INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Annibale A, Coolen ACC, Planell-Morell N. 2015 Quantifying noise in mass spectrometry and yeast two-hybrid protein interaction detection experiments. *J. R. Soc. Interface* **12**: 20150573.
http://dx.doi.org/10.1098/rsif.2015.0573

Received: 27 June 2015
Accepted: 12 August 2015

Subject Areas:
systems biology, mathematical physics

Keywords:
protein interaction networks, detection bias/sampling, network ensemble, network inference

Author for correspondence:
A. Annibale
e-mail: alessia.annibale@kcl.ac.uk

Quantifying noise in mass spectrometry and yeast two-hybrid protein interaction detection experiments

A. Annibale[1], A. C. C. Coolen[1,2,3] and N. Planell-Morell[1]

[1]Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK
[2]Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, London SE1 1UL, UK
[3]London Institute for Mathematical Sciences, 22 South Audley Street, London W1K 2NY, UK

Protein interaction networks (PINs) are popular means to visualize the proteome. However, PIN datasets are known to be noisy, incomplete and biased by the experimental protocols used to detect protein interactions. This paper aims at understanding the connection between true protein interactions and the protein interaction datasets that have been obtained using the most popular experimental techniques, i.e. mass spectronomy and yeast two-hybrid. We start from the observation that the adjacency matrix of a PIN, i.e. the binary matrix which defines, for every pair of proteins in the network, whether or not there is a link, has a special form, that we call separable. This induces precise relationships between the moments of the degree distribution (i.e. the average number of links that a protein in the network has, its variance, etc.) and the number of short loops (i.e. triangles, squares, etc.) along the links of the network. These relationships provide powerful tools to test the reliability of datasets and hint at the underlying biological mechanism with which proteins and complexes recruit each other.

1. Introduction

Protein interactions are a biological phenomenon that controls a large part of the functionality of a cell. Protein interaction networks (PINs) are graphical representations of the complex patterns of interactions that appear in the proteome, which enable quantitative studies of the underlying biology via mathematical tools and complex networks theory.

Mathematically, a PIN is a graph where nodes $i = 1 \ldots N$ represent proteins and links represent their interactions. This graph is encoded in an adjacency matrix $\mathbf{a} = \{a_{ij}\}$, whose entries denote whether there is a link between proteins i and j ($a_{ij} = 1$) or not ($a_{ij} = 0$). However, there is ambiguity in its definition, arising from the non-binary nature of the underlying biochemistry. For example, three proteins may form a complex, but may not interact in pairs. Assigning binary values to intrinsically non-binary interactions requires further prescriptions, which vary across experimental protocols and lead in practice to different graphs. Moreover, different experiments measure protein interactions in different ways, which causes further biases [1–3]. For quantitative studies of the effects of sampling biases on networks see [4–10].

In this paper, we seek to establish the connection between true biological protein interactions and protein interaction datasets produced by the most popular experimental techniques, mass spectronomy (MS) and yeast two-hybrid (Y2H). We argue that the most natural network matrix representation of the proteome has a separable form, which induces precise relationships between the degree distribution and the density of short loops. These relationships provide simple tests to assess the reliability and quality of different datasets, and provide hints on the underlying (evolutionary) mechanisms with which proteins and complexes recruit each other. Our study also provides a theoretical framework to discriminate between 'party' and 'date' hubs in PINs (e.g. [11] and references therein) and addresses several intriguing questions concerning

THE ROYAL SOCIETY PUBLISHING

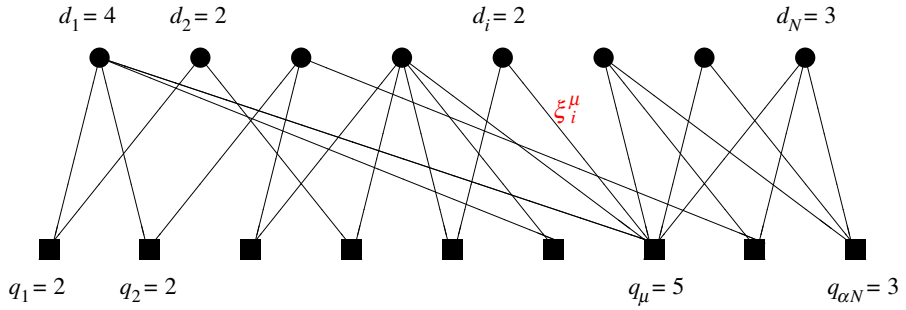

Figure 1. Bipartite graph representation of protein interactions. The protein species $i = 1 \ldots N$ are drawn as circles, and their complexes $\mu = 1 \ldots \alpha N$ as squares. In the bipartite graph representation of protein interactions, d_i is the degree of protein i or protein promiscuity (denoting the number of complexes it participates in), and q_μ is the degree of complex μ or complex size (denoting the number of protein species it contains). The bipartite graph gives more detailed information than the conventional PIN with protein nodes and pairwise links only. For instance, one distinguishes easily between different types of 'hub' proteins: 'date hub' proteins connect to many degree-2 complexes, whereas 'party hub' proteins connect to a high degree complex. (Online version in colour.)

the universality of protein and complex statistics across species. For example, given N protein species in a cell, what is the number of complexes they typically form, i.e. to what extent is the ratio α complexes/proteins conserved across different species? Is the distribution of complex sizes peaked around 'typical' values, or does it have long tails? How is this mirrored in the protein promiscuities, i.e. the propensities of proteins to participate in multiple complexes? Does the power-law behaviour of the degree distribution of PINs perhaps result from tails in the distribution of complex sizes and protein promiscuities?

We tackle the above questions using a mathematical approach that is entirely based on statistical properties of graph ensembles. In §2, we define our models as distinct separable graph ensembles which mimic PINs, each reflecting different possible mechanisms for complex genesis. In §3, we give an overview of the main results and their application to real PINs measured by MS and Y2H experiments. Section 4 defines the mathematical set-up of our analysis and in §§5–7, we give a full derivation of results, that are tested on synthetically generated networks in §8. We end our paper with a summary of our conclusions, and suggest pathways for further research.

2. Definitions and basic properties

2.1. The bipartite graph representation of the proteome

Proteins are large and complicated heteropolymers, which can bind in specific combinations to form stable molecular complexes. We consider a set of N protein species, labelled by $i = 1 \ldots N$. We assume that the number of stable complexes p scales as $p = \alpha N$, where $\alpha > 0$, and we label the complexes by $\mu = 1 \ldots \alpha N$. We can represent this system as a bipartite graph [12], with two sets of nodes (figure 1). The set ν_p represents proteins species (drawn as circles), the set ν_c represents complexes (drawn as squares), and a link between protein species $i \in \nu_p$ and complex $\mu \in \nu_c$ is drawn if protein i participates in complex μ. This graph is defined by the $N \times \alpha N$ connectivity matrix $\boldsymbol{\xi} = \{\xi_i^\mu\}$, where $\xi_i^\mu = 1$ if there is a link between i and μ, and $\xi_i^\mu = 0$ otherwise. For simplicity, we do not allow for complexes with more than one occurrence of any given protein species.

In the bipartite graph, one has two types of node degrees: the degree $d_i(\boldsymbol{\xi}) = \sum_\mu \xi_i^\mu$ (or 'promiscuity') of each protein i gives the number of different complexes in which it is involved, and the degree $q_\mu(\boldsymbol{\xi}) = \sum_i \xi_i^\mu$ (or 'size') of each complex μ gives the number of protein species of which it is formed (figure 1). For a given bipartite graph $\boldsymbol{\xi}$, we define the protein degree distribution, or promiscuity distribution, as

$$p(d|\boldsymbol{\xi}) = N^{-1} \sum_{i=1}^{N} \delta_{d,d_i(\boldsymbol{\xi})}, \tag{2.1}$$

where δ_{xy} is the Kronecher function, defined as 1 for $x = y$ and 0 otherwise. This counts the frequency of occurrence of a protein with promiscuity d, and it is normalized by the total number of proteins N. Similarly, we define the complex degree distribution, or complex size distribution, as

$$p(q|\boldsymbol{\xi}) = (\alpha N)^{-1} \sum_{\mu=1}^{\alpha N} \delta_{q,q_\mu(\boldsymbol{\xi})}, \tag{2.2}$$

which counts the frequency of occurrence of a complex containing q different protein species, divided by the total number of complexes αN.

As the number of links stemming from the proteins has to equate the number of links stemming from complexes, we have

$$\sum_{\mu=1}^{\alpha N} q_\mu(\boldsymbol{\xi}) = \sum_{i=1}^{N} d_i(\boldsymbol{\xi}) \quad \forall\ \boldsymbol{\xi}.$$

This leads to the identity

$$\langle d(\boldsymbol{\xi}) \rangle = \alpha \langle q(\boldsymbol{\xi}) \rangle,$$

where $\langle d(\boldsymbol{\xi}) \rangle = \sum_d d p(d|\boldsymbol{\xi})$ is the average promiscuity, i.e. the first moment of the distribution of promiscuities, and $\langle q(\boldsymbol{\xi}) \rangle = \sum_q q p(q|\boldsymbol{\xi})$ is the average complex size, i.e. the first moment of the distribution of complex sizes.

2.2. Protein interactions as detected by experiments

Protein detection experiments seek to measure for each pair (i, j) of protein species whether they interact in any complex, and assign an undirected link between nodes i and j if they do. This leads to a graphical representation of protein interactions in terms of a monopartite graph, where there is only one type of nodes, which represent proteins. The graph can be represented by an adjacency matrix $\boldsymbol{a} = \{a_{ij}\}$ whose entries are $a_{ij} = 1$ if there is a link between proteins i, j and 0 otherwise. It is reasonable to expect that PIN adjacency matrices resulting from detection experiments are in the form

$$a_{ii} = 0 \quad \forall\ i \tag{2.3}$$

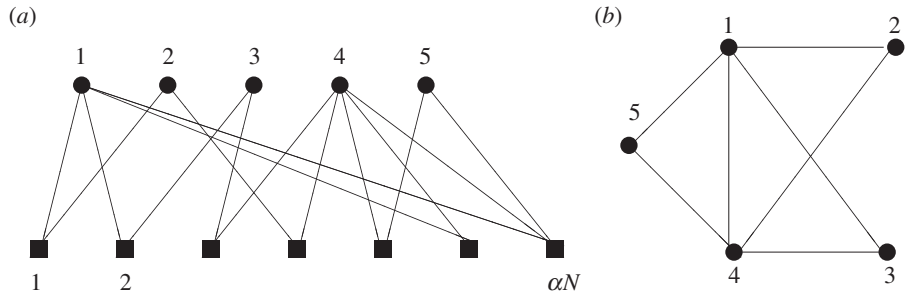

Figure 2. Bipartite graph representation $\boldsymbol{\xi}$ of protein interactions (*a*) and corresponding monopartite graph representation $\boldsymbol{a}$, obtained by marginalizing the bipartite graph (*b*). A dimer, i.e. a complex made up of two proteins *i*, *j* in the bipartite graph corresponds to a link between proteins *i*, *j* in the monopartite graph. A trimer, i.e. a complex made up of three proteins in the bipartite graph corresponds to a triangle in the monopartite graph. Similarly, larger complexes correspond to clique motifs.

and

$$a_{ij} = \theta\left(\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu\right) \quad \forall\, i \neq j, \tag{2.4}$$

where $\theta(x)$ is a binary function that takes value 1 if $x > 0$ and 0 otherwise. We call the form of the matrix $\boldsymbol{a}$ 'separable' as the dependence of its entries on the indices *i*,*j* is factorized, as a consequence of the fact that protein interactions are mediated by complexes. Graph $\boldsymbol{a}$ can be thought of as a marginalized version of the bipartite graph $\boldsymbol{\xi}$ whereby complexes are 'integrated-out', i.e. summed over. Figure 2 gives an illustration of the relationship between the two graphical representations, and shows that protein complexes in $\boldsymbol{\xi}$ are inextricably related to loops in $\boldsymbol{a}$. In particular, the presence of large complexes will boost the number of short loops in PIN $\boldsymbol{a}$.

If PINs detected experimentally displayed properties too far away from those observed in (2.4), this might signal the presence of strong biases in the experimental protocols for the PIN detection and one should ask what exactly the detection experiment is measuring. A key feature we will exploit in our analysis is that, due to the sparsity of links ξ_i^μ, average properties of random graphs (2.4) are identical, to leading orders in N, to those of the related *weighted* random graphs

$$c_{ii} = 0 \quad \forall\, i \tag{2.5}$$

and

$$c_{ij} = \sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu \quad \forall\, i \neq j, \tag{2.6}$$

for which average properties are much easier to calculate. Graphs $\boldsymbol{c}$ have the same structure as $\boldsymbol{a}$ but links are weighted, with each weight $c_{ij} = \sum_{\mu \leq \alpha N} \xi_i^\mu \xi_j^\mu \in \mathrm{IN}$ representing the *number* of complexes in which proteins *i* and *j* participate simultaneously. We will mainly focus on the following properties of the random graphs $\boldsymbol{c}$:

— The degree distribution

$$p(k|\boldsymbol{c}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{k,k_i(c)} \tag{2.7}$$

denoting the probability of observing a node in graph $\boldsymbol{c}$ having degree $k_i(\boldsymbol{c}) = \sum_j c_{ij}$ equal to k. This distribution should not be confused with the promiscuity distribution $p(d|\boldsymbol{\xi})$. The latter is a property of the bipartite graphs $\boldsymbol{\xi}$, whereas the former is a property of the monopartite graphs $\boldsymbol{c}$.

— The density of loops of length 3 and 4, defined as

$$m_3(\boldsymbol{c}) = \frac{1}{N}\sum_{ijk} c_{ij} c_{jk} c_{ki} \tag{2.8}$$

and

$$m_4(\boldsymbol{c}) = \frac{1}{N}\sum_{ijk\ell} c_{ij} c_{jk} c_{k\ell} c_{\ell i} \tag{2.9}$$

denoting, respectively, the density of closed non-intersecting paths of length 3 and 4, along the links of network $\boldsymbol{c}$.

We do not go beyond loops of length 4 because real networks are known to be 'small world', with pairs of nodes typically connected by paths of small length. With larger path lengths, one can typically link every node to any other and the number of loops through any node will increase significantly as we increase the length of loops. Hence, one expects the relevant biological information to be encoded in the short loops statistics.

2.3. Link distribution in the bipartite graph

As we generally do not know the microscopic bipartite graph $\boldsymbol{\xi}$, we will regard the $\boldsymbol{\xi}$'s as random variables, drawn from a distribution $p(\boldsymbol{\xi})$, that we need to postulate. First of all, we will assume that the ξ's are independent, so that their distribution factorizes over the protein and complex indices

$$p(\boldsymbol{\xi}) = \prod_{i\mu} p(\xi_i^\mu).$$

This is the easiest assumption that we can make and will need to be checked *a posteriori*. Next, we need to postulate $p(\xi_i^\mu)$. As each ξ_i^μ is a binary variable which can only take values 0 and 1, we only need to specify $p(\xi_i^\mu = 1)$, and $p(\xi_i^\mu = 0) = 1 - p(\xi_i^\mu = 1)$. We have three natural choices for $p(\xi_i^\mu = 1)$, mimicking (i) a complex-driven, (ii) a protein-driven and (iii) a mixed mechanism for complex genesis.

(i) A natural choice is to assume that complexes have given sizes $\{q_\mu\}$, distributed according to $P(q) = (\alpha N)^{-1}\sum_\mu \delta_{q,q_\mu}$, which are determined, for example, by the functions that they are called to carry out inside the cell, and the likelihood of a protein *i* making part of a complex μ is given by the number

q_μ of proteins participating in complex μ divided by the total number of proteins,

$$p(\xi_i^\mu = 1) = \frac{q_\mu}{N} \quad \text{and} \quad p(\xi_i^\mu = 0) = 1 - \frac{q_\mu}{N}. \quad (2.10)$$

In random graphs $\boldsymbol{\xi}$ built according to this prescription, for large N, each complex size $q_\mu(\boldsymbol{\xi})$ is a Poissonian random variable with average q_μ, and all protein promiscuities $d_i(\boldsymbol{\xi})$ will be Poissonian variables with the same average $\langle d\rangle = \alpha\langle q\rangle$, with $\langle q\rangle = \sum_q P(q)q$ (see appendix A). Hence, in this ensemble, complex sizes are prescribed on average, i.e. $\langle q_\mu(\boldsymbol{\xi})\rangle = q_\mu \; \forall \mu$, and protein promiscuities are homogeneous, Poissonian variables with average determined from the average size of complexes, i.e. $\langle d_i(\boldsymbol{\xi})\rangle = \alpha\langle q\rangle \; \forall \; i$. As in this ensemble proteins' recruitment to complexes is determined by functions, we will refer to this ensemble as 'function-driven' or more briefly, 'q-ensemble'.

(ii) An alternative choice is to assume that proteins have given propensities to interact $\{d_i\}$, distributed according to $P(d) = N^{-1}\sum_i \delta_{d,d_i}$, which are determined, for example, by the number of their binding sites, polarization, etc., and the likelihood of a protein i making part of complex μ is given by the number of complexes which involve protein i divided by the total number of complexes,

$$p(\xi_i^\mu = 1) = \frac{d_i}{\alpha N} \quad \text{and} \quad p(\xi_i^\mu = 0) = 1 - \frac{d_i}{\alpha N}. \quad (2.11)$$

For large graphs $\boldsymbol{\xi}$ drawn from ensemble (2.11), each protein promiscuity $d_i(\boldsymbol{\xi})$ is a Poissonian variable with average d_i, whereas all complex sizes $q_\mu(\boldsymbol{\xi})$ are Poisson variables with the same average $\langle q\rangle = \langle d\rangle/\alpha$, with $\langle d\rangle = \sum_d P(d)d$ (appendix A). In this ensemble, it is therefore assumed that protein binding is driven by protein promiscuities, and we will refer to it as 'protein-driven' or 'd-ensemble'.

(iii) A third obvious choice is to assume that protein promiscuities $\{d_i\}$ and complex sizes $\{q_\mu\}$ are distributed according to given $P(d)$ and $P(q)$, respectively, and the likelihood of protein i participating in complex μ is controlled by both protein promiscuity and complex size

$$p(\xi_i^\mu = 1) = \frac{q_\mu d_i}{\alpha N\langle q\rangle} \quad \text{and}$$
$$p(\xi_i^\mu = 0) = 1 - \frac{q_\mu d_i}{\alpha N\langle q\rangle}. \quad (2.12)$$

Large graphs $\boldsymbol{\xi}$ drawn from this ensemble will have all protein promiscuities and complex sizes constrained on average, i.e. $\langle d_i(\boldsymbol{\xi})\rangle = d_i$ and $\langle q_\mu(\boldsymbol{\xi})\rangle = q_\mu$, with $\{d_i\}$ and $\{q_\mu\}$ distributed according to $P(d)$ and $P(q)$. In this third scenario, protein-binding statistics is driven both by complex functionality and protein promiscuity factors, and we will refer to this as the 'mixed ensemble'.

The mixed ensemble (2.12) reduces to (2.10) for the choice of homogeneous protein promiscuities $P(d) = \delta_{d,\alpha\langle q\rangle}$, and to (2.11) for the choice of homogeneous complex size $P(q) = \delta_{q,\langle q\rangle}$. By determining which of the above ensembles reflects better biological reality, we will learn about the mechanisms with which complexes and proteins recruit each other.

