\frac{\mathrm{d}\Omega}{2\pi}\mathrm{e}^{\mathrm{i}\Omega q + q\mathrm{e}^{-\mathrm{i}\Omega}}\,Q^{k-1}(q,\Omega)\mathrm{e}^{-Q(q,\Omega)}
\tag{A 21}
$$

and

$$
\mathcal{L}(k|q) = \overline{\delta}_{k,0}\frac{q!}{q^q(k-1)!}\int_{-\pi}^{\pi}\frac{\mathrm{d}\Omega}{2\pi}\mathrm{e}^{\mathrm{i}\Omega(q-1) + q\mathrm{e}^{-\mathrm{i}\Omega}}\,Q^{k-1}(q,\Omega)\mathrm{e}^{-Q(q,\Omega)}.
\tag{A 22}
$$

Summation over k reveals that $\sum_{k\geq 0}\mathcal{J}(k|q) = \sum_{k\geq 0}\mathcal{L}(k|q) = 1$ for all $q > 0$, which leads to the final result:

$$
\begin{aligned}
&W(k,k'|x,y,z)\\
&= \frac{\sum_{q,q'>0} x(q)x(q')\{p(q)p(q')z(q,q')\mathcal{J}(k|q)\mathcal{J}(k'|q') + \overline{k}\,W(q,q')y(q,q')\mathcal{L}(k|q)\mathcal{L}(k'|q')\}}{\sum_{q,q'>0}x(q)x(q')\{p(q)p(q')z(q,q') + \overline{k}\,W(q,q')y(q,q')\}}\\
&= \frac{\sum_{q,q'>0} x(q)x(q')\{p(q)p(q')z(q,q')\mathcal{J}(k|q)\mathcal{J}(k'|q') + \overline{k}\,W(q,q')y(q,q')\mathcal{L}(k|q)\mathcal{L}(k'|q')\}}{\overline{k}(x,y,z)\sum_q p(q)x(q)}.
\end{aligned}
\tag{A 23}
$$

with $\overline{k}(x,y,z)$ as given in equation (2.10). The marginals of $W(k,k'|x,y,z)$ are obtained trivially by summing equation (A 23) over k', giving

$$
W(k|x,y,z) = \frac{\sum_{q,q'>0} x(q)x(q')\{p(q)p(q')z(q,q')\mathcal{J}(k|q) + \overline{k}\,W(q,q')y(q,q')\mathcal{L}(k|q)\}}{\overline{k}(x,y,z)\sum_q p(q)x(q)}.
\tag{A 24}
$$

A.4. Explicit expression for the factors $\mathcal{J}(k|q)$

To carry out the integral in equations (A 21) and (A 22), we first write $Q(q,\Omega)$ as $Q(q,\Omega) = a(q) + b(q)q\mathrm{e}^{-\mathrm{i}\Omega}$, with

$$
a(q) = \sum_{q'\geq 0} p(q')x(q')z(q,q') \quad \text{and } b(q) = \frac{\overline{k}}{qp(q)}\sum_{q'\geq 0} x(q')y(q,q')\,W(q,q').
\tag{A 25}
$$

We note that, owing to $\sum_{q'} W(q,q') = (q/\overline{k})p(q)$, we can be sure that $a(q) \in [0,1]$ and $b(q) \in [0,1]$. Substitution into equations (A 21) and (A 22) and integration over Ω, for $q > 0$ and $k > 0$, then leads to

$$
\begin{aligned}
&\mathcal{J}(k|q) = \mathrm{e}^{-a(q)}\frac{q!}{q^q(k-1)!}\sum_{n=0}^{k-1}\binom{k-1}{n}a^{k-1-n}(q)b^n(q)q^n\\
&\int_{-\pi}^{\pi}\frac{\mathrm{d}\Omega}{2\pi}\mathrm{e}^{\mathrm{i}\Omega(q-n) + q(1-b(q))\mathrm{e}^{-\mathrm{i}\Omega}} = \mathrm{e}^{-a(q)}\sum_{n=0}^{\min\{k-1,q\}}\binom{q}{n}\frac{a^{k-1-n}(q)}{(k-1-n)!}b^n(q)(1-b(q))^{q-n}
\end{aligned}
\tag{A 26}
$$