2.4. Accounting for binding sites

In all PINs, each protein is reduced to a simple network node, in spite of the fact that proteins are in reality complex chains of amino acids with several binding domains. Here we show that the ensembles introduced in the previous section can accommodate the presence of multiple binding sites when these are equally reactive. Let us first assume that each protein has d functional reactive amino acid endgroups. When two such proteins bind, the resulting dimer has $2d-2$ unused reactive endgroups, a trimer has $3d-4$ endgroups and a k-mer has $kd-2(k-1) = (d-2)k+2$ endgroups. If all endgroups are equally reactive, the *a priori* probability that a protein i is part of a complex μ is given by

$$p(\xi_i^\mu = 1) = \frac{d[(d-2)q_\mu + 2]}{Z} \simeq \frac{q_\mu d}{\alpha N\langle q\rangle}, \quad (2.13)$$

where the last approximate equality holds for $d \gg 1$ and $Z = \sum_\mu q_\mu d = \alpha N\langle q\rangle d$. This corresponds to ensemble (2.10), with the choice $d = \alpha\langle q\rangle$. If proteins have different endgroups d_i,

$$p(\xi_i^\mu = 1) \simeq \frac{d_i[(d-2)q_\mu + 2]}{\alpha N\langle q\rangle d} \simeq \frac{d_i q_\mu}{\alpha N\langle q\rangle}, \quad (2.14)$$

where $d = N^{-1}\sum_i d_i$, leading to ensemble (2.12). If the variability of q_μ is small, $q_\mu \simeq \langle q\rangle$,

$$p(\xi_i^\mu = 1) = \frac{d_i}{\alpha N}, \quad (2.15)$$

and we retrieve (2.11). The assumption of unbiased interactions between proteins with varying individual binding affinities has been supported in [13], and in recent structural analysis on residue-type-independent interactions [14].

3. Overview of results

In this section, we summarize the main results of this paper, that will be derived in full details in the next sections, suitable for readers with a mathematical or a statistical background. Readers with a less quantitative background who are interested in the biological applications of the mathematical framework introduced above, can stop at the end of this section and skip the mathematical details presented later.

3.1. Test relationships

The main result of this paper is that graphs with elements $c_{ii} = 0 \; \forall \, i$ and

$$c_{ij} = \sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu \quad \forall \; i \neq j,$$

where the $\boldsymbol{\xi}$ are drawn from either the 'function-driven' distribution (2.10) or the 'protein-driven' distribution (2.11), display special relationships between the moments $\langle k\rangle$, $\langle k^2\rangle$, etc. of their degree distribution $p(k)$ and the density m_3, m_4, etc. of their loops of length 3, 4, etc. Remarkably, these relationships are completely independent of α, $P(q)$ and $P(d)$ and follow solely from the separable nature of the matrix c_{ij}. In addition, they are identical, to orders $\mathcal{O}(N^0)$, to those found in the binary matrices $\boldsymbol{a}$ with elements $a_{ij} = \theta(\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu) \; \forall \, i \neq j$ and $a_{ii} = 0 \; \forall \, i$. For the q-ensemble we have the following relationships:

$$m_3 = \langle k^2\rangle - \langle k\rangle^2 - \langle k\rangle \quad (3.1)$$

Table 1. List of the publicly available experimental protein interaction datasets as used in this study, together with their main quantitative characteristics (number of proteins N, average degree $\langle k \rangle$ and largest degree k_{max}) and references.

species	N	$\langle k \rangle$	k_{max}	method	references
C. elegans	2528	2.96	99	Y2H	[15]
C. jejuni	1324	17.5	207	Y2H	[16]
E. coli	2457	7.05	641	MS	[17]
H. pylori	724	3.87	55	Y2H	[18]
H. sapiens I	1499	3.37	125	Y2H	[19]
H. sapiens II	1655	3.71	95	Y2H	[20]
H. sapiens III	2268	5.67	314	MS	[21]
M. loti	1803	3.43	401	Y2H	[22]
P. falciparum	1267	4.17	51	Y2H	[23]
S. cerevisiae I	991	1.82	24	Y2H	[24]
S. cerevisiae II	787	1.91	55	Y2H	[25]
S. cerevisiae III	3241	2.69	279	Y2H	[25]
S. cerevisiae IV	1576	4.58	62	MS	[26]
S. cerevisiae VI	1358	4.73	53	MS	[27]
S. cerevisiae VIII	2551	16.77	955	MS	[28]
S. cerevisiae IX	2708	5.25	141	MS	[29]
S. cerevisiae X	1630	11.15	127	MS	[30]
Synechocystis	1903	3.25	51	Y2H	[31]
T. pallidum	724	10.01	285	Y2H	[32]

and

$$m_4 = \langle k^3 \rangle - 3\langle k^2 \rangle + 2\langle k \rangle + \langle k \rangle(\langle k^2 \rangle - \langle k \rangle - 2\langle k \rangle^2)$$
$$= \langle k^3 \rangle - 3\langle k^2 \rangle + 2\langle k \rangle - \langle k \rangle^3 - 3\langle k \rangle m_3, \quad (3.2)$$

whereas for the d-ensemble we have

$$m_4 = \frac{m_3^2}{\langle k \rangle}. \quad (3.3)$$

Remarkably, we obtain different sets of relationships for the two ensembles, meaning that the two ensembles do not represent equivalent descriptions of the bipartite representation of the proteome, as one may have naively expected. One can then check whether real PINs come closer to satisfy the test relationships from the q-ensemble or the ones from the d-ensemble. This will hint at the underlying mechanism with which proteins and complexes recruit each other. Next, we apply the above results to real publicly available protein interaction datasets, obtained via MS and Y2H experiments. The detailed quantitative features of the various datasets and their references are listed in table 1.

3.2. Application to mass spectrometry datasets

Seven of the experimental PIN datasets in table 1 were obtained by MS experiments, and they involved three distinct biological species, namely *Saccharomyces cerevisiae*, *Homo sapiens* and *Escherichia coli*. Each set takes the form of an $N \times N$ matrix of binary entries a_{ij}, but with different values of N. In figure 3, we show the results of our analytical predictions for the densities of length-3 and length-4 loops, as given by the formulae for the function- and protein-driven ensembles, versus their measured values in the MS datasets. Before looking at the performance of PIN data with respect to the above test formulae, let us briefly look at what we would expect in fully random networks, of the Erdös–Rényi (ER) type, which are not in the family of separable graphs. Let us denote with $m_{\ell q}$ and $m_{\ell d}$ the density of loop of length ℓ as predicted by the function- and the protein-driven ensemble, respectively, and with $m_{\ell \mathrm{m}}$ their measured value. In fully random graphs, the degree distribution is Poissonian and one has $\langle k^2 \rangle = \langle k \rangle + \langle k \rangle^2$ and $\langle k^3 \rangle = \langle k \rangle(1 + 3\langle k \rangle + \langle k^2 \rangle)$. Furthermore, measured values of loop densities are typically $m_{\ell \mathrm{m}} = \mathcal{O}(N^{-1})$ for $\ell = 3,4$. Hence, the r.h.s. of (3.1) and (3.2) vanish giving $m_{3q} = \mathcal{O}(N^{-1})$ and $m_{4q} = \mathcal{O}(N^{-1})$, whereas (3.3) gives $m_{3d} = \sqrt{\langle k \rangle m_{4\mathrm{m}}} = \mathcal{O}(N^{-1/2})$ and $m_{4d} = m_{3\mathrm{m}}^2/\langle k \rangle = \mathcal{O}(N^{-2})$. Hence, one has $m_{3d} \gg m_{3q} \backsim m_{3\mathrm{m}}$ and $m_{4\mathrm{m}} \backsim m_{4q} \gg m_{4d}$.

We now turn to the interpretation of plots in figure 3. First off, we note that the theoretical predictions from the function-driven ensemble lead to values of the number of short loops consistently higher than those predicted by the protein-driven ensemble, even for loops of length 3. This is remarkably different from the behaviour expected in random graphs of the ER type and is consistent with the behaviour of separable random graphs, where a function-driven complex genesis induces large cliques in the PINs, which boosts short loops. By contrast, a protein-driven complex genesis induces a homogeneous distribution for the complex sizes, which suppresses the presence of large cliques, hence of short loops, in the PINs. Notably, the densities of length-4 loops of all MS datasets are in between those of the d-ensemble (which thereby

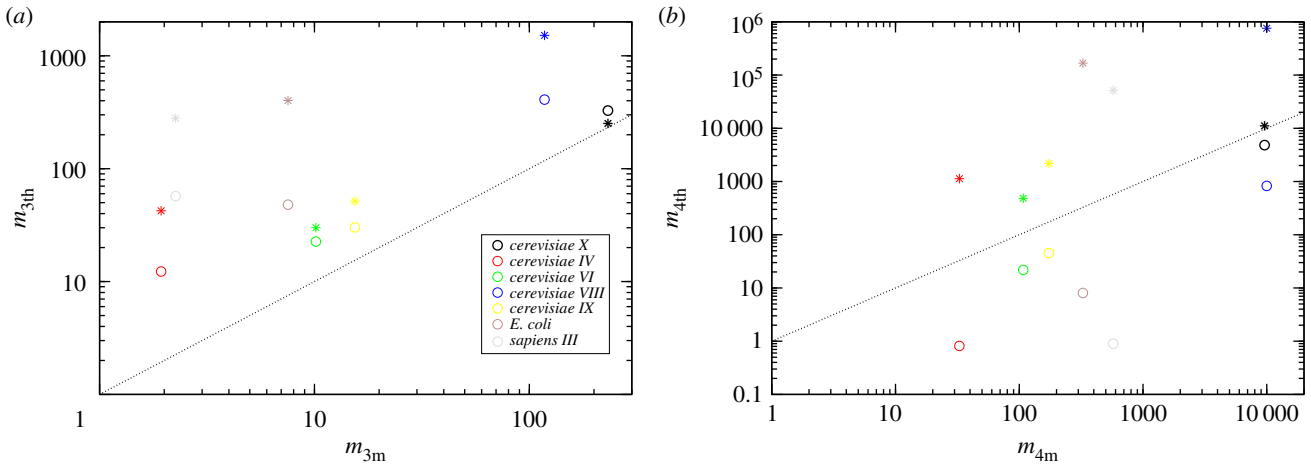

Figure 3. (*a*) Theoretical predictions m_{3th} for the densities of length-3 loops in the PINs, as obtained from the *q*-ensemble (stars) and the *d*-ensemble (circles), plotted versus the values m_{3m} measured in the different MS datasets. (*b*) Theoretical predictions m_{4th} for the densities of length-4 loops in the same PINs, obtained from the *q*-ensemble (stars) and the *d*-ensemble (circles), plotted versus the measured values m_{4m}. The diagonals are shown as guides to the eye. (Online version in colour.)

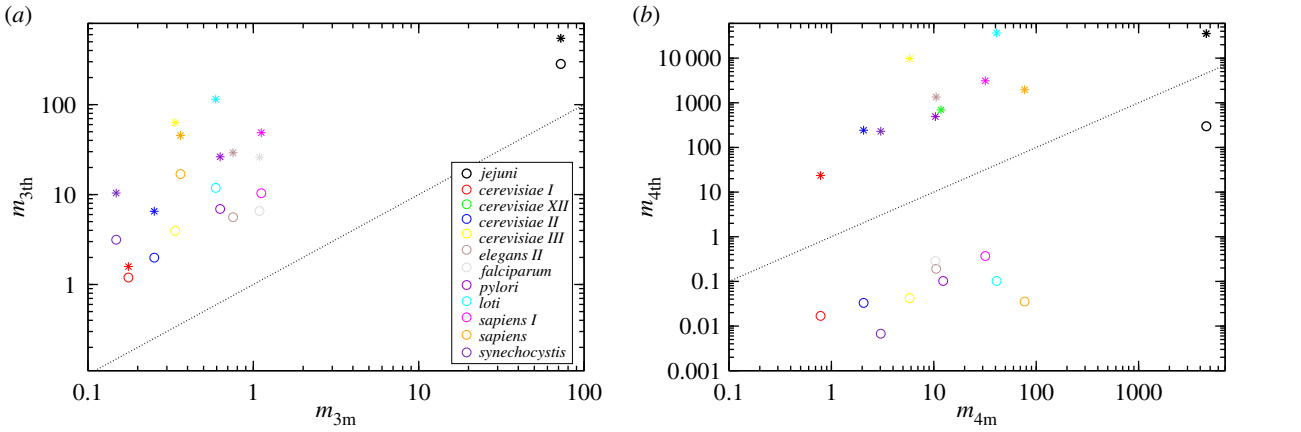

Figure 4. (*a*) Theoretical predictions m_{3th} for the densities of length-3 loops in the PINs, as obtained from the *q*-ensemble (stars) and the *d*-ensemble (circles), plotted versus the values m_{3m} measured in the different Y2H datasets. (*b*) Theoretical predictions m_{4th} for the densities of length-4 loops in the same PINs, obtained from the *q*-ensemble (stars) and the *d*-ensemble (circles), plotted versus the measured values m_{4m}. The diagonals are shown as guides to the eye. (Online version in colour.)

acts as a lower bound) and those of the *q*-ensemble (which acts as an upper bound). This suggests a compatibility of data from MS experiments with the expected separable form of the proteome network. However, the measured densities of length-3 loops are consistently lower than the values compatible with a separable structure of the proteome.

To shed light on this result, it is useful to compare this behaviour with those of random graphs with the same degree distribution as our PINs. As the latter is non-Poissonian, the theoretical predictions for the loop densities given by the function-driven ensemble are now $m_{3q} = \mathcal{O}(1)$ and $m_{4q} = \mathcal{O}(1)$, whereas both m_{3m} and m_{4m} are still $\mathcal{O}(N^{-1})$, yielding the same m_{3d} and m_{4d} as in ER networks. This now leads to $m_{3q} \gg m_{3d} \gg m_{3m}$ and $m_{4q} \gg m_{4m} \gg m_{4d}$.

In real PINs, we observe the same patterns of inequalities; however, the measured values of loop densities are $\mathcal{O}(1)$ hence much larger than their expected values in random graphs with the same degree distribution. In particular, several datasets (including *cerevisiae VI* and *cerevisiae IX*) show a loop density m_3 of the same magnitude as the one predicted by the separable models and dataset *cerevisiae X* is in excellent agreement with the predicted value. This suggests a degree of compatibility with the proposed separable model, which enforces many more loops than in a typical random network with the same degree distribution, and also a large degree of noise which tends to randomize interactions.

3.3. Applications to yeast two-hybrid datasets

We tested similarly the compatibility of Y2H data with a separable structure of the proteome, by checking whether the measured values for the network observables m_3 and m_4 fall within what appeared to be (in separable graph ensembles) theoretical bounds set by the function- and protein-driven ensembles. We now used the 12 PIN datasets in table 1 that were obtained from Y2H experiments. Results are shown in figure 4. We observe that Y2H datasets exhibit generally fewer short loops than MS dataset. This may be due to the fact that, at variance with MS datasets, which use immunoprecipitations to sample from a functioning biological network, Y2H experiments sample proteins from the entire potential biophysical network, a much larger interaction space than any given one cell-type/tissue sampled by MS, leading to an undersampling of links and therefore to underestimation of connectivity and loops. Qualitatively

Y2H datasets show similar trends to MS, with measured values of m_4 somewhat compatible with separable models, and values for m_3 that fall below. This is quite remarkable, as MS and Y2H experiments are known to measure interactions in very different ways. However, quantitatively, m_3 is generally an order of magnitude less than that predicted by separable models, with the exception of dataset *jejuni* which stays closer to the predicted values. This suggests that Y2H has a lower level of compatibility with a separable structure of the proteome.

3.4. Origin of fat tails in the degree distribution of protein interaction network

In previous sections, we have introduced the promiscuity distribution $P(d)$, the complex size distribution $P(q)$ and the degree distribution $p(k)$. The former distributions are properties of the bipartite graph $\boldsymbol{\xi}$, whereas the latter is a property of the marginalized graph $\boldsymbol{c}$, or $\boldsymbol{a}$. The degree distribution $p(k)$ can be computed directly from the graph $\boldsymbol{a}$ and for PINs it typically displays a fat tail, $p(k) \simeq Ck^{-\mu}$ for large k with $2 < \mu < 3$ [33–36].

By contrast, $P(d)$ and $P(q)$ cannot be measured directly, but they are related to $p(k)$. This allows one in principle to infer the tail behaviour of the promiscuity and complex size distributions from the tail of the degree distribution of the PIN, which can be easily computed. The relationship between the above distributions depends on the mechanism driving complex genesis, i.e. (i) function-driven, (ii) protein-driven or (iii) mixed.

(i) For function-driven ensembles, one has for large q

$$P(q) \simeq \left(\frac{C}{\alpha}\right) q^{-\mu-1} \tag{3.4}$$

and $P(d) = \delta_{\langle d\rangle,\alpha\langle q\rangle}$. This shows that the complex size distribution $P(q)$ decays faster than the degree distribution of the associated PIN $\boldsymbol{a}$, so fat tails in the degree distribution $p(k)$ of PINs can emerge from less heterogeneous complex size distributions. In particular, complex size distributions $P(q)$ with a finite second moment (but diverging higher moments) give scale-free degree distributions $p(k)$. This is consistent with the intuition that, while large hubs are often observed in PINs, super-complexes of the same number of proteins are unlikely to be stable. Indeed, many interactions in hubs are 'date' types, as opposed to 'party' type [11]. Our framework allows us to discriminate between different types of hub proteins, and suggests that heterogeneous (i.e. power law) behaviour in PINs may emerge from homogeneous protein 'dating' and moderately heterogeneous protein 'partying'.

(ii) For protein-driven ensembles, one has for large d,

$$P(d) \simeq C' d^{-\mu} \tag{3.5}$$

with

$$C' = C\left(\frac{\alpha}{\langle d\rangle}\right)^{\mu-1} = C\langle q\rangle^{1-\mu}, \tag{3.6}$$

whereas $P(q) = \delta_{\langle q\rangle,\langle d\rangle/\alpha}$. Hence, any tail in the promiscuity distribution will produce the same tail in the degree distribution of $\boldsymbol{a}$, but with a rescaled amplitude. Fat tails in the degree distribution of PINs can thus arise from equally heterogeneous 'dating' interactions between proteins, combined with a homogeneous distribution of 'party' interactions. Short loops are boosted by broad distributions of complex sizes, as large complexes in the bipartite graph induce large cliques in the network $\boldsymbol{a}$. The d-ensemble (4.2), which attributes any heterogeneity in $p(k)$ to heterogeneity of protein binding promiscuities, generates separable PIN graphs $\boldsymbol{a}$ with the least number of loops. Conversely, the q-ensemble (4.1), which attributes all heterogeneity in $p(k)$ to heterogeneity in complex sizes, generates separable PIN graphs $\boldsymbol{a}$ with the largest number of loops.

(iii) For the mixed ensembles, it is easier to write relationships in terms of the distribution $W(q)$, related to $P(q)$ via $W(q) = qP(q)/\langle q\rangle$. Assuming $W(q)$ has a power-law tail, with a finite first moment (as in both cases previously considered), i.e. $W(q) \simeq Kq^{-\gamma}$ with $\gamma > 2$, one has $P(d) \sim C'd^{-\mu}$ with $\mu \leq \gamma$, where $C' = C(\langle q^2\rangle/\langle q\rangle)^{1-\mu}$ for $\gamma > \mu$ and $K\langle d\rangle + C'(\langle q^2\rangle/\langle q\rangle)^{\mu-1} = C$ for $\gamma = \mu$. This means that if $W(q)$ decays faster than $p(k)$ (as for protein-driven recruitment), then the tail in $p(k)$ must arise from the tail in $P(d)$, although heterogeneities in $P(q)$ will affect the amplitude of the power-law tail in $P(d)$. Conversely, if $P(d)$ is as broad as $W(q)$, then both $P(q)$ and $P(d)$ contribute to the tail of $p(k)$, whose amplitude will be the sum of the amplitudes of the tails of the two distributions.

4. Mathematical set-up

The remainder of this paper is devoted to the derivation of results presented in §3. We start by casting the objectives of our study in mathematical language. Formally, we can write the bipartite network distributions for the three different type of protein–complex recruitment (i.e. complex-driven, protein-driven and mixed) as

$$p(\boldsymbol{\xi}) = \prod_{i\mu}\left[\frac{q_\mu}{N}\delta_{\xi_i^\mu,1} + \left(1 - \frac{q_\mu}{N}\right)\delta_{\xi_i^\mu,0}\right], \tag{4.1}$$

$$p(\boldsymbol{\xi}) = \prod_{i\mu}\left[\frac{d_i}{\alpha N}\delta_{\xi_i^\mu,1} + \left(1 - \frac{d_i}{\alpha N}\right)\delta_{\xi_i^\mu,0}\right] \tag{4.2}$$

and

$$p(\boldsymbol{\xi}) = \prod_{i\mu}\left[\frac{d_i q_\mu}{\alpha N\langle q\rangle}\delta_{\xi_i^\mu,1} + \left(1 - \frac{d_i q_\mu}{\alpha N\langle q\rangle}\right)\delta_{\xi_i^\mu,0}\right], \tag{4.3}$$

respectively. Bipartite graphs drawn from (4.1) were found to have modular topologies, and to accomplish parallel information processing for suitable values of the parameter α [37,38], and their connection to PINs has been pointed out in [39]. As explained above, the three ensembles become equivalent when complex sizes and protein promiscuities are both homogeneous, i.e. $q_\mu = \langle q\rangle\ \forall\,\mu$ and $d_i = \alpha\langle q\rangle\ \forall\, i$. In that case, the recruitment process between proteins and complexes is fully random. We are interested in the properties of the random graph ensemble

$$p(\boldsymbol{a}) = \left\langle \left[\prod_{i<j}\delta_{a_{ij},\theta\left(\sum_{\mu\leq\alpha N}\xi_i^\mu\xi_j^\mu\right)}\right]\left[\prod_i \delta_{a_{ii},0}\right]\right\rangle_{\boldsymbol{\xi}}, \tag{4.4}$$

where $\langle\cdot\rangle_{\boldsymbol{\xi}} = \sum_{\boldsymbol{\xi}} \cdot\, p(\boldsymbol{\xi})$ and $p(\boldsymbol{\xi})$ is given by (4.1)–(4.3). Some properties of (4.4) will turn out not to depend on the choices made for the distributions of complex sizes and protein promiscuities, and this leads to powerful benchmarks against which to test available PIN datasets.

A key step of our analysis is that averages over (4.4) can often be replaced by averages over the ensemble of *weighted* graphs (2.6)

$$p(\boldsymbol{c}) = \left\langle \left[\prod_{i<j} \delta_{c_{ij},\sum_{\mu\leq\alpha N}\xi_i^\mu\xi_j^\mu}\right]\left[\prod_i \delta_{c_{ii},0}\right]\right\rangle_{\boldsymbol{\xi}}. \tag{4.5}$$

For finite q_μ, d_i and α, one finds that in large networks generated via (4.1)–(4.3), the probability of seeing $c_{ij} > 1$ is of order $\mathcal{O}(N^{-2})$, and the values of many macroscopic observables in the **a** and $\boldsymbol{c}$ ensembles will, to leading order in N, be identical.

5. Network properties generated by the q-ensemble

In this section, we study the statistical properties of the ensembles (4.5) and (4.4) upon generating the bipartite protein interaction graph $\boldsymbol{\xi}$ from ensemble (4.1), where complexes recruit proteins.

5.1. Link probabilities

For the graphs $\boldsymbol{c}$ of (4.5), we find the following expectation values of individual bonds:

$$\langle c_{ij}\rangle = \sum_{\mu=1}^{\alpha N}\langle\xi_i^\mu\xi_j^\mu\rangle_\xi = \sum_{\mu=1}^{\alpha N}\left(\frac{q_\mu}{N}\right)^2 = \frac{\alpha}{N}\langle q^2\rangle, \tag{5.1}$$

where the brackets on the r.h.s. denote averaging over the complex size distribution $P(q)$. The likelihood of an individual bond is (see appendix B)

$$\begin{aligned} p(c_{ij}) &= \langle\delta_{c_{ij},\sum_{\mu\leq\alpha N}\xi_i^\mu\xi_j^\mu}\rangle_\xi \\ &= \delta_{c_{ij},0} + \frac{\alpha\langle q^2\rangle}{N}(\delta_{c_{ij},1}-\delta_{c_{ij},0}) + \left(\frac{\alpha^2\langle q^2\rangle^2}{2N^2} - \frac{1}{2}\frac{\alpha\langle q^4\rangle}{N^3}\right) \\ &\quad\times(\delta_{c_{ij},2}-2\delta_{c_{ij},1}+\delta_{c_{ij},0}) + \frac{\alpha^3\langle q^2\rangle^3}{6N^3} \\ &\quad\times(\delta_{c_{ij},3}-3\delta_{c_{ij},2}+3\delta_{c_{ij},1}-\delta_{c_{ij},0}) + \mathcal{O}(N^{-4}), \end{aligned} \tag{5.2}$$

so we find for the first few probabilities:

$$p(0) = 1 - \frac{\alpha\langle q^2\rangle}{N} + \frac{\alpha^2\langle q^2\rangle^2}{2N^2} - \frac{\alpha\langle q^4\rangle}{2N^3} - \frac{\alpha^3\langle q^2\rangle^3}{6N^3} + \mathcal{O}(N^{-4}) \tag{5.3}$$

and

$$p(1) = \frac{\alpha\langle q^2\rangle}{N} - \frac{\alpha^2\langle q^2\rangle^2}{N^2} + \frac{\alpha\langle q^4\rangle}{N^3} + \frac{\alpha^3\langle q^2\rangle^3}{2N^3} + \mathcal{O}(N^{-4}), \tag{5.4}$$

and hence

$$\begin{aligned} \sum_{\ell>1} p(\ell) &= 1 - p(0) - p(1) = \mathcal{O}(N^{-2}) \text{ and} \\ \sum_{\ell>1} \ell p(\ell) &= \langle c_{ij}\rangle - p(1) = \mathcal{O}(N^{-2}). \end{aligned} \tag{5.5}$$

The probability to have $c_{ij} \neq 0$ is of order $\mathcal{O}(N^{-1})$, so the graphs generated by (4.5) are finitely connected. Moreover, although the graphs $\boldsymbol{c}$ are in principle weighted, for large N the number of links per node that are not in $\{0,1\}$ will be vanishingly small.

5.2. Densities of short loops

We now turn to the calculation of expectation values for different observables in ensemble (4.5). First, we calculate the average number of ordered and oriented loops of length-3 per node, which are (see appendix B)

$$m_3 = \left\langle \frac{1}{N}\sum_{ijk} c_{ij}c_{jk}c_{ki}\right\rangle_\xi = \frac{1}{N}\sum_{\mu\nu\rho=1}^{\alpha N}\sum_{i\neq j\neq k}\langle\xi_i^\mu\xi_j^\mu\xi_j^\nu\xi_k^\nu\xi_k^\rho\xi_i^\rho\rangle_\xi \tag{5.6}$$

$$= \alpha\langle q^3\rangle + \mathcal{O}(N^{-1}). \tag{5.7}$$

Calculating the density of loops m_L for lengths $L > 3$ can be simplified by returning to the bipartite graph $\boldsymbol{\xi}$. We define a star S_n to be a simple $(n+1)$-node tree in $\boldsymbol{\xi}$, of which the central node belongs to ν_c (the complexes), and the n leaves belong to ν_p (the proteins). Thus, S_2 stars represent protein dimers, S_3 stars represent protein trimers, and so on. Each link in $\boldsymbol{c}$ corresponds to at least one S_2 star in the bipartite graph (which, in turn, can be a subset of any S_n star with $n > 2$). Therefore, the total number of S_2 stars in the bipartite graph,

$$\begin{aligned} \sum_\mu\sum_{i\neq j}\langle\xi_i^\mu\xi_j^\mu\rangle &= \sum_\mu\sum_{i\neq j}\langle\xi_i^\mu\rangle\langle\xi_j^\mu\rangle = \sum_{i\neq j}\sum_\mu\frac{q_\mu^2}{N^2} \\ &= \alpha(N-1)\langle q^2\rangle, \end{aligned} \tag{5.8}$$

has to equate in leading order the total number of links $N\langle k\rangle$ in graph $\boldsymbol{c}$, yielding

$$\langle q^2\rangle = \frac{\langle k\rangle}{\alpha} + \mathcal{O}(N^{-1}), \tag{5.9}$$

which is indeed in agreement with the result of the direct calculation $\langle k\rangle = N^{-1}\sum_{ij}\langle c_{ij}\rangle$, using (5.1). Similarly, we can obtain the number of loops of length 3, calculated earlier, by realizing that these loops arise when we have in the bipartite graph either a star S_3 (which can be a subset of any S_n with $n > 3$) or a combination of three S_2 stars, where every leaf is shared by two stars. The contribution of the number of S_3 stars per node to the number of loops of length 3 is

$$\begin{aligned} \frac{1}{N}\sum_\mu\sum_{i\neq j\neq k(\neq i)}\langle\xi_i^\mu\xi_j^\mu\xi_k^\mu\rangle &= \frac{1}{N}\sum_\mu\sum_{i\neq j\neq k(\neq i)}\langle\xi_i^\mu\rangle\langle\xi_j^\mu\rangle\langle\xi_k^\mu\rangle \\ &= \frac{1}{N}\sum_\mu\sum_{i\neq j\neq k(\neq i)}\frac{q_\mu^3}{N^3} = \alpha\langle q^3\rangle + \mathcal{O}(N^{-1}). \end{aligned} \tag{5.10}$$

The contribution of the combination of three S_2 stars, where each leaf is shared by two stars, is

$$\begin{aligned} \frac{1}{N}\sum_{[\mu,\nu,\rho]}\sum_{[i,j,k]}\langle\xi_i^\mu\xi_j^\mu\xi_j^\nu\xi_k^\nu\xi_k^\rho\xi_i^\rho\rangle &= \frac{1}{N}\sum_{[\mu,\nu,\rho]}\sum_{[i,j,k]}\frac{q_\mu^2q_\nu^2q_\rho^2}{N^6} \\ &= \frac{1}{N}\alpha^3\langle q^2\rangle^3 + \mathcal{O}(N^{-1}), \end{aligned} \tag{5.11}$$

with the square brackets $[i,j,k]$ denoting that the three indices are distinct. The expected density of length 3 loops is the sum of an $\mathcal{O}(1)$ contribution from S_3 stars, plus an $\mathcal{O}(N^{-1})$ contribution from combinations of three S_2 stars that share leaves.

For large N the second contribution vanishes, and we recover $m_3 = \alpha\langle q^3\rangle$. Likewise, the $\mathcal{O}(1)$ contribution to the density of length-4 loops comes from S_4 stars in the bipartite graph, which consist of five sites (four leaves and one central node) and four links, each with probability $\mathcal{O}(N^{-1})$. Combinations of two S_3 stars with two shared leaves, or of S_2 stars, always involve a number of links at least equal to the number of nodes and therefore yield sub-leading contributions. Hence, the density of loops of length 4 is

$$m_4 = \frac{1}{N}\sum_{\mu}\sum_{[i,j,k,\ell]}\langle \xi_i^\mu \xi_j^\mu \xi_k^\mu \xi_\ell^\mu\rangle = \alpha\langle q^4\rangle + \mathcal{O}(N^{-1}). \tag{5.12}$$

More generally, the average density of loops of arbitrary length L is given by

$$m_L = \alpha\langle q^L\rangle + \mathcal{O}(N^{-1}). \tag{5.13}$$

For large N the ratio α and the distribution $P(q)$ of complex sizes apparently determine in full the statistics of loops of arbitrary length in $\boldsymbol{c}$, if the protein interactions are described by (4.1).

Finally, we note that if m_L gives the number of ordered and oriented loops of length L per node, the number of unordered and unoriented closed paths of length L equals $\bar{m}_L = m_L/6$, as there are L possible nodes to start a closed path from, and two possible orientations.

5.3. The degree distribution

It follows from (5.9) and (5.13) that by measuring the average degree $\langle k\rangle$ and the densities m_L of loops of length L we can compute all the moments of the distribution of complex sizes $P(q)$:

$$\langle q^2\rangle = \frac{\langle k\rangle}{\alpha} \quad \text{and} \quad \forall L > 2\text{: } \langle q^L\rangle = \frac{m_L}{\alpha}. \tag{5.14}$$

This would allow us to calculate $P(q)$ in full via its generating function, provided α and $\langle q\rangle$ are known. However, counting the number of loops of arbitrary length in a graph is computationally challenging, and α and $\langle q\rangle$ are generally unknown. However, it is possible to express $P(q)$ for large N in terms of the degree distribution $P(k)$ of $\boldsymbol{c}$. Specifically, in appendix C, we show that

$$\lim_{N\to\infty} p(k) = \frac{\int_0^\infty \mathrm{d}y\, P(y)\, \mathrm{e}^{-y} y^k}{k!}, \tag{5.15}$$

where

$$P(y) = \mathrm{e}^{-\alpha\langle q\rangle}\sum_{\ell\geq 0}\frac{(\alpha\langle q\rangle)^\ell}{\ell!}\sum_{q_1\ldots q_l\geq 0} W(q_1)\ldots W(q_\ell)\,\delta\left[y - \sum_{r\leq\ell} q_r\right], \tag{5.16}$$

and $W(q) = qP(q)/\langle q\rangle$ is the likelihood to draw a link attached to a complex node of degree q in the bipartite graph $\boldsymbol{\xi}$. Formula (5.15) is easily interpreted. The degree of node i in $\boldsymbol{c}$ is given by the second neighbours of i in $\boldsymbol{\xi}$; the number ℓ of first neighbours of node i will thus be a Poissonian variable with average $\alpha\langle q\rangle$, and each of its ℓ first neighbours will have a degree q_r drawn from $W(q_r)$. Clearly, any tail in the distribution $W(q)$ will induce a tail in the distribution $p(k)$, with (as we will show below) the same exponent, but an amplitude that is reduced by a factor $\alpha\langle q\rangle$.

One can complement (5.15) with a reciprocal relationship that gives $P(q)$ in terms of $p(k)$. To achieve this, we define the generating functions $Q_1(z) = \sum_k p(k)\mathrm{e}^{-kz}$, $Q_2(z) = \int_0^\infty \mathrm{d}y\, P(y)\mathrm{e}^{-yz}$ and $Q_3(z) = \sum_q W(q)\mathrm{e}^{-zq}$. We then see from expression (5.15) for $p(k)$ that

$$\begin{aligned} Q_1(z) &= \int_0^\infty \mathrm{d}y\, P(y)\, \mathrm{e}^{-y}\sum_{k\geq 0}\frac{(y\mathrm{e}^{-z})^k}{k!} \\ &= \int_0^\infty \mathrm{d}y\, P(y)\, \mathrm{e}^{y[\mathrm{e}^{-z}-1]} = Q_2(1-\mathrm{e}^{-z}) \end{aligned} \tag{5.17}$$

and

$$\begin{aligned} Q_2(z) &= \mathrm{e}^{-\alpha\langle q\rangle}\sum_{\ell\geq 0}\frac{(\alpha\langle q\rangle)^\ell}{\ell!}\sum_{q_1\ldots q_\ell\geq 0} W(q_1)\ldots W(q_\ell)\,\mathrm{e}^{-z\sum_{r\leq\ell} q_r} \\ &= \mathrm{e}^{-\alpha\langle q\rangle}\sum_{\ell\geq 0}\frac{(\alpha\langle q\rangle Q_3(z))^\ell}{\ell!} = \mathrm{e}^{\alpha\langle q\rangle[Q_3(z)-1]}. \end{aligned} \tag{5.18}$$

The first identity can be rewritten as $Q_1(-\log(1-y)) = Q_2(y)$. Inserting this into (5.18) allows us to express the desired $Q_3(z)$ as

$$Q_3(z) = 1 + \frac{\log Q_2(z)}{\alpha\langle q\rangle} = 1 + \frac{\log Q_1(-\log(1-z))}{\alpha\langle q\rangle}, \tag{5.19}$$

which translates into

$$\sum_{q>0} P(q) q\mathrm{e}^{-zq} = \langle q\rangle + \frac{1}{\alpha}\log\sum_k p(k)(1-z)^k. \tag{5.20}$$

We can now extract the asymptotic form of $P(q)$ from that of $p(k)$. The generating functions $Q_1(z)$ of degree distributions that exhibit prominent tails, i.e. $p(k) \simeq Ck^{-\mu}$ for large k with $2 < \mu < 3$ (as observed in PINs [33–36]), are for small z of the form

$$Q_1(z) = 1 - \langle k\rangle z + C\Gamma(1-\mu)z^{\mu-1} + \cdots, \tag{5.21}$$

where Γ is Euler's gamma function [40]. For small z, we may use $1 - z \simeq \mathrm{e}^{-z}$ to rewrite (4.19) as

$$\log Q_1(z) \simeq \alpha\langle q\rangle[Q_3(z) - 1]. \tag{5.22}$$

Combining this with (5.21) then gives, for small z,

$$-\langle k\rangle z + C\Gamma(1-\mu)z^{\mu-1} \simeq \alpha\langle q\rangle[Q_3(z) - 1]. \tag{5.23}$$

Hence, for small z, $Q_3(z)$ has the same form as $Q_1(z)$,

$$Q_3(z) = 1 - \frac{\langle k\rangle}{\alpha\langle q\rangle}z + \frac{C}{\alpha\langle q\rangle}\Gamma(1-\mu)z^{\mu-1}. \tag{5.24}$$

Therefore, $W(q)$ behaves asymptotically in the same way as $p(k)$, i.e. $W(q) \simeq (C/\alpha\langle q\rangle)q^{-\mu}$. This, in turn, gives

$$P(q) \simeq \left(\frac{C}{\alpha}\right) q^{-\mu-1}. \tag{5.25}$$

5.4. Relationships that are independent of $P(q)$ and α

The first two moments of $p(k)$ are given, to leading order in N, by (see appendix C)

$$\langle k\rangle = \alpha\langle q^2\rangle + \mathcal{O}(N^{-1}), \tag{5.26}$$

which is in agreement with (5.9), and

$$\langle k^2\rangle = \alpha\langle q^2\rangle + \alpha\langle q^3\rangle + \alpha^2\langle q^2\rangle^2. \tag{5.27}$$

The latter is easily interpreted in terms of the underlying bipartite graph: $\langle k^2\rangle$ is the average density of paths of