and, similarly,

$$
\mathcal{L}(k|q) = \mathrm{e}^{-a(q)}\sum_{n=0}^{\min\{k-1,q-1\}}\binom{q-1}{n}\frac{a^{k-1-n}(q)}{(k-1-n)!}b^n(q)(1-b(q))^{q-1-n}.
\tag{A 27}
$$

Clearly, $\mathcal{J}(k|q) \geq 0, \mathcal{L}(k|q) \geq 0$ for all (k,q). Since the factors (A 21) and (A 22) also satisfy the normalization $\sum_{k\geq 0} \mathcal{J}_{kq} = 1, \sum_{k\geq 0} \mathcal{L}_{kq} = 1$ for all $q > 0$, they can be interpreted as conditional probabilities, as suggested by our chosen notation.

A.5. Tests

To test our expression (A 23) we set $x(k) = x$, $y(k) = y$ and $z(k,k') = z$, and try to recover from equation (A 24) via identity (3.1), our earlier results on the degree distribution for unbiased sampling. We now find $a(q) = xz$, $b(q) = xy$ and $\bar{k}(x, y, z) = x(z + \bar{k}y)$, which implies that

$$\mathcal{J}(k|q) = \mathrm{e}^{-xz} x^{k-1} \sum_{n=0}^{\min\{q,k-1\}} \binom{q}{n} y^n (1 - xy)^{q-n} \frac{z^{k-1-n}}{(k-1-n)!} \tag{A 28}$$

and

$$\mathcal{L}(k|q) = \mathcal{J}(k|q-1). \tag{A 29}$$

Let us inspect the following cases:

— *Perfect sampling*, i.e. $x = y = 1$ and $z = 0$.
Now there should be no difference between the kernel $W(k,k')$ and the observed kernel $W(k,k'|1,1,0)$ of the sample. Here, we see that equation (A 19) simplifies to $Q(q,\Omega) = q\mathrm{e}^{-\mathrm{i}\,\Omega}$; hence $a(q) = 0$ and $b(q) = 1$ leading to $\mathcal{L}(k|q) = \delta_{q,k}$ and therefore to the correct identity $W(k,k'|x,y,z) = W(k,k')$.

— *Unbiased node and/or link undersampling*, i.e. $xy < 1$ and $z = 0$.
Now we have $\bar{k}(x, y, 0) = \bar{k}xy$ and

$$\mathcal{J}(k|q) = x^{k-1} \binom{q}{k-1} y^{k-1} (1 - xy)^{q-k+1} I(q \geq k - 1), \tag{A 30}$$

which gives

$$W(k|x, y, 0) = \frac{1}{\bar{k}} \sum_{k'\geq k} p(k')k' \binom{k'-1}{k-1} (xy)^{k-1} (1 - xy)^{k'-k} \tag{A 31}$$

and therefore we recover the correct expression

$$\begin{aligned} p(k|x, y, 0) &= \frac{\bar{k}xy}{k} W(k|x, y, 0) \\ &= (xy)^k \sum_{k'\geq k} p(k') \binom{k'}{k'-k} (1 - xy)^{k'-k}. \end{aligned} \tag{A 32}$$