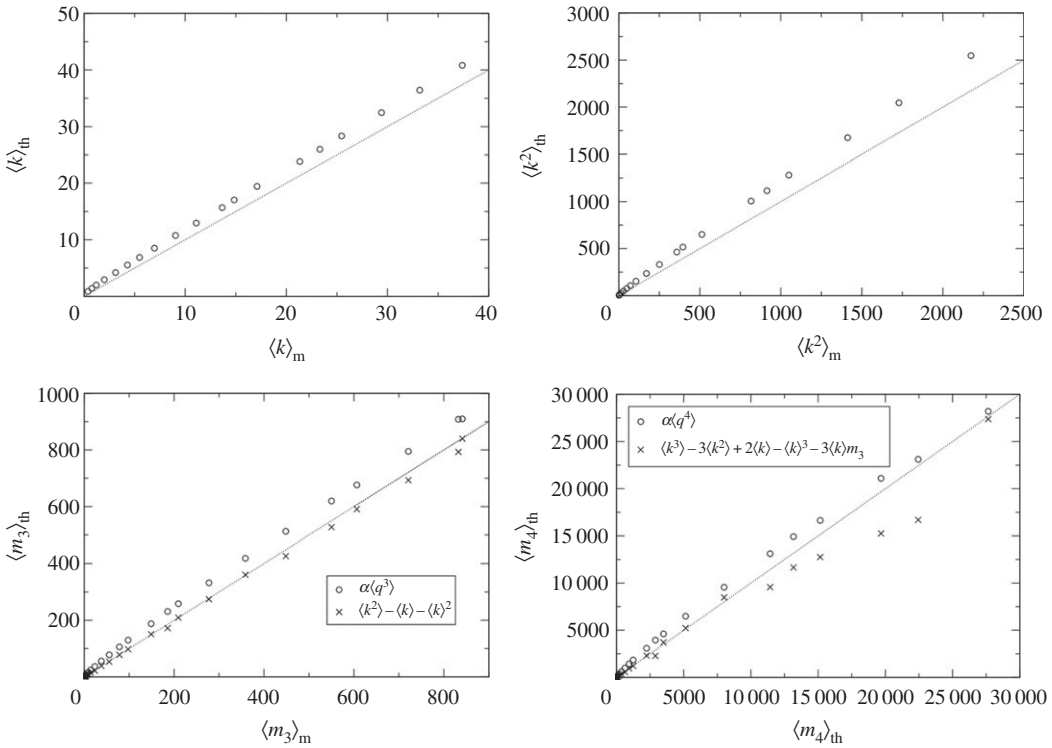

Figure 5. Symbols: theoretical $\langle \ldots \rangle_{\rm th}$ versus measured $\langle \ldots \rangle_{\rm m}$ values of observables $\langle k\rangle$, $\langle k^2\rangle$, m_3 and m_4 in synthetically random graphs $\boldsymbol{c}$ with $N = 3000$, defined via (4.1), (4.5) for a power-law distributed complex size distribution $P(q)$. Theoretical values are given by formulae (5.26) for $\langle k\rangle$, (5.27) for $\langle k^2\rangle$, (5.13) and (3.1) for m_3 and (5.13) and (3.2) for m_4. Dotted lines: the diagonals (as guides to the eye).

length two, so it has a contribution from $\langle k\rangle = \alpha\langle q^2\rangle$ because of backtracking, plus a contribution from pairs of S_2 stars that share a node, whose density is

$$\frac{1}{N}\sum_{[ijk]}\sum_{\mu\neq\nu}\langle \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_k^\nu\rangle = \frac{1}{N}\sum_{[ijk]}\sum_{\mu\neq\nu}\frac{q_\mu^2}{N^2}\frac{q_\nu^2}{N^2} = \alpha^2\langle q^2\rangle^2, \quad (5.28)$$

plus a contribution from S_3 stars, whose density is $\alpha\langle q^3\rangle$ (as shown earlier). Combining (5.27) with (5.14) gives us a relationship between average and width of the degree distribution of $\boldsymbol{c}$ and its density of length-3 loops. Remarkably, this relationship is completely independent of α and $P(q)$:

$$m_3 = \langle k^2\rangle - \langle k\rangle^2 - \langle k\rangle. \quad (5.29)$$

This identity and others, which all depend only on the separable underlying nature of the PIN and the assumption of complex-driven recruitment of proteins to complexes, can be derived more systematically from (5.20) by expanding both sides as power series in z and comparing the expansion coefficients. This gives a hierarchy of relationships between moments of $p(k)$ and $P(q)$, and hence (via (5.13)) between moments of $p(k)$ and densities of loops of increasing length, that are all completely independent of α and $P(q)$. At order z^2, one recovers (5.29). The next order z^3 leads to

$$\begin{aligned} m_4 &= \langle k^3\rangle - 3\langle k^2\rangle + 2\langle k\rangle + \langle k\rangle(\langle k^2\rangle - \langle k\rangle - 2\langle k\rangle^2) \\ &= \langle k^3\rangle - 3\langle k^2\rangle + 2\langle k\rangle - \langle k\rangle^3 - 3\langle k\rangle m_3. \end{aligned} \quad (5.30)$$

To test these asymptotic identities in finite systems, we generate random graphs $\boldsymbol{c}$ of size $N = 3000$ according to (4.1) and (4.5), and we compared the measured values of m_3 and m_4 in these random graphs with the predictions of formulae (5.29) and (5.30), respectively. We show the results in figure 5.

5.5. Link between *a* and *c* graph definitions

In conventional experimental PIN databases, one records only whether or not protein pairs interact, not the *number* of complexes in which they interact. Hence, protein interactions are normally represented in terms of the adjacency matrix $\boldsymbol{a} = \{a_{ij}\}$, which is related to the weighted matrix $\boldsymbol{c} = \{c_{ij}\}$ via $a_{ij} = \theta(c_{ij})\ \forall\ (i \neq j)$, with the convention for the step function $\theta(0) = 0$. We therefore have $p(a_{ij}) = \langle\delta_{c_{ij},0}\rangle\delta_{a_{ij},0} + (1 - \langle\delta_{c_{ij},0}\rangle)\delta_{a_{ij},1}$. However, the links $\{a_{ij}\}$ are correlated. In appendix D, we derive the relationship between the expected values of different graph observables for the two graph ensembles $p(\boldsymbol{a})$ and $p(\boldsymbol{c})$. Denoting averages in the $\boldsymbol{a}$ ensemble as $\langle \ldots \rangle_a$, and using the usual notation $\langle \ldots \rangle$ for averages in the $\boldsymbol{c}$ ensemble, one finds that for large N the first two moments of the degree distributions and the first two loop densities in the two ensembles are identical:

$$\begin{aligned} \langle k\rangle_a &= \frac{1}{N}\sum_{ij}\langle a_{ij}\rangle_a = \frac{1}{N}\sum_{ij}[1 - \langle\delta_{c_{ij},0}\rangle] = \alpha\langle q^2\rangle + \mathcal{O}(N^{-1}) \\ &= \langle k\rangle + \mathcal{O}(N^{-1}), \end{aligned} \quad (5.31)$$

$$\begin{aligned} \langle k^2\rangle_a &= \frac{1}{N}\sum_{i\neq j\neq k}\langle a_{ij}a_{jk}\rangle_a = \alpha\langle q^2\rangle + \alpha\langle q^3\rangle + \alpha^2\langle q^2\rangle^2 + \mathcal{O}(N^{-1}) \\ &= \langle k^2\rangle + \mathcal{O}(N^{-1}), \end{aligned} \quad (5.32)$$

$$m_3^a = \frac{1}{N}\sum_{i\neq j\neq k(\neq i)} \langle a_{ij}a_{jk}a_{ki}\rangle_a = \alpha\langle q^3\rangle + \mathcal{O}(N^{-1})$$
$$= m_3 + \mathcal{O}(N^{-1}) \tag{5.33}$$

and
$$m_4^a = \frac{1}{N}\sum_{[i,j,k,\ell]} \langle a_{ij}a_{jk}a_{k\ell}a_{\ell i}\rangle_a = \alpha\langle q^4\rangle + \mathcal{O}(N^{-1})$$
$$= m_4 + \mathcal{O}(N^{-1}). \tag{5.34}$$

Square brackets underneath summations again indicate distinct indices, which excludes backtracking in the counting of length-4 loops. Apparently, the ensembles $p(\boldsymbol{a})$ and $p(\boldsymbol{c})$ are asymptotically equivalent with regard to the statistics of these four quantities. We will see in the next section that this equivalence holds also for the 'dual' ensemble (4.2). To test the above claims, we compute and show in figure 6 the above observables in synthetic graphs $\boldsymbol{c}$ and $\boldsymbol{a}$ generated randomly from (4.4) and (4.5), where the random bipartite interaction graph $\boldsymbol{\xi}$ is drawn from (4.1).

6. Network properties generated by the *d*-ensemble

In this section, we will derive properties for the network ensembles (4.4) and (4.5) upon assuming that the statistics of the underlying bipartite PIN are given by (4.2), i.e. are protein-driven as opposed to complex-driven. Despite the superficial similarity between definitions (4.1) and (4.2), the expectations of graph observables in the two ensembles are found to be remarkably different.

6.1. Link probabilities

We start by calculating the link expectation values in the weighted graphs $c_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$:

$$\langle c_{ij}\rangle = \sum_\mu \langle \xi_i^\mu \xi_j^\mu\rangle = \frac{d_i d_j}{\alpha N}. \tag{6.1}$$

Hence the random graphs $\boldsymbol{c}$ are again finitely connected, now with

$$\langle k\rangle = \frac{1}{N}\sum_{ij}\langle c_{ij}\rangle = \frac{\langle d\rangle^2}{\alpha}. \tag{6.2}$$

Averages over d refer to the distribution $P(d)$ of protein promiscuities in the bipartite graph $\boldsymbol{\xi}$. The result (6.2) can also be written as $\langle k\rangle = \alpha\langle q\rangle^2$, and is thus notably different from the earlier expression $\langle k\rangle = \alpha\langle q^2\rangle$ found in the q-ensemble. The link likelihood is calculated in appendix B, and shows again that $p(c_{ij} > 1) = \mathcal{O}(N^{-2})$.

6.2. Densities of short loops

We can calculate the density of length-3 loops similar to how this was done for the q-ensemble in the previous section. Again these are given, to order $\mathcal{O}(1)$, by the S_3 stars in the bipartite graph, as the contribution from combinations of S_2 stars is as before $\mathcal{O}(N^{-1})$. Here we obtain

$$m_3 = \frac{1}{N}\sum_{[ijk]}\sum_\mu \langle \xi_i^\mu \xi_j^\mu \xi_k^\mu\rangle = \frac{1}{N}\sum_{[ijk]}\sum_\mu \frac{d_i d_j d_k}{\alpha^3 N^3} = \frac{\langle d\rangle^3}{\alpha^2}. \tag{6.3}$$

For loops of arbitrary length L, this generalizes to

$$m_L = \frac{\langle d\rangle^L}{\alpha^{L-1}}. \tag{6.4}$$

Interestingly, the densities m_L of short loops and the average connectivity $\langle k\rangle$ depend on $P(d)$ only through its first moment. Promiscuity heterogeneity apparently cannot affect the densities of short loops. In the present ensemble, these densities must therefore be identical to what would be found in a randomly wired bipartite graph. This prediction will be confirmed in simulations.

6.3. The degree distribution

In appendix C, we calculate the asymptotic degree distribution of $\boldsymbol{c}$ for the protein-driven complex recruitment model (4.2), giving

$$p(k) = \lim_{N\to\infty}\frac{1}{N}\sum_i \delta_{k,\sum_j c_{ij}}$$
$$= \sum_{d\geq 0} P(d)\sum_\ell \left(\frac{\mathrm{e}^{-d}d^\ell}{\ell!}\right)\left(\frac{\mathrm{e}^{-\ell\langle d\rangle/\alpha}(\ell\langle d\rangle/\alpha)^k}{k!}\right). \tag{6.5}$$

This result is again understood easily: the number of neighbours of a node i is a Poissonian variable ℓ, with average d, where d is now drawn from $P(d)$. Each of the ℓ first neighbours will have a degree which is a Poissonian variable with average $\langle d\rangle/\alpha$, so the number k of second neighbours of i in the bipartite graph is a Poisson variable with average $\ell\langle d\rangle/\alpha$. Equation (6.5) shows that a tail in the promiscuity distribution $P(d)$ will induce a tail in the degree distribution $p(k)$ of $\boldsymbol{c}$. The link between the two distributions is again most easily expressed via generating functions. Upon defining $Q_1(z) = \sum_k p(k)\mathrm{e}^{-zk}$ and $Q_4(z) = \sum_d P(d)\mathrm{e}^{-zd}$, we obtain from (6.5)

$$Q_1(z) = \sum_{d\geq 0} P(d)\mathrm{e}^{-pd}\frac{\sum_\ell (d\mathrm{e}^{\langle d\rangle(\mathrm{e}^{-z}-1)/\alpha})^\ell}{\ell!}$$
$$= Q_4(1-\mathrm{e}^{\langle d\rangle(\mathrm{e}^{-z}-1)/\alpha}). \tag{6.6}$$

For $z \simeq 0$ this gives

$$Q_1(z) \simeq Q_4\left(\frac{z\langle d\rangle}{\alpha}\right). \tag{6.7}$$

Hence, if $p(k)$ decays for large k as $p(k) \simeq Ck^{-\mu}$ with $2 < \mu < 3$, then via (5.21) we infer that

$$Q_4\left(\frac{z\langle d\rangle}{\alpha}\right) \simeq 1 - \langle k\rangle z + C\Gamma(1-\mu)z^{\mu-1}. \tag{6.8}$$

Equivalently,

$$Q_4(x) \simeq 1 - \frac{\alpha\langle k\rangle x}{\langle d\rangle} + C\Gamma(1-\mu)\left(\frac{\alpha}{\langle d\rangle}\right)^{\mu-1}x^{\mu-1}. \tag{6.9}$$

This implies that for large d the promiscuity distribution will be of the form $P(d) \simeq C'd^{-\mu}$, where

$$C' = C\left(\frac{\alpha}{\langle d\rangle}\right)^{\mu-1} = C\langle q\rangle^{1-\mu}. \tag{6.10}$$

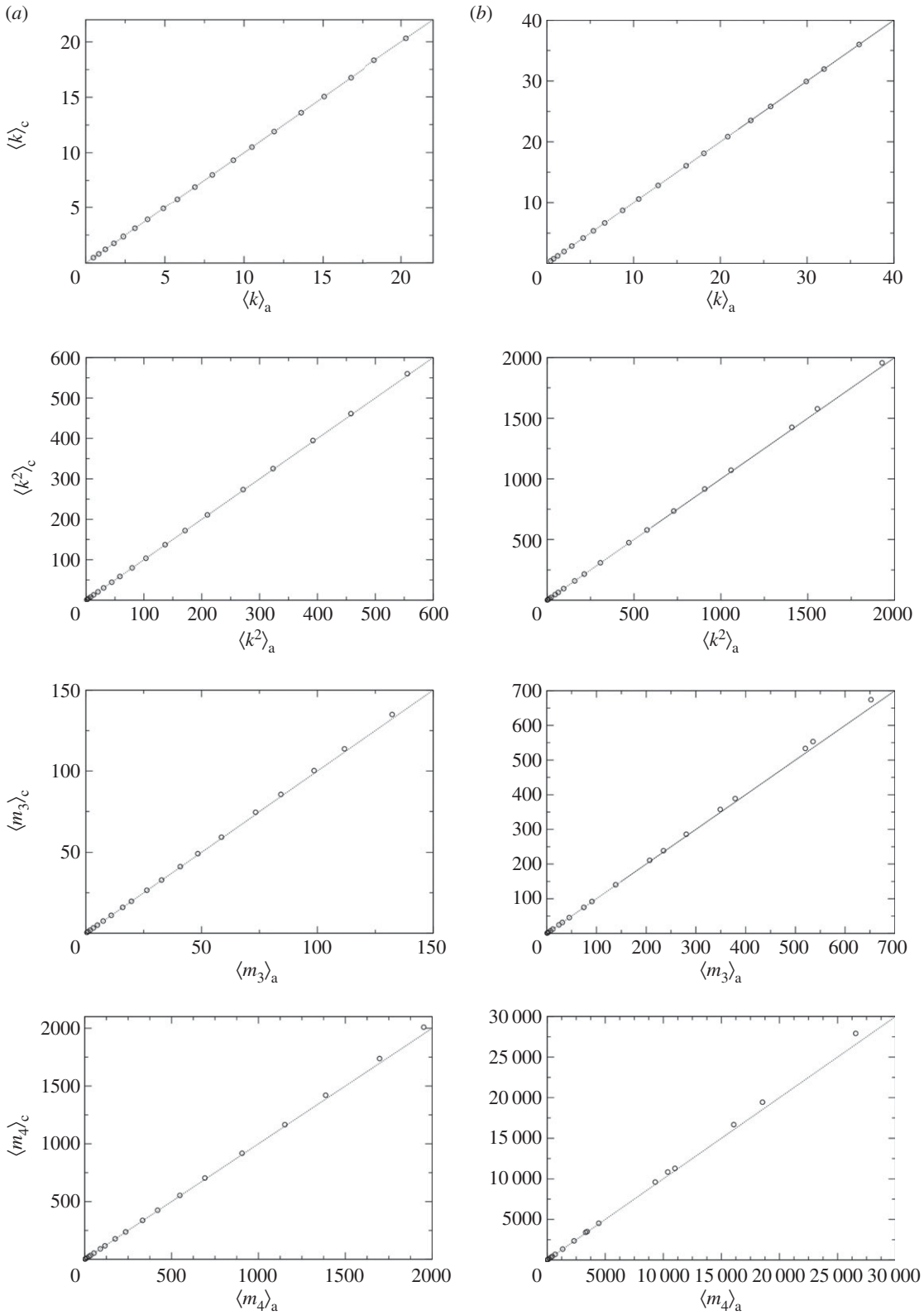

Figure 6. Symbols: $\langle k\rangle$, $\langle k^2\rangle$, m_3 and m_4 as measured in synthetic graphs ***c*** drawn from (4.5) with $N = 3000$, shown versus corresponding values found in the binary graphs ***a*** drawn from (4.4). Bipartite interaction graphs $\boldsymbol{\xi}$ are drawn from (4.1), with complex size distributions $P(q)$ that are Poissonian (*a*) or power law (*b*). Dotted lines: the diagonals (shown as guides to the eye). As expected, the values measured in the weighted graphs ***c*** are consistently higher than in the binary ones, but one finds that these deviations get smaller for increasing network sizes N.

6.4. Relationships that are independent of $P(d)$ and α

The first two moments of the degree distribution $p(k)$ of the separable PIN networks ***c*** are

$$\langle k\rangle = \sum_k kp(k) = \sum_d P(d) \sum_\ell \mathrm{e}^{-d} \frac{d^\ell}{\ell!} \frac{\ell\langle d\rangle}{\alpha} = \frac{\langle d\rangle^2}{\alpha} \tag{6.11}$$

and

$$\langle k^2\rangle = \sum_k k^2 p(k) = \sum_d P(d) \sum_\ell \mathrm{e}^{-d} \frac{d^\ell}{\ell!} \left[\left(\frac{\ell\langle d\rangle}{\alpha}\right)^2 + \frac{\ell\langle d\rangle}{\alpha}\right]$$
$$= \frac{\langle d\rangle^2}{\alpha} + \frac{\langle d\rangle^3}{\alpha^2} + \frac{\langle d\rangle^2\langle d^2\rangle}{\alpha^2}. \tag{6.12}$$

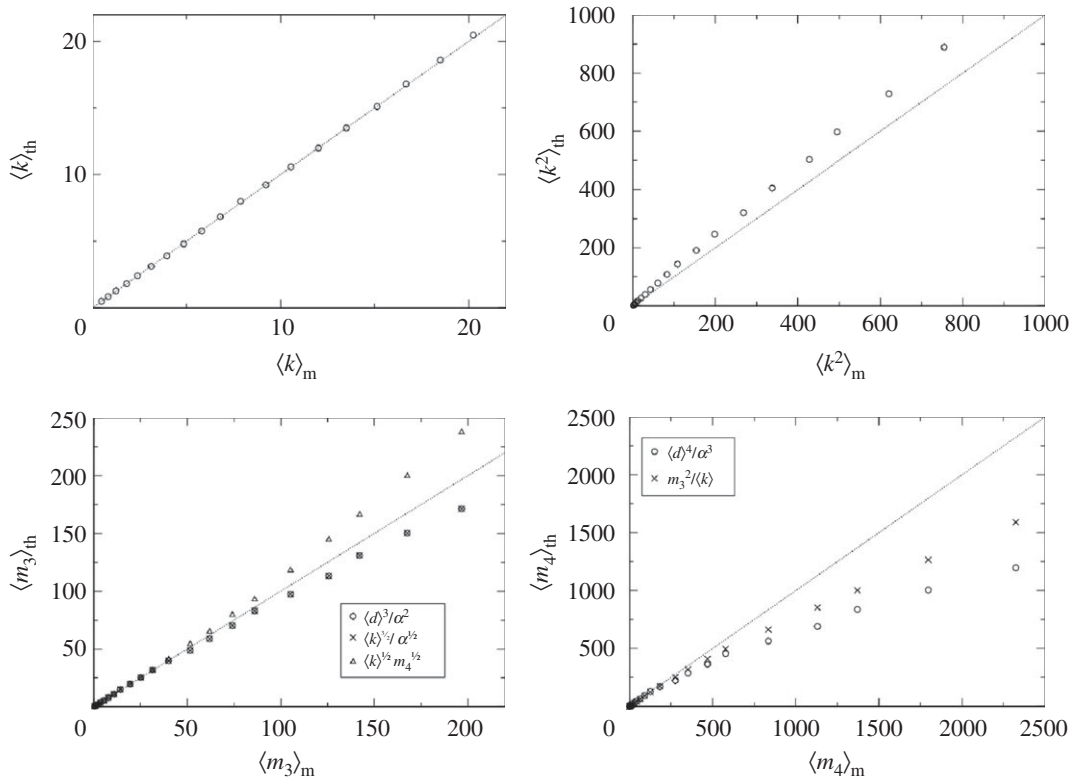

Figure 7. Symbols: theoretical $\langle\ldots\rangle_{\rm th}$ versus measured $\langle\ldots\rangle_{\rm m}$ values of observables $\langle k\rangle$, $\langle k^2\rangle$, m_3 and m_4 in synthetic random graphs $\boldsymbol{c}$ with $N = 3000$, defined via (4.1), (4.5) for a power-law distributed promiscuity distribution $P(d)$. Theoretical values are given by formulae (6.18) for $\langle k\rangle$, (6.12) for $\langle k^2\rangle$, (6.3), (6.14) and (6.21) for m_3 and (6.4) and (6.21) for m_4. Dotted lines: the diagonals (shown as guides to the eye).

Combination of (6.18), (6.12) and (6.13) now yields the relationship

$$\frac{\langle d^2\rangle}{\alpha} = \frac{\langle k^2\rangle - \langle k\rangle - m_3}{\langle k\rangle}, \tag{6.13}$$

which still involves $\langle d^2\rangle$ and α. We can also find an alternative expression for the density of loops of length 3 by combining (6.18) and (6.13)

$$m_3 = \frac{\langle k\rangle^{3/2}}{\sqrt{\alpha}}. \tag{6.14}$$

Unfortunately, neither of our two expressions for m_3, (6.13) nor (6.14), are useful, because the protein promiscuities distribution $P(d)$ and the ratio α are generally unknown. Access to information on these quantities via future detection experiments may therefore be extremely welcome in support of theoretical modelling of protein interaction datasets. To make progress, we need to derive relationships for graph observables that are independent of α and $P(d)$. We note that (6.14) yields

$$\forall L \geq 3: \quad \frac{m_{L+1}}{m_L} = \frac{\langle d\rangle}{\alpha}. \tag{6.15}$$

This can be rewritten using (6.18) as

$$\forall L \geq 3: \quad \frac{m_{L+1}}{m_L} = \sqrt{\frac{\langle k\rangle}{\alpha}}. \tag{6.16}$$

On the other hand, we know from (6.14) that $m_3/\langle k\rangle = \sqrt{\langle k\rangle/\alpha}$. Combining the above formulae allows us to establish the following relationship, that now is completely independent of $P(d)$ and α:

$$m_4 = \frac{m_3^2}{\langle k\rangle}. \tag{6.17}$$

Again we have tested the various formulae in synthetically generated graphs (figure 7).

6.5. Link between *a* and *c* graph definitions

As a final step, we check whether the observables m_3 and m_4 are indeed the same for the two PIN definitions (4.4) and (4.5), with the bipartite graph of our protein-driven ensemble (4.2), as protein detection experiments provide the binary matrix $\boldsymbol{a}$ as opposed to the weighted graph $\boldsymbol{c}$ for which (6.21) was derived. Again we denote averages relating to $\boldsymbol{a}$ as $\langle\ldots\rangle_a$, and those relating to $\boldsymbol{c}$ as $\langle\ldots\rangle$. For the moments of the degree distributions, we find the differences to be negligible:

$$\langle k\rangle_a = \frac{1}{N}\sum_{ij}\langle a_{ij}\rangle_a = \frac{\langle d\rangle^2}{\alpha} + \mathcal{O}(N^{-1}) = \langle k\rangle + \mathcal{O}(N^{-1}) \tag{6.18}$$

and

$$\begin{aligned}\langle k^2\rangle_a &= \frac{1}{N}\sum_{i\neq j\neq k}\langle a_{ij}a_{jk}\rangle_a \\ &= \frac{\langle d\rangle^2}{\alpha} + \frac{\langle d\rangle^3}{\alpha^2} + \frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2} + \mathcal{O}(N^{-1}) \\ &= \langle k^2\rangle + \mathcal{O}(N^{-1}).\end{aligned} \tag{6.19}$$

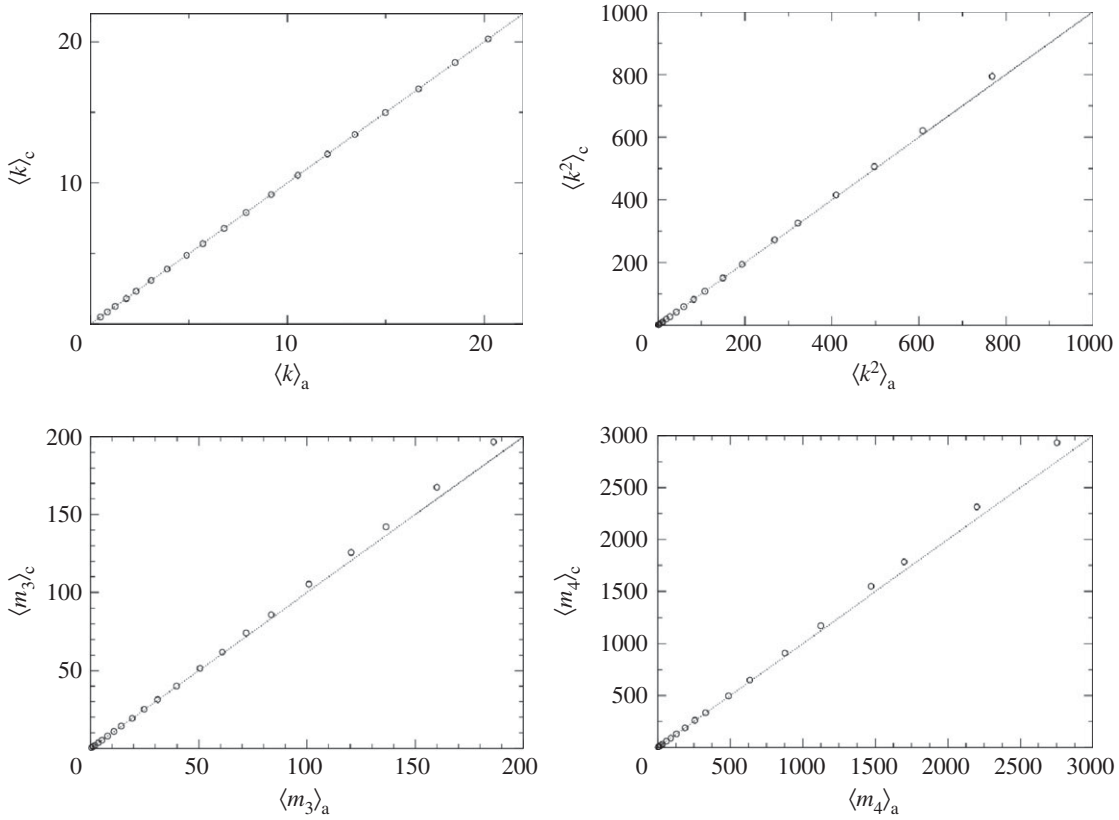

Figure 8. Symbols: $\langle k\rangle$, $\langle k^2\rangle$, m_3 and m_4 as measured in synthetic graphs $\boldsymbol{c}$ drawn from (4.5) with $N = 3000$, shown versus corresponding values found in the binary graphs $\boldsymbol{a}$ drawn from (4.4). Bipartite interaction graphs $\boldsymbol{\xi}$ are drawn from (4.2), with protein promiscuity distributions $P(d)$ that have a power-law form. Dotted line: the diagonals (shown as guides to the eye). As expected, the values measured in the weighted graphs $\boldsymbol{c}$ are consistently higher than in the binary ones, but these deviations get smaller for increasing network sizes N.

The same is true for the densities of loops of length 3 and 4:

$$m_3^a = \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle a_{ij}a_{jk}a_{ki}\rangle = \frac{\langle d\rangle^3}{\alpha^2} + \mathcal{O}(N^{-1}) = m_3 + \mathcal{O}(N^{-1}) \tag{6.20}$$

and

$$m_4^a = \frac{1}{N}\sum_{[i,j,k,\ell]}\langle a_{ij}a_{jk}a_{k\ell}a_{\ell i}\rangle = \frac{\langle d\rangle^4}{\alpha^3} + \mathcal{O}(N^{-1}) = m_4 + \mathcal{O}(N^{-1}). \tag{6.21}$$

This equivalence between the ensembles $p(\boldsymbol{a})$ and $p(\boldsymbol{c})$ when calculating the main average values of graph observables for large N implies that large protein interaction adjacency matrices can in practice be regarded as having a separable structure. Again, we check our relationships (6.12), (6.18), (6.20) and (6.21), against synthetically generated graphs and show results in figure 8.

7. Macroscopic observables in the mixed ensemble

The two bipartite graph ensembles (4.1) and (4.22) considered so far led to Poissonian distributions either for the protein promiscuities d_i (in the q-ensemble) or for the complex sizes q_μ (in the d-ensemble). It is possible to model heterogeneity in both d_i and q_μ using the mixed ensemble (4.3). Owing to the similarities with previous calculations we can and will be more brief in this section. For ensemble (4.3), the expectation values of individual links in the weighted graph $\boldsymbol{c}$ are

$$\langle c_{ij}\rangle = \sum_\mu \langle \xi_i^\mu \xi_j^\mu\rangle = \sum_\mu \frac{d_i d_j q_\mu^2}{\alpha^2\langle q\rangle^2 N^2} = \frac{d_i d_j\langle q^2\rangle}{\alpha\langle q\rangle^2 N} + \mathcal{O}(N^{-3/2}), \tag{7.1}$$

and the average connectivity follows as

$$\langle k\rangle = \frac{1}{N}\sum_{ij}\langle c_{ij}\rangle = \frac{\langle d\rangle^2\langle q^2\rangle}{\alpha\langle q\rangle^2} + \mathcal{O}(N^{-1/2}) = \alpha\langle q^2\rangle + \mathcal{O}(N^{-1/2}). \tag{7.2}$$

Full details are found in appendix B. As in previous ensembles, the leading contribution to the density of length-3 loops comes from the S_3 stars in the bipartite graphs, now giving

$$m_3 = \frac{1}{N}\sum_{[ijk]}\sum_\mu \langle \xi_i^\mu \xi_j^\mu \xi_k^\mu\rangle = \frac{1}{N}\sum_{[ijk]}\sum_\mu \frac{d_i d_j d_k q_\mu^3}{\alpha^3\langle q\rangle^3 N^3} \simeq \frac{\langle d\rangle^3\langle q^3\rangle}{\alpha^2\langle q\rangle^3} = \alpha\langle q^3\rangle. \tag{7.3}$$

As before, the heterogeneity in the d affects neither the average connectivity $\langle k\rangle$ nor the density of triangles m_3, both are as they were in the q-ensemble. This is confirmed numerically (figure 9). The degree distribution for large N

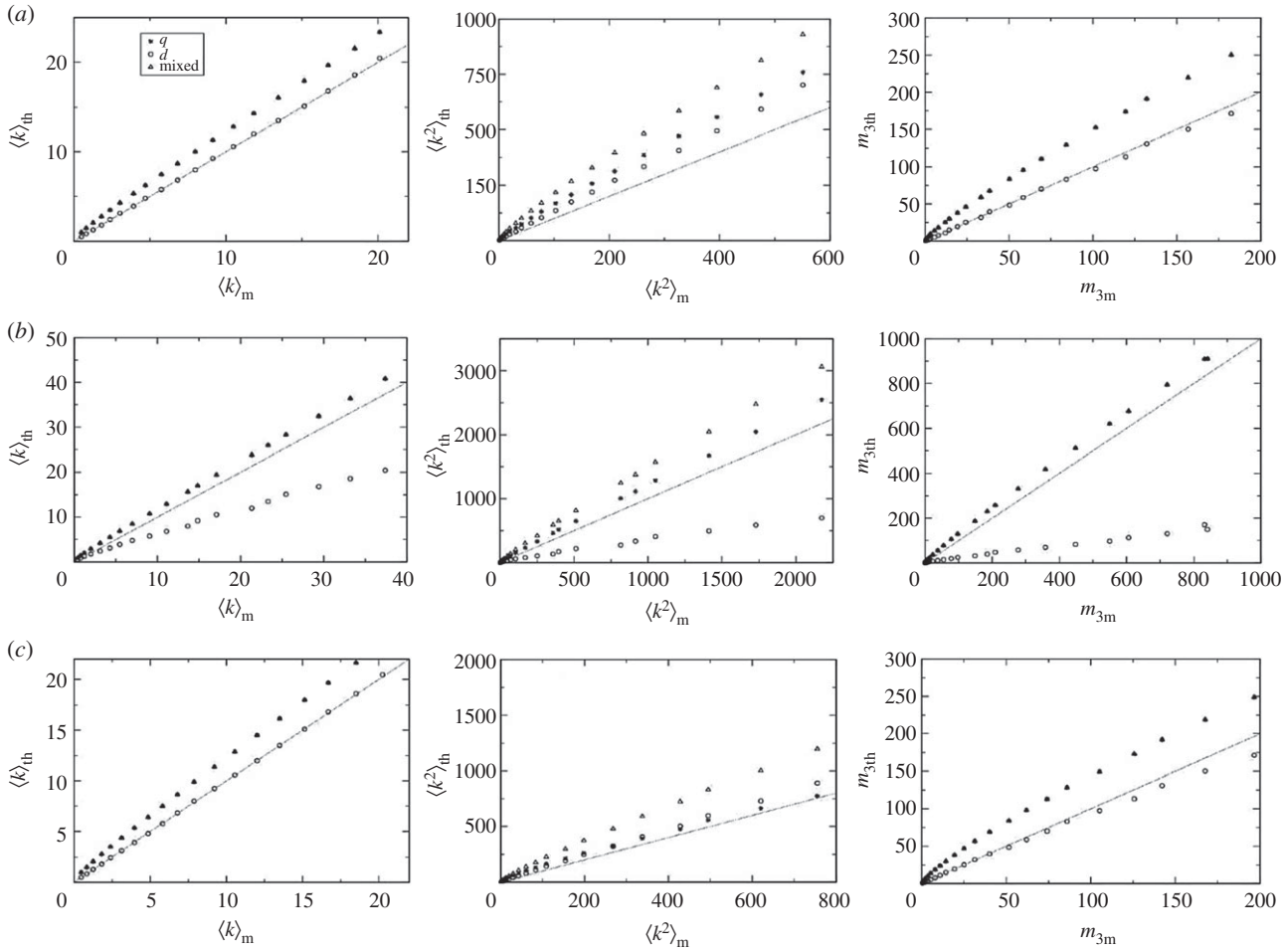

Figure 9. Symbols: theoretical $\langle \ldots \rangle_{\mathrm{th}}$ versus measured $\langle \ldots \rangle_{\mathrm{m}}$ values of observables $\langle k \rangle$, $\langle k^2 \rangle$, and m_3 in synthetic random graphs ***a*** with $N = 3000$ and $\alpha = 0.5$, generated either via random wiring (*a*), *q*-preferential attachment (*b*) or *d*-preferential attachment (*c*). Dotted lines: the diagonals (shown as guides to the eye).

in the ensemble $p(\boldsymbol{c})$ is calculated in appendix C, giving

$$p(k) = \frac{\int_0^\infty \mathrm{d}y\, P(y)\, \mathrm{e}^{-y} y^k}{k!}, \tag{7.4}$$

where

$$P(y) = \sum_d P(d) \mathrm{e}^{-d} \sum_{\ell \geq 0} \frac{d^\ell}{\ell!} \sum_{q_1 \ldots q_\ell \geq 0} W(q_1) \ldots W(q_\ell)\, \delta\left[y - \sum_{r \leq \ell} q_r\right]. \tag{7.5}$$

Again it is possible to relate the asymptotic behaviour of $p(k)$ to that of $P(d)$ and $W(q)$, by inspecting the relationship between the relevant generating functions. Using our previous definitions for $Q_1(z)$, $Q_2(z)$, $Q_3(z)$, and $Q_4(z)$, we obtain via (7.4) and (7.5)

$$\begin{aligned} Q_1(z) &= \int \mathrm{d}y P(y) \sum_k \frac{\mathrm{e}^{-y}(y\mathrm{e}^{-z})^k}{k!} = \int \mathrm{d}y P(y) \mathrm{e}^{-y(1-\mathrm{e}^{-z})} \\ &= Q_2(1 - \mathrm{e}^{-z}) \end{aligned} \tag{7.6}$$

and

$$\begin{aligned} Q_2(z) &= \sum_d P(d) \mathrm{e}^{-d} \sum_\ell \frac{d^\ell}{\ell!} \prod_{r=1}^{\ell} \left(\sum_{q_r} W(q_r) \mathrm{e}^{-z q_r} \right) \\ &= \sum_d P(d) \mathrm{e}^{-d} \sum_\ell \frac{d^\ell}{\ell!} Q_3^\ell(z) \\ &= \sum_d P(d) \mathrm{e}^{-d[1-Q_3(z)]} = Q_4(1 - Q_3(z)). \end{aligned} \tag{7.7}$$

Expanding (7.6) for small z tells us that $Q_1(z) \simeq Q_2(z)$. Substitution into (7.7) subsequently gives

$$Q_1(z) \simeq Q_4(1 - Q_3(z)). \tag{7.8}$$

Assuming $W(q)$ to have a power-law tail, but with a finite first moment (as in all cases previously considered), i.e. $W(q) \simeq K q^{-\gamma}$ with $\gamma > 2$, its generating function $Q_3(z)$, can be written as

$$Q_3(z) = 1 - \frac{\langle q^2 \rangle z}{\langle q \rangle} + \mathcal{O}(z^\delta), \tag{7.9}$$

where $\delta = \min\{2, \gamma - 1\}$. Insertion into (7.8) then leads to

$$Q_1(z) \simeq Q_4\left(\frac{z\langle q^2 \rangle}{\langle q \rangle} - \mathcal{O}(z^\delta) \right). \tag{7.10}$$

If $p(k) = C k^{-\mu}$, with $2 < \mu < 3$, we may use our earlier result (5.21) and get

$$Q_4\left(x - \mathcal{O}\left(\frac{x\langle q \rangle}{\langle q^2 \rangle} \right)^\delta \right) \simeq 1 - \frac{\langle k \rangle \langle q \rangle x}{\langle q^2 \rangle} + C\Gamma(1-\mu) \left(\frac{\langle q \rangle}{\langle q^2 \rangle} \right)^{\mu - 1} x^{\mu - 1}. \tag{7.11}$$

If $\gamma > \mu$ we have $\delta > \mu - 1$, so we can neglect the second term in the argument of Q_4, and conclude that the promiscuity distribution has the asymptotic form $P(d) = C' d^{-\mu}$ where $C' = C(\langle q^2 \rangle / \langle q \rangle)^{1-\mu}$. This means that if $W(q)$ decays faster than $p(k)$ (as in §6), then the tail in $p(k)$ must arise from the tail in $P(d)$. Note, however, that heterogeneities in

$P(q)$ will affect the amplitude of the power-law tail in $P(d)$, which will be smaller by a factor $(\langle q^2\rangle/\langle q\rangle^2)^{1-\mu}$ compared with the case where $P(q)=\delta_{q,\langle q\rangle}$, where we had $C'=C\langle q\rangle^{1-\mu}$. Conversely, if $\gamma=\mu$ we have $\delta=\mu-1$, and writing the $\mathcal{O}(z^\delta)$ term explicitly in (7.10) gives

$$Q_4(z\langle q^2\rangle/\langle q\rangle - K\Gamma(1-\mu)z^{\mu-1}) = 1-\langle k\rangle z + C\Gamma(1-\mu)z^{\mu-1}. \tag{7.12}$$

Expanding both sides in powers of z and equating prefactors tells us that either $C'=0$ and $C=K\langle d\rangle$ (i.e. $K=C/\alpha\langle q\rangle$, which retrieves the case in §5), or $\delta=\mu$ with $K\langle d\rangle + C'(\langle q^2\rangle/\langle q\rangle)^{\mu-1}=C$. Hence, if $P(d)$ is as broad as $W(q)$, then both contribute to the tail in $p(k)$, whose amplitude will be the sum of the amplitudes of the tails in $P(q)$ and $P(d)$. We see in (7.12) that $\gamma<\mu$ is not possible, i.e. $W(q)$ needs to decay at least as fast as $p(k)$.