— *Unbiased bond oversampling*, i.e. $x = y = 1$ and $z > 0$.
Now $\bar{k}(1, 1, z) = \bar{k} + z$ and $\mathcal{J}(k|q) = \mathrm{e}^{-z} \frac{z^{k-1-q}}{(k-1-q)!} I(k \geq q + 1)$, which results in

$$\begin{aligned} p(k|1, 1, z) &= \frac{\bar{k}}{k} W(k|1, 1, z) = \frac{1}{k} \left\{ z \sum_q p(q)\mathcal{J}(k|q) + \sum_q p(q)q\mathcal{J}(k|q - 1) \right\} \\ &= \frac{\mathrm{e}^{-z}}{k} \left\{ \sum_{q=0}^{k-1} p(q) \frac{z^{k-q}}{(k-1-q)!} + \sum_{q=1}^{k} p(q)q \frac{z^{k-q}}{(k-q)!} \right\} = \mathrm{e}^{-z} \sum_{\ell=0}^{k} \frac{p(k-\ell)z^{\ell}}{\ell!}, \end{aligned} \tag{A 33}$$

which is indeed the correct result identified earlier.

REFERENCES

1 Prasad, T. S. K. *et al.* 2009 Human protein reference database—2009 update. *Nucleic Acids Res.* **37**, D767–D772. (doi:10.1093/nar/gkn892)

2 Fernandes, L. P., Annibale, A., Kleinjung, J., Coolen, A. C. C. & Fraternali, F. 2010 Protein networks reveal detection bias and species consistency when analysed by information-theoretic methods. *PLoS ONE* **5**, e12083. (doi:10.1371/journal.pone.0012083)

3 Stumpf, M. P. H. & Wiuf, C. 2005 Sampling properties of random graphs: the degree distribution. *Phys. Rev. E* **72**, 036118. (doi:10.1103/PhysRevE.72.036118)

4 Stumpf, M. P. H., Wiuf, C. & May, R. M. 2005 Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA* **1.2**, 4221–4224. (doi:10.1073/pnas.0501179102)

5 Han, J. D. J., Dupuy, D., Bertin, N., Cusick, M. E. & Vidal, M. 2005 Effect of sampling on topology predictions

of protein–protein interaction networks. *Nat. Biotechnol.* **23**, 839–844. (doi:10.1038/nbt1116)

6 Lee, S. H., Kim, P.-J. & Jeong, H. 2006 Statistical properties of sampled networks. *Phys. Rev. E* **73**, 016102. (doi:10.1103/PhysRevE.73.016102)

7 De Silva, E., Thorne, T., Ingram, P., Agrafioti, I., Swire, J., Wiuf, C. & Stumpf, M. P. H. 2006 The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol.* **4**, 39. (doi:10.1186/1741-7007-4-39)

8 Viger, F., Barrat, A., Dall' Asta, L., Zhang, C. H. & Kolaczyk, E. D. 2007 What is the real size of a sampled network? The case of the internet. *Phys. Rev. E* **75**, 056111. (doi:10.1103/PhysRevE.75.056111)

9 Solokov, I. M. & Eliazar, I. I. 2010 Sampling from scale-free networks and the matchmaking paradox. *Phys. Rev. E* **81**, 026107. (doi:10.1103/PhysRevE.81.026107)

10 Annibale, A., Coolen, A. C. C., Fernandes, L. P., Fraternali, F. & Kleinjung, J. 2009 Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure. *J. Phys. A Math. Theor.* **42**, 485001. See http://arxiv.org/abs/0908.1759.

11 Coolen, A. C. C., De Martino, A. & Annibale, A. 2009 Constrained Markovian dynamics of random graphs. *J. Stat. Phys.* **136**, 1035–1067. (doi:10.1007/s10955-009-9821-2)

12 Coolen, A. C. C., Fraternali, F., Annibale, A., Fernandes, L. P. & Kleinjung, J. In press. Modelling biological networks via tailored random graphs. *Handb. Stat. Syst. Biol.*

13 Simonis, N. *et al.* 2009 Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. *Nat. Methods* **6**, 47–54. (doi:10.1038/nmeth.1279)

14 Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. 2007 Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 1–10. (doi:10.1038/msb4100180)

15 Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E. & Vert, J. P. 2007 Classification of microarray data using gene networks. *BMC Bioinf.* **8**, 35. (doi:10.1186/1471-2105-8-35)

16 Taylor, I. W. *et al.* 2009 Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27**, 199–204. (doi:10.1038/nbt.1522)