In appendix C, we calculate the first two moments of the degree distribution $p(k)$ of the ensemble $p(\boldsymbol{c})$. This recovers (7.2) for the first moment, and for the second moment gives

$$\langle k^2\rangle = \frac{\alpha\langle q^2\rangle + \alpha\langle q^3\rangle + \langle d^2\rangle\langle k\rangle^2}{\langle d\rangle^2}. \tag{7.13}$$

Substituting (7.2) and (7.3) into (7.13) then leads to

$$m_3 = \frac{\langle k^2\rangle - \langle k\rangle - \langle k\rangle^2\langle d^2\rangle}{\langle d\rangle^2}. \tag{7.14}$$

The density of length-3 loops depends again on the first two moments of the degree distribution $p(k)$, but is also seen to depend on the first two moments of the promiscuity distribution $P(d)$, which is unknown. Hence, this relationship cannot serve as a test of PIN data quality. It is nevertheless useful for comparing the mixed ensemble with the d- and the q-ensembles in synthetically generated data.

8. Numerical comparison of the three bipartite generative ensembles

Here we compare the ability of our bipartite ensembles (4.1)–(4.4) to predict properties of the associated binary PIN graphs, for synthetic networks that are generated from any of these ensembles. We focus on comparing homologous formulae for the observables $\langle k\rangle$, $\langle k^2\rangle$, m_3 and m_4. The synthetic matrices $\boldsymbol{a}=\{a_{ij}\}$ with $a_{ij}\in\{0,1\}$ are defined as before via $a_{ij}=\theta(\sum_\mu \xi_i^\mu \xi_j^\mu)$, with $\theta(0)=0$, and the links of the bipartite graph $\boldsymbol{\xi}$ are generated from the following three protocols. In the first protocol, links between nodes (i,μ) are drawn randomly and independently, until their total number reaches a prescribed limit. In the second protocol, we assign the links preferentially to complexes with large sizes. In the third protocol, we assign links preferentially to proteins with large promiscuities.

In figure 9, we show along the vertical axes the values of $\langle k\rangle$ (left) predicted by the three ensembles, via formulae (5.26), (6.2) and (7.2), the predicted values of $\langle k^2\rangle$ (middle), via (5.27), (6.12) and (7.13), and the predicted triangle density m_3 (right), via (5.29), (6.13) and (7.14). All are shown together with the corresponding values that were measured in $\boldsymbol{a}$, along the horizontal axes. As expected, the d-ensemble outperforms the other ensembles when links are drawn according to d-preferential attachment, whereas the q-ensemble performs better for graphs generated via q-preferential attachment. The mixed ensemble performs very similar to the q-ensemble in terms of counting triangles, as expected from the reasoning in §7. Deviations between the q and the mixed ensembles are most evident in the second moment of the degree distribution, where the mixed ensemble always leads to values well above those of the q- and the d-ensembles. We found in §6 that the d-ensemble is indistinguishable from a fully random ensemble when calculating $\langle k\rangle$ and m_3, which explains why the d-ensemble predicts the values of these two observables perfectly. The other two ensembles are more sensitive to finite size effects, as any heterogeneity in the q will boost the number of loops.

In figure 10, we show the values of m_3 and m_4 predicted by those formulae that involve only measurable graph observables, for the synthetically generated graphs used in figure 9. The prediction of m_3 is now obtained from (5.29) and (6.21), for the q- and d-ensembles, respectively, and m_4 is evaluated using (5.30) and (6.21). In figure 11, we plot the degree distribution $p(k)$ of graphs with identical values for the number of nodes ($N=3000$) and the number of links $L=N\alpha\langle q\rangle$, generated synthetically via the three chosen protocols, together with the distributions $P(q)$ of complex sizes and $P(d)$ of protein promiscuities. As explained in §7, tails in the degree distribution $p(k)\sim k^{-\mu}$ can arise either from a complex size distribution $P(q)\sim q^{-\mu-1}$ and a homogeneous promiscuity distribution, or from having an equally fat tail in the promiscuity distribution $P(d)\sim d^{-\mu}$ together with less heterogeneous complex sizes $P(q)\sim q^{-\gamma-1}$ with $\gamma>\mu$.

9. Conclusion

In this paper, we propose a bipartite network representation of protein interactions, where the two node types represent proteins and complexes. A protein–protein interaction network can then be regarded as the result of a 'marginalization' of the bipartite network, whereby the complexes are integrated out (i.e. summed over). This leads to a weighted PIN $\boldsymbol{c}$ with a separable structure. Adjacency matrices of PINs $\boldsymbol{a}$ are then simply the binary versions of the separable $\boldsymbol{c}$, obtained by the entry truncations $a_{ij}=\theta(c_{ij})$, with the convention $\theta(0)=0$. One of the central results of this work is that for sufficiently large networks there is an equivalence between the two graph ensembles $p(\boldsymbol{c})$ and $p(\boldsymbol{a})$, inasmuch as macroscopic statistical properties are concerned, such as densities of short loops and degree distributions. This allows us to regard the conventional protein interaction adjacency matrices as if they were to have a separable structure, and induces precise relationships between expectation values of macroscopic graph observables which, remarkably, only depend on measurable quantities and on the underlying mechanism with which proteins and complexes recruit each other. They are independent of inaccessible microscopic details of proteins and their complexes.

We considered the two extreme complex recruitment scenarios, one where recruitment is driven solely by protein promiscuities, and another where it is driven by complex sizes. Preferential attachment to large complexes (the q-ensemble) favours the presence of large cliques in PINs, which boosts the number of short loops. Hence we can reasonably expect that the predictions on short loop densities from the

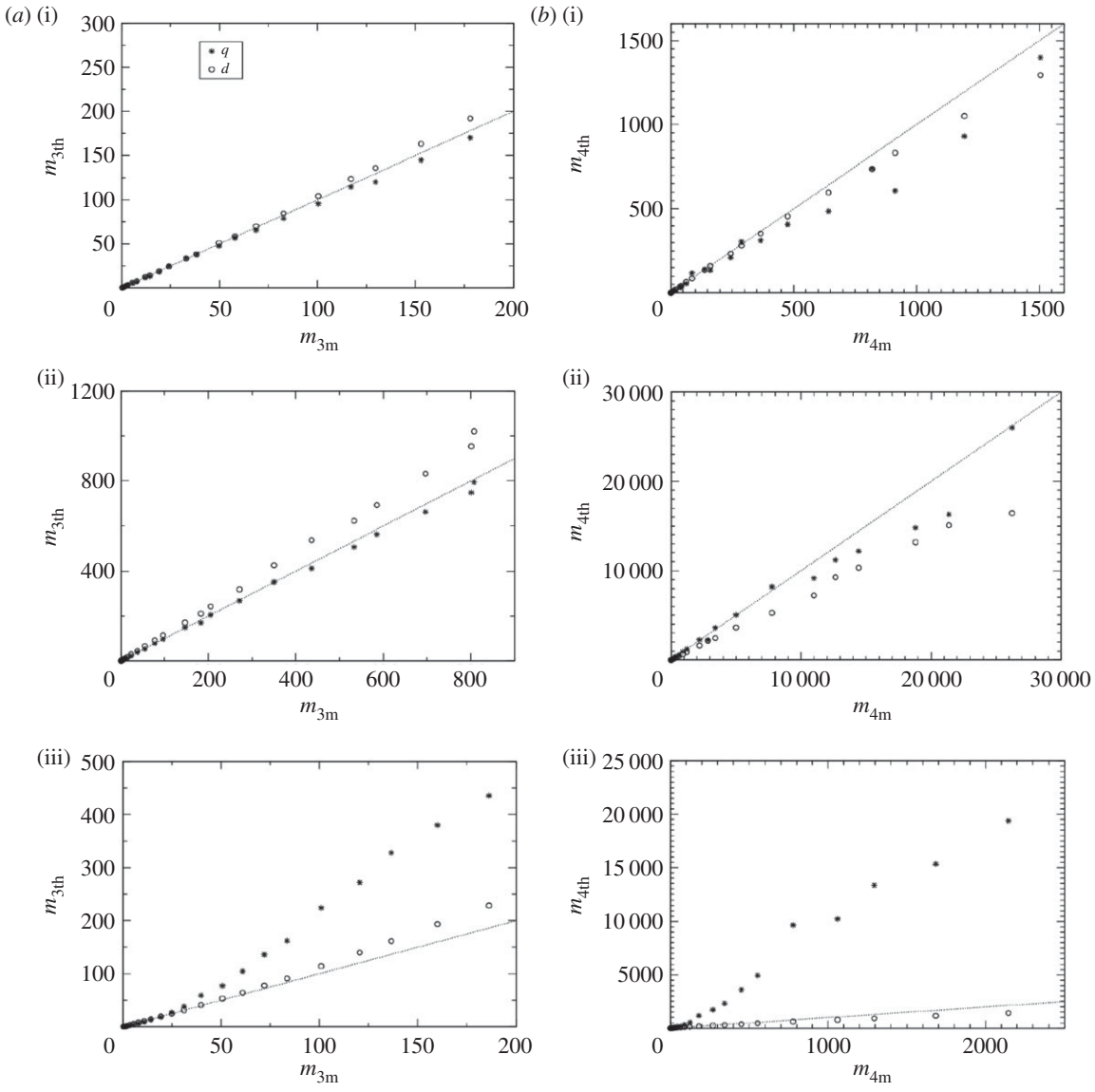

Figure 10. Predicted versus real m_3 (*a*) and m_4 (*b*) for random bipartite graphs with $N = 3000$ and $\alpha = 0.5$ generated via random wiring (*a*(i),*b*(i)), *q*-preferential (*a*(ii),*b*(ii)) and *d*-preferential (*a*(iii),*b*(iii)), calculated by using formulae (5.29), (5.30), (6.21) and observables appearing in the formulae computed directly from the network.

q-ensemble will overestimate the real number of loops. Conversely, preferential attachment based only on protein promiscuities (the *d*-ensemble) leads to homogeneous complex sizes, which suppresses large cliques in PINs, leading to an underestimation of short loop densities. Remarkably, real protein interaction data from MS and Y2H experiments show a density of length-4 loops in between the predictions of the *d*-ensemble and those of the *q*-ensemble, suggesting a degree of compatibility of these experimental data with a separable structure of the proteome. By contrast, both MS and Y2H datasets show densities or length-3 loops that are consistently smaller than all our theoretical predictions and closer to expectation in random graphs with identical degree distributions, suggesting the presence of a noise level which randomizes interactions. We note that MS values generally show a higher degree of compatibility with a separable structure of the proteome than Y2H.

We believe that, by providing a systematic and practical framework for understanding protein interaction experiments, our approach may represent a valuable step towards establishing a more solid connection between protein interaction datasets and the underlying biology. Universal bounds on observables in PINs may become powerful tools for data quality testing. As future work, it would be useful to apply the present framework to datasets with different features, including ribosomal data, large-scale datasets resulting from the union of known datasets (e.g. [41]), more nuanced descriptions of PINs as those involving alternative splicing, as well as adapting the present framework to include multiple measurements from repeated experiments [42]. Improved versions of the present models, which fit the experimental data better, may open a route to infer quantities such as the ratio α, and the distributions of protein promiscuities and complex sizes. Such quantities are hardly available in the current PIN datasets, and are generally difficult to access experimentally. The present work has revealed that the asymptotic forms of these distributions can be extracted from the tails of the PIN degree distributions. Recent protein–complex datasets such as [43,44] may provide useful sources to test inference capabilities of the present framework. Finally, our method my shed some light on the way protein and complexes recruit one another, in particular, whether this recruitment is driven by

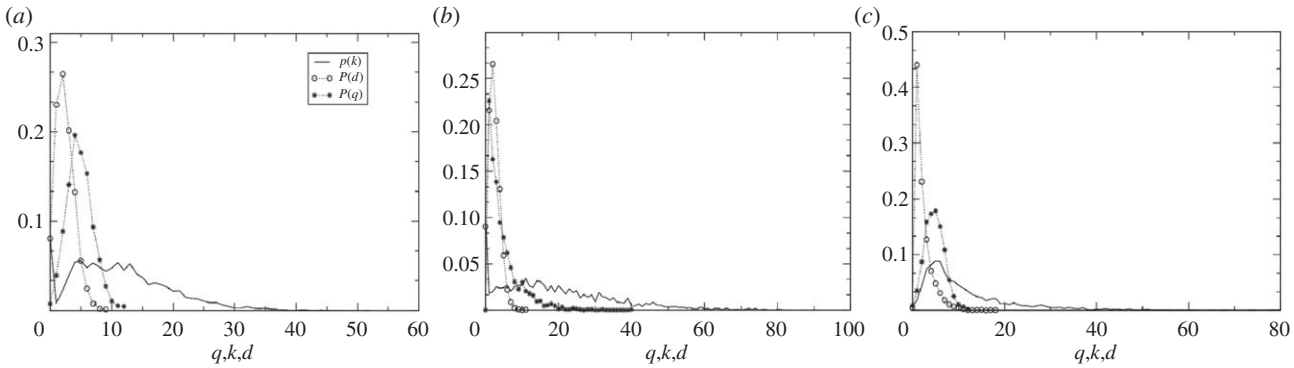

Figure 11. Distributions $P(q)$ of complex sizes, $P(d)$ or protein promiscuities and $p(k)$ of the degrees in ***a*** (distinguished by markers as shown in the legend), for random bipartite graphs with $N = 3000$, $\alpha = 0.5$ and $\langle q \rangle = 4.8$, which have been generated via random wiring (*a*), via *q*-preferential attachment (*b*) or via *d*-preferential attachment (*c*).

proteins or by complexes, and may enable us to discriminate between 'party hub' and 'date hub' interactions.

Competing interests. We declare we have no competing interests.

Funding. A.C.C.C. is grateful for support from the UK's Biotechnology and Biological Sciences Research Council (BBSRC).

Acknowledgements. A.A. acknowledges Alessandro Pandini and Sun Chung for providing protein interaction datasets. Kate Roberts is acknowledged for interesting discussions during the early stages of this work. Authors are grateful to all referees for many useful comments.

Appendix A. Promiscuities and complex size distributions in ensembles $P_q(\boldsymbol{\xi})$ and $P_d(\boldsymbol{\xi})$

For graphs drawn from ensemble (4.1), where complex sizes $\{q_\mu\}$ are drawn from a given distribution $P(q)$, the distribution of protein promiscuities is, for large N,

$$p(d) = \lim_{N\to\infty} \langle \delta_{d,\sum_\mu \xi_i^\mu} \rangle = \lim_{N\to\infty} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega d} \langle e^{-i\omega \sum_\mu \xi_i^\mu} \rangle$$
$$= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega d + \alpha\langle q\rangle (e^{-i\omega}-1)} = \frac{e^{-\alpha\langle q\rangle} (\alpha\langle q\rangle)^d}{d!}. \qquad (A\,1)$$

For graphs drawn from ensemble (4.2), where protein promiscuities $\{d_i\}$ are drawn from a given distribution $P(q)$, the distribution of complex sizes is, for large N,

$$p(q) = \lim_{N\to\infty} \langle \delta_{q,\sum_i \xi_i^\mu} \rangle = \lim_{N\to\infty} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega q} \langle e^{-i\omega \sum_i \xi_i^\mu} \rangle$$
$$= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega q + (\langle d\rangle/\alpha)(e^{-i\omega}-1)} = \frac{e^{-\langle d\rangle/\alpha} (\langle d\rangle/\alpha)^q}{q!}. \qquad (A\,2)$$

Appendix B. Link probabilities in the weighted protein interaction network

In this appendix, we derive the likelihood to have a link in the weighted PIN $c_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$, when the ξ_i^μ are drawn from the ensembles (4.1)–(4.3).

B.1. The *q*-ensemble

In the *q*-ensemble, we have

$$p(c_{ij}) = \left\langle \delta_{c_{ij},\sum_{\mu\le\alpha N} \xi_i^\mu \xi_j^\mu} \right\rangle_{\boldsymbol{\xi}} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \prod_{\mu=1}^{\alpha N} \langle e^{-i\omega \xi_i^\mu \xi_j^\mu} \rangle_{\boldsymbol{\xi}}$$
$$= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \prod_{\mu=1}^{\alpha N} \left\{ \frac{q_\mu^2}{N^2} e^{-i\omega} + \left(1 - \frac{q_\mu^2}{N^2}\right) \right\} = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij} + \sum_{\mu=1}^{\alpha N} (q_\mu^2/N^2)[e^{-i\omega}-1] - (1/2)\sum_{\mu=1}^{\alpha N} (q_\mu^4/N^4)[e^{-i\omega}-1]^2}$$
$$= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega c_{ij}} \left[1 + \frac{\alpha\langle q^2\rangle}{N} (e^{-i\omega}-1) - \frac{1}{2}\frac{\alpha\langle q^4\rangle}{N^3} (e^{-i\omega}-1)^2 + \frac{\alpha^2\langle q^2\rangle^2}{2N^2} (e^{-i\omega}-1)^2 \right.$$
$$\left. + \frac{\alpha^3\langle q^2\rangle^3}{6N^3} (e^{-i\omega}-1)^3 + \mathcal{O}(N^{-4}) \right]$$
$$= \delta_{c_{ij},0} + \frac{\alpha\langle q^2\rangle}{N} (\delta_{c_{ij},1} - \delta_{c_{ij},0}) + \left(\frac{\alpha^2\langle q^2\rangle^2}{2N^2} - \frac{1}{2}\frac{\alpha\langle q^4\rangle}{N^3} \right) (\delta_{c_{ij},2} - 2\delta_{c_{ij},1} + \delta_{c_{ij},0})$$
$$+ \frac{\alpha^3\langle q^2\rangle^3}{6N^3} (\delta_{c_{ij},3} - 3\delta_{c_{ij},2} + 3\delta_{c_{ij},1} - \delta_{c_{ij},0}) + \mathcal{O}(N^{-4}). \qquad (B\,1)$$

From this, one reads off directly the values of $p(c_{ij}=0)$, $p(c_{ij}=1)$ and $p(c_{ij}\geq 2)$. The density of triangles is obtained writing (B 2) as

$$m_3=(N-1)(N-2)\sum_{\mu\nu\rho=1}^{\alpha N}\langle\xi^\mu\xi^\nu\rangle\langle\xi^\nu\xi^\rho\rangle\langle\xi^\rho\xi^\mu\rangle, \tag{B 2}$$

and using

$$\begin{aligned}\langle\xi^\mu\xi^\nu\rangle&=\langle\xi^\mu\rangle\langle\xi^\nu\rangle+\delta_{\mu\nu}\langle\xi^\mu\rangle(1-\langle\xi^\mu\rangle)\\&=\frac{q_\mu q_\nu}{N^2}+\delta_{\mu\nu}\frac{q_\mu}{N}\left(1-\frac{q_\mu}{N}\right).\end{aligned} \tag{B 3}$$

This gives

$$\begin{aligned}m_3&=\frac{1}{N}\left[1+\mathcal{O}\left(\frac{1}{N}\right)\right]\sum_{\mu\nu\rho=1}^{\alpha N}q_\mu q_\nu q_\rho\left[\frac{q_\nu}{N}+\delta_{\mu\nu}\left(1-\frac{q_\nu}{N}\right)\right]\left[\frac{q_\rho}{N}+\delta_{\nu\rho}\left(1-\frac{q_\rho}{N}\right)\right]\left[\frac{q_\mu}{N}+\delta_{\rho\mu}\left(1-\frac{q_\mu}{N}\right)\right]\\&=\frac{1}{N}\left(1+\mathcal{O}\left(\frac{1}{N}\right)\right)\sum_{\mu\nu\rho=1}^{\alpha N}q_\mu q_\nu q_\rho\left\{\frac{q_\mu q_\nu q_\rho}{N^3}+3\delta_{\mu\nu}\frac{q_\rho q_\mu}{N^2}\left(1-\frac{q_\mu}{N}\right)+3\delta_{\mu\nu}\delta_{\nu\rho}\frac{q_\mu}{N}\left(1-\frac{q_\mu}{N}\right)^2+\delta_{\mu\nu}\delta_{\nu\rho}\delta_{\rho\mu}\left(1-\frac{q_\mu}{N}\right)^3\right\}\\&=\frac{1}{N}\sum_{\mu=1}^{\alpha N}\left(1-\frac{q_\mu}{N}\right)^3q_\mu^3+\mathcal{O}\left(\frac{1}{N}\right)=\alpha\langle q^3\rangle+\mathcal{O}(N^{-1}).\end{aligned} \tag{B 4}$$

B.2. The *d*-ensemble

In the *d*-ensemble, we obtain

$$\begin{aligned}p(c_{ij})&=\langle\delta_{c_{ij},\sum_\mu\xi_i^\mu\xi_j^\mu}\rangle=\int_{-\pi}^{\pi}\frac{d\omega}{2\pi}e^{i\omega c_{ij}+(d_id_j/\alpha N)(e^{-i\omega}1)-1/2(d_i^2d_j^2/(\alpha N)^3)(e^{-i\omega}-1)^2}\\&=\int_{-\pi}^{\pi}\frac{d\omega}{2\pi}e^{i\omega c_{ij}}\left[1+\frac{d_id_j}{\alpha N}(e^{-i\omega}-1)+\frac{1}{2}\left(\frac{d_id_j}{\alpha N}\right)^2(e^{-i\omega}-1)^2-\frac{1}{2}\frac{(d_id_j)^2}{(\alpha N)^3}(e^{-i\omega}-1)^2\right.\\&\qquad\left.+\frac{1}{6}\left(\frac{d_id_j}{\alpha N}\right)^3(e^{-i\omega}-1)^3+\cdots\right],\end{aligned} \tag{B 5}$$

which gives

$$\begin{aligned}p(c_{ij}=0)&=1-\frac{d_id_j}{\alpha N}+\frac{1}{2}\left(\frac{d_id_j}{\alpha N}\right)^2-\frac{1}{6}\left(\frac{d_id_j}{\alpha N}\right)^3-\frac{1}{2}\frac{d_i^2d_j^2}{(\alpha N)^3}\\p(c_{ij}=1)&=\frac{d_id_j}{\alpha N}-\left(\frac{d_id_j}{\alpha N}\right)^2+\frac{1}{2}\left(\frac{d_id_j}{\alpha N}\right)^3+\frac{d_i^2d_j^2}{(\alpha N)^3}\\p(c_{ij}\geq 2)&=\mathcal{O}(N^{-2}).\end{aligned} \tag{B 6}$$

B.3. The mixed ensemble

For the mixed ensemble, the link likelihood is found to be

$$\begin{aligned}p(c_{ij})&=\langle\delta_{c_{ij},\sum_\mu\xi_i^\mu\xi_j^\mu}\rangle=\int_{-\pi}^{\pi}\frac{d\omega}{2\pi}e^{i\omega c_{ij}+\sum_\mu(d_id_jq_\mu^2/\alpha^2\langle q\rangle^2N^2)(e^{-i\omega}-1)}=\int_{-\pi}^{\pi}\frac{d\omega}{2\pi}e^{i\omega c_{ij}+(d_id_j\langle q^2\rangle/\alpha\langle q\rangle^2N)(e^{-i\omega}-1)}\\&=e^{-d_id_j\langle q^2\rangle/\alpha\langle q\rangle^2N}\int_{-\pi}^{\pi}\frac{d\omega}{2\pi}e^{i\omega c_{ij}}\left[1+\frac{d_id_j\langle q^2\rangle}{\alpha N\langle q\rangle^2}e^{-i\omega}+\frac{1}{2}\left(\frac{d_id_j\langle q^2\rangle}{\alpha N\langle q\rangle^2}\right)^2e^{-2i\omega}+\ldots\right],\end{aligned} \tag{B 7}$$

giving

$$\begin{aligned}p(c_{ij}=0)&=1-\frac{d_id_j\langle q^2\rangle}{\alpha\langle q\rangle^2N}+\frac{1}{2}\left(\frac{d_id_j\langle q^2\rangle}{\alpha\langle q\rangle^2N}\right)^2+\mathcal{O}(N^{-3})\\p(c_{ij}=1)&=\frac{d_id_j\langle q^2\rangle}{\alpha\langle q\rangle^2N}\left(1-\frac{d_id_j\langle q^2\rangle}{\alpha\langle q\rangle^2N}\right)+\mathcal{O}(N^{-3})\\p(c_{ij}\geq 2)&=\mathcal{O}(N^{-2}).\end{aligned} \tag{B 8}$$

Appendix C. Calculation of the degree distribution $p(k)$

In this appendix, we calculate the degree distribution of the weighted PIN $c_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$, in which the entries ξ_i^μ are drawn from the bipartite ensembles (4.1)–(4.3).

C.1. The q-ensemble

In the q-ensemble, we can calculate $p(k)$ as follows:

$$
\begin{aligned}
p(k) &= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \left\langle \frac{1}{N} \sum_i e^{-i\omega \sum_j c_{ij}} \right\rangle_\xi = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \langle e^{-i\omega \sum_{j>1} c_{1j}} \rangle_\xi \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \langle e^{-i\omega \sum_\mu \xi_1^\mu \sum_{j>1} \xi_j^\mu} \rangle_\xi = \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_\mu \langle e^{-i\omega \xi_1^\mu \sum_{j>1} \xi_j^\mu} \rangle_\xi \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_\mu \left\{ 1 + \frac{q_\mu}{N} [\langle e^{-i\omega \xi^\mu} \rangle_{\xi^\mu}^{N-1} - 1] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \prod_\mu \left\{ 1 + \frac{q_\mu}{N} \left[\left(1 + \frac{q_\mu}{N} (e^{-i\omega} - 1) \right)^{N-1} - 1 \right] \right\} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \sum_\mu (q_\mu/N)[\exp[q_\mu (e^{-i\omega} - 1)] - 1] + \mathcal{O}(N^{-1})} \\
&= \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \alpha \langle q[\exp[q(e^{-i\omega} - 1)] - 1] \rangle} + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k + \alpha \langle q e^{-q} \exp[q e^{-i\omega}] \rangle} + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^\ell}{\ell!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} \langle q e^{-q} \exp[q e^{-i\omega}] \rangle^\ell + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^\ell}{\ell!} \left\langle \prod_{r \leq \ell} (q_r e^{-q_r}) \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k} e^{e^{-i\omega} \sum_{r \leq l} q_r} \right\rangle_{q_1 \ldots q_\ell} + \mathcal{O}(\mathrm{N}-) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^\ell}{\ell!} \left\langle \prod_{r \leq \ell} (q_r e^{-q_r}) \sum_{s \geq 0} \frac{(\sum_{r \leq \ell} q_r)^s}{s!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} e^{i\omega k - i\omega s} \right\rangle_{q_1 \ldots q_\ell} + \mathcal{O}(N^{-1}) \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^\ell}{\ell!} \left\langle \prod_{r \leq \ell} (q_r e^{-q_r}) \frac{(\sum_{r \leq \ell} q_r)^k}{k!} \right\rangle_{q_1 \ldots q_\ell} + \mathcal{O}(N^{-1}).
\end{aligned}
\tag{C 1}
$$

Hence, for large network sizes $N \to \infty$ we obtain

$$
\begin{aligned}
\lim_{N \to \infty} p(k) &= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^\ell}{\ell!} \left\langle \left(\prod_{r \leq \ell} q_r \right) e^{-\sum_{r \leq \ell} q_r} \left(\left(\sum_{r \leq \ell} q_r \right)^k / k! \right) \right\rangle_{q_1 \ldots q_\ell} \\
&= e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{\alpha^\ell}{\ell!} \sum_{q_1 \ldots q_\ell \geq 0} p(q_1) \ldots p(q_\ell) q_1 \ldots q_\ell e^{-\sum_{r \leq \ell} q_r} \left(\left(\sum_{r \leq \ell} q_r \right)^k / k! \right).
\end{aligned}
\tag{C 2}
$$

We can rewrite this in terms of the distribution $W(q) = qP(q)/\langle q \rangle$, which denotes the likelihood to draw a link attached to a node of degree q in the bipartite graph,

$$
\lim_{N \to \infty} p(k) = e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha \langle q \rangle)^\ell}{\ell!} \sum_{q_1 \ldots q_\ell \geq 0} W(q_1) \ldots W(q_\ell) e^{-\sum_{r \leq \ell} q_r} \left(\left(\sum_{r \leq \ell} q_r \right)^k / k! \right), \tag{C 3}
$$

and upon defining

$$
P(y) = e^{-\alpha\langle q \rangle} \sum_{\ell \geq 0} \frac{(\alpha \langle q \rangle)^\ell}{\ell!} \sum_{q_1 \ldots q_\ell \geq 0} W(q_1) \ldots W(q_\ell) \, \delta \left[y - \sum_{r \leq \ell} q_r \right], \tag{C 4}
$$

we finally get to

$$
\lim_{N \to \infty} p(k) = \int_0^\infty dy \, \frac{P(y) \, e^{-y} y^k}{k!}. \tag{C 5}
$$

The interpretation is that if we draw ℓ from a Poisson distribution with $\langle \ell \rangle = \alpha \langle q \rangle$, and then draw ℓ variables q_r from $W(q_r)$, we find k as a Poissonian variable with $\langle k \rangle = \sum_{r \leq \ell} q_r$.

Clearly $p(k)$ is normalized, and for its first moment we find

$$\langle k\rangle = \int_0^\infty \mathrm{d}y\, P(y)\, y = \mathrm{e}^{-\alpha\langle q\rangle} \sum_{\ell\geq 0} \frac{(\alpha\langle q\rangle)^\ell}{\ell!} \sum_{q_1\ldots q_\ell\geq 0} W(q_1)\ldots W(q_\ell) \sum_{r\leq \ell} q_r$$
$$= \mathrm{e}^{-\alpha\langle q\rangle} \sum_{\ell>0} \frac{(\alpha\langle q\rangle)^\ell}{(\ell-1)!} \sum_q W(q) q = \alpha\langle q^2\rangle. \tag{C6}$$

For the second moment, we obtain

$$\langle k^2\rangle = \langle k\rangle + \int_0^\infty \mathrm{d}y\, P(y)\, y^2$$
$$= \alpha\langle q^2\rangle + \mathrm{e}^{-\alpha\langle q\rangle} \sum_{\ell\geq 0} \frac{(\alpha\langle q\rangle)^\ell}{\ell!} \sum_{q_1\ldots q_\ell\geq 0} W(q_1)\ldots W(q_\ell) \sum_{r,s\leq \ell} q_r q_s$$
$$= \alpha\langle q^2\rangle + \mathrm{e}^{-\alpha\langle q\rangle} \left(\sum_q W(q) q\right)^2 \sum_{\ell>0} \frac{(\alpha\langle q\rangle)^\ell}{\ell!} \ell^2$$
$$+ \mathrm{e}^{-\alpha\langle q\rangle} \left[\sum_q W(q) q^2 - \left(\sum_q W(q) q\right)^2\right] \sum_{\ell>0} \frac{(\alpha\langle q\rangle)^\ell}{(\ell-1)!}$$
$$= \alpha\langle q^2\rangle + \mathrm{e}^{-\alpha\langle q\rangle} \sum_{\ell\geq 0} \frac{(\alpha\langle q\rangle)^\ell}{\ell!} \sum_{\mathrm{q}_1\ldots \mathrm{q}_\ell\geq 0} W(q_1)\ldots W(q_\ell) \sum_{r,s\leq \ell} q_r q_s$$
$$= \alpha\langle q^2\rangle + \alpha\left[\langle q^3\rangle - \frac{\langle q^2\rangle^2}{\langle q\rangle}\right] + \frac{\langle q^2\rangle^2}{\langle q\rangle^2} \mathrm{e}^{-\alpha\langle q\rangle} \sum_{\ell>0} \frac{(\alpha\langle q\rangle)^\ell}{\ell!} \ell^2$$
$$= \alpha\langle q^2\rangle + \alpha\left[\langle q^3\rangle - \frac{\langle q^2\rangle^2}{\langle q\rangle}\right] + \frac{\langle q^2\rangle^2}{\langle q\rangle^2} [\alpha^2\langle q\rangle^2 + \alpha\langle q\rangle]$$
$$= \alpha\langle q^2\rangle + \alpha\langle q^3\rangle + \alpha^2\langle q^2\rangle^2. \tag{C7}$$

This is in agreement with results from a direct calculation:

$$\langle k^2\rangle = \frac{1}{N} \sum_{i\neq j\neq k} \langle c_{ij} c_{k\ell}\rangle = \frac{1}{N} \sum_{i\neq j} \langle c_{ij} c_{ji}\rangle + \frac{1}{N} \sum_{[ijk]} \langle c_{ij} c_{jk}\rangle$$
$$= \frac{1}{N} \sum_{i\neq j} \sum_{\mu\nu} \langle \xi_i^\mu \xi_j^\mu \xi_i^\nu \xi_j^\nu\rangle + \frac{1}{N} \sum_{[ijk]} \sum_{\mu\nu} \langle \xi_i^\mu \xi_j^\mu \xi_j^\nu \xi_k^\nu\rangle$$
$$= \frac{1}{N} \sum_{i\neq j} \sum_\mu \langle \xi_i^\mu \xi_j^\mu\rangle + \frac{1}{N} \sum_{[ijk]} \sum_{\mu\neq\nu} \langle \xi_i^\mu \xi_j^\mu\rangle \langle \xi_j^\nu \xi_k^\nu\rangle$$
$$+ \frac{1}{N} \sum_{[ijk]} \sum_\mu \langle \xi_i^\mu \xi_j^\mu \xi_k^\nu\rangle + \mathcal{O}(N^{-1})$$
$$= \frac{1}{N} \sum_{i\neq j} \sum_\mu \frac{q_\mu^2}{N^2} + \frac{1}{N} \sum_{[ijk]} \sum_{\mu\neq\nu} \frac{q_\mu^2}{N^2} \frac{q_\nu^2}{N^2} + \frac{1}{N} \sum_{[ijk]} \sum_\mu \frac{q_\mu^3}{N^3} + \mathcal{O}(N^{-1})$$
$$= \alpha\langle q^2\rangle + (\alpha\langle q^2\rangle)^2 + \alpha\langle q^3\rangle + \mathcal{O}(N^{-1})$$
$$= \langle k\rangle + \langle k\rangle^2 + \alpha\langle q^3\rangle + \mathcal{O}(N^{-1}). \tag{C8}$$

C.2. The *d*-ensemble

We can calculate the asymptotic degree distribution in the *d*-ensemble as follows:

$$p(k) = \lim_{N\to\infty} \frac{1}{N} \sum_i \langle \delta_{k,\sum_j c_{ij}}\rangle_\xi = \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \langle \mathrm{e}^{-\mathrm{i}\omega \sum_\mu \xi_i^\mu \sum_j \xi_j^\mu}\rangle_\xi$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \prod_\mu \left[1 + \frac{d_i}{\alpha N} \left(\prod_j \left\langle \mathrm{e}^{-\mathrm{i}\omega \xi_j^\mu}\right\rangle - 1\right)\right]$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \prod_\mu \left[1 + \frac{d_i}{\alpha N} (\mathrm{e}^{(\langle d\rangle/\alpha)(\mathrm{e}^{-\mathrm{i}\omega}-1)} - 1)\right]$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k + d_i(\mathrm{e}^{(\langle d\rangle/\alpha)(\mathrm{e}^{-\mathrm{i}\omega}-1)} - 1)}$$
$$= \sum_d P(d) \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{i\omega k + d(\mathrm{e}^{(\langle d\rangle/\alpha)(\mathrm{e}^{-\mathrm{i}\omega}-1)} - 1)}$$
$$= \sum_d P(d) \mathrm{e}^{-d} \sum_\ell \frac{d^\ell}{\ell!} \mathrm{e}^{-\ell(\langle d\rangle/\alpha)} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k + \ell(\langle d\rangle/\alpha)\mathrm{e}^{-\mathrm{i}\omega}}$$
$$= \sum_d P(d) \sum_\ell \mathrm{e}^{-d} \frac{d^\ell}{\ell!} \mathrm{e}^{-\ell(\langle \mathrm{d}\rangle/\alpha)} \frac{(\ell\langle d\rangle/\alpha)^k}{k!}. \tag{C9}$$

C.3. The mixed ensemble

In the mixed ensemble, we have the asymptotic degree distribution

$$p(k) = \lim_{N\to\infty} \frac{1}{N} \sum_i \langle \delta_{k,\sum_j c_{ij}}\rangle_\xi = \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \langle \mathrm{e}^{-\mathrm{i}\omega \sum_\mu \xi_i^\mu \sum_j \xi_j^\mu}\rangle_\xi$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \prod_\mu \left[1 + \frac{d_i q_\mu}{\alpha\langle q\rangle N} \left(\prod_j \left\langle \mathrm{e}^{-\mathrm{i}\omega \xi_j^\mu}\right\rangle - 1\right)\right]$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \prod_\mu \left[1 + \frac{d_i q_\mu}{\alpha\langle q\rangle N} (\mathrm{e}^{q_\mu(\mathrm{e}^{-\mathrm{i}\omega}-1)} - 1)\right]$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_i \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k + \sum_\mu (d_i q_\mu/\alpha\langle q\rangle N)(\mathrm{e}^{q_\mu(\mathrm{e}^{-\mathrm{i}\omega}-1)} - 1)}$$
$$= \sum_d P(d) \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k + (d/\langle q\rangle)\langle q(\mathrm{e}^{q(\mathrm{e}^{-\mathrm{i}\omega}-1)} 1)\rangle_q}$$
$$= \sum_d P(d) \mathrm{e}^{-d} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k + (d/\langle q\rangle)\langle q\mathrm{e}^{-q} \exp[q\mathrm{e}^{-\mathrm{i}\omega}]\rangle_q}$$
$$= \sum_d P(d) \mathrm{e}^{-d} \sum_{\ell\geq 0} \frac{(d/\langle q\rangle)^\ell}{\ell!} \int_{-\pi}^{\pi} \frac{d\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega k} \langle q\mathrm{e}^{-q} \exp[q\mathrm{e}^{-\mathrm{i}\omega}]\rangle_q^\ell$$
$$= \sum_d P(d) \mathrm{e}^{-d} \sum_{\ell\geq 0} \frac{d^\ell}{\ell!} \left\langle \prod_{r\leq \ell} \left(\frac{q_r \mathrm{e}^{-q_r}}{\langle q\rangle}\right) \frac{(\sum_{r\leq \ell} q_r)^k}{k!} \right\rangle_{q_1\ldots q_\ell}. \tag{C10}$$

We can rewrite this expression in terms of the associated distribution $W(q) = qP(q)/\langle q\rangle$ as

$$p(k) = \sum_d P(d) \mathrm{e}^{-\mathrm{d}} \sum_{\ell\geq 0} \frac{d^\ell}{\ell!} \left\langle \prod_{r\leq \ell} \left(\frac{q_r \mathrm{e}^{-q_r}}{\langle q\rangle}\right) \frac{(\sum_{r\leq \ell} q_r)^k}{k!} \right\rangle_{q_1\ldots q_\ell}$$
$$= \sum_d P(d) \mathrm{e}^{-\mathrm{d}} \sum_{\ell\geq 0} \frac{d^\ell}{\ell!} \sum_{q_1\ldots q_\ell\geq 0} \mathrm{W}(q_1)\ldots \mathrm{W}(q_\ell) \mathrm{e}^{-\sum_{r\leq \ell} q_r} \frac{(\sum_{r\leq \ell} q_r)^k}{k!} \tag{C11}$$

or, equivalently, as

$$p(k) = \int_0^\infty dy\, P(y)\, \frac{\mathrm{e}^{-y} y^k}{k!}, \tag{C 12}$$

where

$$P(y) = \sum_d P(d)\mathrm{e}^{-d} \sum_{\ell\geq 0} \frac{d^\ell}{\ell!} \sum_{q_1 \ldots q_\ell \geq 0} W(q_1)\ldots W(q_\ell)\, \delta\left[y - \sum_{r\leq \ell} q_r\right]. \tag{C 13}$$

The first two moments of $p(k)$ are

$$\langle k\rangle = \int_0^\infty dy P(y) y = \sum_{\mathrm{d}} P(d)\mathrm{e}^{-d} \sum_{\ell\geq 0} \frac{d^\ell}{\ell!} \sum_{q_1\ldots q_\ell \geq 0} W(q_1)\ldots W(q_\ell) \sum_{r\leq \ell} q_r$$
$$= \sum_d P(d)\mathrm{e}^{-d} \sum_{\ell>0} \frac{d^\ell}{(\ell-1)!} \sum_q W(q) q = \langle d\rangle \frac{\langle q^2\rangle}{\langle q\rangle} = \alpha\langle q^2\rangle \tag{C 14}$$

and

$$\langle k^2\rangle = \langle k\rangle + \int_0^\infty dy P(y) y^2$$
$$= \alpha\langle q^2\rangle + \sum_d P(d) e^{-d} \sum_{\ell\geq 0} \frac{d^\ell}{\ell!} \sum_{q_1\ldots q_\ell \geq 0} W(q_1)\ldots W(q_\ell) \sum_{r,s\leq \ell} q_r q_s$$
$$= \alpha\langle q^2\rangle + \sum_d P(d) e^{-d} \sum_{\ell>0} \frac{d^\ell}{\ell!} \left[\ell \sum_q W(q) q^2 + \ell(\ell-1)\left(\sum_q W(q) q\right)^2\right]$$
$$= \alpha\langle q^2\rangle + \frac{\langle q^3\rangle}{\langle q\rangle}\langle d\rangle + \langle d^2\rangle \frac{\langle q^2\rangle^2}{\langle q\rangle^2} = \alpha\langle q^2\rangle + \alpha\langle q^3\rangle + \frac{\langle d^2\rangle}{\langle d\rangle^2}\langle k\rangle^2. \tag{C 15}$$

For loops of length 3, we proceed in the same way, obtaining

$$\frac{1}{N(N-1)(N-2)} \sum_{i\neq j\neq k(\neq i)} \langle \delta_{c_{ij},0}\delta_{c_{jk},0}\rangle$$
$$= \frac{1}{N(N-1)(N-2)} \sum_{i\neq j\neq k(\neq i)} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \prod_\mu \langle \mathrm{e}^{\mathrm{i}\xi_j^\mu(\xi_i^\mu\omega + \xi_k^\mu\omega')}\rangle$$
$$= \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \prod_\mu \left\{1 + \frac{q_\mu^2}{N^2}\left[(\mathrm{e}^{\mathrm{i}\omega} + \mathrm{e}^{\mathrm{i}\omega'} - 2) + \frac{q_\mu}{N}(\mathrm{e}^{\mathrm{i}(\omega+\omega')} - \mathrm{e}^{\mathrm{i}\omega} - \mathrm{e}^{\mathrm{i}\omega'} + 1)\right]\right\}$$
$$= \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \mathrm{e}^{(\alpha\langle q^2\rangle/N)(\mathrm{e}^{\mathrm{i}\omega} + \mathrm{e}^{\mathrm{i}\omega'} - 2) + (\alpha\langle q^3\rangle/N^2)(\mathrm{e}^{\mathrm{i}(\omega+\omega')} - \mathrm{e}^{\mathrm{i}\omega} - \mathrm{e}^{\mathrm{i}\omega'} + 1) - (\alpha\langle q^4\rangle/2N^3)(\mathrm{e}^{\mathrm{i}\omega} + \mathrm{e}^{\mathrm{i}\omega'} - 2)^2}$$
$$= 1 - 2\frac{\alpha\langle q^2\rangle}{N} + \frac{\alpha\langle q^3\rangle}{N^2} + 2\frac{\alpha^2\langle q^2\rangle^2}{N^2} - 2\frac{\alpha\langle q^4\rangle}{N^3} - 2\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3} - \frac{4}{3}\frac{\alpha^3\langle q^2\rangle^3}{N^3}. \tag{D 3}$$

Appendix D. The link between observables in the *a* and *c* networks

In this appendix, we inspect the relationship between expectation values of various observables in the ensembles $p(\boldsymbol{a})$ and $p(\boldsymbol{c})$.

D.1. The *q*-ensemble

Denoting averages in the $\boldsymbol{a}$ ensemble as $\langle \ldots \rangle_a$, we have, for the q-ensemble of bipartite graphs

$$\langle k\rangle_a = \frac{1}{N}\sum_{ij}\langle a_{ij}\rangle_a = \frac{1}{N}\sum_{ij}\left\langle \theta\left[c_{ij} - \frac{1}{2}\right]\right\rangle$$
$$= \frac{1}{N}\sum_{ij}[1 - \langle \delta_{c_{ij},0}\rangle] = \alpha\langle q^2\rangle + \mathcal{O}(N^{-1}) = \langle k\rangle + \mathcal{O}(N^{-1}) \tag{D 1}$$

and

$$\langle k^2\rangle_a = \frac{1}{N}\sum_{i\neq j\neq k}\langle a_{ij}a_{jk}\rangle = \frac{1}{N}\sum_{ij}\langle a_{ij}\rangle + \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle a_{ij}a_{jk}\rangle$$
$$= \frac{1}{N}\sum_{ij}\langle(1-\delta_{c_{ij},0})\rangle + \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle(1-\delta_{c_{ij},0})(1-\delta_{c_{jk},0})\rangle$$
$$= \frac{1}{N}\sum_{ij}\frac{\alpha\langle q^2\rangle}{N} + \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}(1 - 2\langle\delta_{c_{ij},0}\rangle + \langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle)$$
$$= (N-1)(N-2) - 2(N-1)(N-2)\left(1 - \frac{\alpha\langle q^2\rangle}{N} + \frac{\alpha^2\langle q^2\rangle^2}{2N^2}\right)$$
$$+ (N-1)(N-2)\left(1 - 2\frac{\alpha\langle q^2\rangle}{N} + \frac{\alpha\langle q^3\rangle}{N^2} + 2\frac{\alpha^2\langle q^2\rangle^2}{N^2}\right) + \alpha\langle q^2\rangle$$
$$= \alpha\langle q^2\rangle + \alpha\langle q^3\rangle + \alpha^2\langle q^2\rangle^2 \equiv \langle k^2\rangle, \tag{D 2}$$

where we used

$$
\begin{aligned}
m_3^a &= \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle a_{ij}a_{jk}a_{ki}\rangle = \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle(1-\delta_{c_{ij},0})(1-\delta_{c_{jk},0})(1-\delta_{c_{ki},0})\rangle \\
&= \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}(1-3\langle\delta_{c_{ij},0}\rangle+3\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle-\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{ki},0}\rangle)\rangle \\
&= (N-1)(N-2)-3(N-1)(N-2)\left(1-\frac{\alpha\langle q^2\rangle}{N}+2\frac{\alpha^2\langle q^2\rangle^2}{2N^2}\right) \\
&\quad +3(N-1)(N-2)\left(1-2\frac{\alpha\langle q^2\rangle}{N}+\frac{\alpha\langle q^3\rangle}{N^2}+2\frac{\alpha^2\langle q^2\rangle^2}{N^2}\right) \\
&\quad -(N-1)(N-2)\left(1-3\frac{\alpha\langle q^2\rangle}{N}+2\frac{\alpha\langle q^3\rangle}{N^2}+\frac{9}{2}\frac{\alpha^2\langle q^2\rangle^2}{N^2}\right) = \alpha\langle q^3\rangle \equiv m_3^c,
\end{aligned}
\tag{D 4}
$$

where we used

$$
\begin{aligned}
&\frac{1}{N(N-1)(N-2)}\sum_{i\neq j\neq k(\neq i)}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{ki},0}\rangle \\
&= \frac{1}{N(N-1)(N-2)}\sum_{i\neq j\neq k(\neq i)}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\langle e^{i\xi_i^\mu(\xi_j^\mu\omega+\xi_k^\mu\omega'')+i\xi_j^\mu\xi_k^\mu\omega'}\rangle \\
&= \int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\left\{1+\frac{q_\mu^2}{N^2}\left[(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)+\frac{q_\mu}{N}(e^{i(\omega+\omega'+\omega'')}-e^{i\omega}-e^{i\omega'}-e^{i\omega'}+2)\right]\right\} \\
&= \int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}e^{\sum_\mu(q_\mu^2/N^2)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)+\sum_\mu(q_\mu^3/N^3)(e^{i(\omega+\omega'+\omega'')}-e^{i\omega}-e^{i\omega'}-e^{i\omega''}+2)-\sum_\mu(q_\mu^4/2N^4)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)^2} \\
&= \int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}e^{(\alpha\langle q^2\rangle/N)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)+(\alpha\langle q^3\rangle/N^2)(e^{i(\omega+\omega'+\omega'')}-e^{i\omega}-e^{i\omega'}-e^{i\omega''}+2)-(\alpha\langle q^4\rangle/2N^3)(+e^{i\omega'}+e^{i\omega''}-3)^2} \\
&= 1-3\frac{\alpha\langle q^2\rangle}{N}+2\frac{\alpha\langle q^3\rangle}{N^2}+\frac{9}{2}\frac{\alpha^2\langle q^2\rangle^2}{N^2}-\frac{9}{2}\frac{\alpha\langle q^4\rangle}{N^3}-6\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3}-\frac{9}{2}\frac{\alpha^3\langle q^2\rangle^3}{N^3}.
\end{aligned}
\tag{D 5}
$$

Finally for loops of length 4, we have

$$
\begin{aligned}
m_4^a &= \frac{1}{N}\sum_{[i,j,k,\ell]}\langle a_{ij}a_{jk}a_{k\ell}a_{\ell i}\rangle \\
&= \frac{1}{N}\sum_{[i,j,k,\ell]}\langle(1-\delta_{c_{ij},0})(1-\delta_{c_{jk},0})(1-\delta_{c_{k\ell},0})(1-\delta_{c_{\ell i},0})\rangle \\
&= \frac{1}{N}\sum_{[i,j,k,\ell]}(1-4\langle\delta_{c_{ij},0}\rangle+4\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle+2\langle\delta_{c_{ij},0}\rangle\langle\delta_{c_{jk},0}\rangle-4\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\rangle)\rangle \\
&\quad +\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\delta_{c_{\ell i},0})\rangle \\
&= (N-1)(N-2)(N-3)\left\{1-4\left(1-\frac{\alpha\langle q^2\rangle}{N}+\frac{\alpha^2\langle q^2\rangle^2}{2N^2}-\frac{\alpha^3\langle q^2\rangle^3}{6N^3}-\frac{\alpha\langle q^4\rangle}{2N^3}\right)\right. \\
&\quad +4\left(1-2\frac{\alpha\langle q^2\rangle}{N}+\frac{\alpha\langle q^3\rangle}{N^2}+2\frac{\alpha^2\langle q^2\rangle^2}{N^2}-\frac{4}{3}\frac{\alpha^3\langle q^2\rangle^3}{N^3}-2\frac{\alpha\langle q^4\rangle}{N^3}-2\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3}\right) \\
&\quad +2\left(1-\frac{\alpha\langle q^2\rangle}{N}+\frac{\alpha^2\langle q^2\rangle^2}{2N^2}-\frac{\alpha^3\langle q^2\rangle^3}{6N^3}-\frac{\alpha\langle q^4\rangle}{2N^3}\right)^2 \\
&\quad -4\left(1-3\frac{\alpha\langle q^2\rangle}{N}+2\frac{\alpha\langle q^3\rangle}{N^2}+\frac{9}{2}\frac{\alpha^2\langle q^2\rangle^2}{N^2}-\frac{9}{2}\frac{\alpha^3\langle q^2\rangle^3}{N^3}-\frac{9}{2}\frac{\alpha\langle q^4\rangle}{N^3}-6\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3}\right) \\
&\quad \left.+\left(1-4\frac{\alpha\langle q^2\rangle}{N}+4\frac{\alpha\langle q^3\rangle}{N^2}-9\frac{\alpha}{N^3}\langle q^4\rangle+8\frac{\alpha^2\langle q^2\rangle^2}{N^2}-16\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3}-\frac{32}{3}\frac{\alpha^3\langle q^2\rangle^3}{N^3}\right)\right\} \\
&= \alpha\langle q^4\rangle \equiv m_4^c,
\end{aligned}
\tag{D 6}
$$

where we used

$$\begin{aligned}
&\frac{1}{N(N-1)(N-2)(N-3)}\sum_{i\neq j\neq k(\neq i)}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\rangle\\
&=\frac{1}{N(N-1)(N-2)(N-3)}\sum_{i\neq j\neq k(\neq i)}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\langle e^{i\xi_j^{\mu}(\xi_i^{\mu}\omega+\xi_k^{\mu}\omega')+i\xi_\ell^{\mu}\xi_k^{\mu}\omega''}\rangle\\
&=\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\left\{1+\frac{q_\mu^2}{N^2}\left[(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)+\frac{q_\mu}{N}(e^{i\omega'}-1)(e^{i\omega}+e^{i\omega''}-2)\right.\right.\\
&\quad\left.\left.+\frac{q_\mu^2}{N^2}e^{i\omega'}(e^{i\omega}-1)(e^{i\omega''}-1)\right\}\right]\\
&=\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}e^{\sum_\mu(q_\mu^2/N^2)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)+\sum_\mu(q_\mu^3/N^3)(e^{i\omega'}-1)(e^{i\omega}+e^{i\omega''}-2))-\sum_\mu(q_\mu^4/2N^4)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)^2}\\
&\quad\times e^{(q_\mu^4/N^4)e^{i\omega'}(e^{i\omega}-1)(e^{i\omega''}-1)}\\
&=\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}e^{(\alpha\langle q^2\rangle/N)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)+(\alpha\langle q^3\rangle/N^2)(e^{i\omega'}-1)(e^{i\omega}+e^{i\omega''}-2))-(\alpha\langle q^4\rangle/2N^3)(e^{i\omega}+e^{i\omega'}+e^{i\omega''}-3)^2}\\
&\quad\times e^{\alpha(\langle q^4\rangle/N^3)e^{i\omega'}(e^{i\omega}-1)(e^{i\omega''}-1)}\\
&=1-3\frac{\alpha\langle q^2\rangle}{N}+2\frac{\alpha\langle q^3\rangle}{N^2}+\frac{9}{2}\frac{\alpha^2\langle q^2\rangle^2}{N^2}-\frac{9}{2}\frac{\alpha\langle q^4\rangle}{N^3}-6\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3}-\frac{9}{2}\frac{\alpha^3\langle q^2\rangle^3}{N^3}
\end{aligned}\tag{D7}$$

and

$$\begin{aligned}
&\frac{1}{N(N-1)(N-2)(N-3)}\sum_{[ijk\ell]}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\delta_{c_{\ell i},0}\rangle\\
&=\frac{1}{N(N-1)(N-2)(N-3)}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4}\prod_{\mu}\langle e^{i\xi_i^{\mu}(\xi_j^{\mu}\omega+\xi_\ell^{\mu}\omega''')+\xi_k^{\mu}(\xi_j^{\mu}\omega'+\xi_\ell^{\mu}\omega'')}\rangle\\
&=\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4}\prod_{\mu}\left\{\left(1-\frac{q_\mu}{N}\right)\left\{\frac{q_\mu}{N}\left[\frac{q_\mu^2}{N^2}e^{i(\omega+\omega')}+\frac{q_\mu}{N}\left(1-\frac{q_\mu}{N}\right)(e^{i\omega'}+e^{i\omega''})+\left(\frac{1-q_\mu}{N}\right)^2\right]\right.\right.\\
&\quad\left.+\left(1-\frac{q_\mu}{N}\right)\right\}+\frac{q_\mu}{N}\left\{\frac{q_\mu^2}{N^2}e^{i(\omega+\omega''\prime)}\left(1-\frac{q_\mu}{N}+\frac{q_\mu}{N}e^{i(\omega'+\omega'')}\right)\right.\\
&\quad\left.\left.+\frac{q_\mu}{N}\left(1-\frac{q_\mu}{N}\right)\left[e^{i\omega}\left(1-\frac{q_\mu}{N}+\frac{q_\mu}{N}e^{i\omega}\right)+e^{i\omega'''}\left(1-\frac{q_\mu}{N}+\frac{q_\mu}{N}e^{i\omega''}\right)\right]+\left(1-\frac{q_\mu}{N}\right)^2\right\}\right.\\
&=\prod_{\mu}\left\{\frac{q_\mu}{N}\left(1-\frac{q_\mu}{N}\right)^2+\left(1-\frac{q_\mu}{N}\right)\left[\frac{q_\mu}{N}\left(1-\frac{q_\mu}{N}\right)^2+\left(1-\frac{q_\mu}{N}\right)\right]\right\}\\
&=\prod_{\mu}\left\{1-4\frac{q_\mu^2}{N^2}+4\frac{q_\mu^3}{N^3}-\frac{q_\mu^4}{N^4}\right\}=e^{-4(\alpha\langle q^2\rangle/N)+4(\alpha\langle q^3\rangle/N^2)-9(\alpha\langle q^4\rangle/N^3)+\mathcal{O}(N^{-4})}\\
&=1-4\frac{\alpha\langle q^2\rangle}{N}+4\frac{\alpha\langle q^3\rangle}{N^2}+8\frac{\alpha^2\langle q^2\rangle^2}{N^2}-9\frac{\alpha\langle q^4\rangle}{N^3}-16\frac{\alpha^2\langle q^2\rangle\langle q^3\rangle}{N^3}-\frac{32}{3}\frac{\alpha^3\langle q^2\rangle^3}{N^3}+\mathcal{O}(N^{-4}).
\end{aligned}\tag{D8}$$

Again, the square brackets underneath the summations indicate that all indices are different, to exclude backtracking in the counting of loops of length 4.

D.2. The *d*-ensemble

For the *d*-ensemble, denoting averages relating to $\boldsymbol{a}$ as $\langle\ldots\rangle_a$, we have

$$\begin{aligned}
\langle k\rangle_a&=\frac{1}{N}\sum_{ij}\langle a_{ij}\rangle_a=\frac{1}{N}\sum_{ij}[1-\langle\delta_{c_{ij},0}\rangle]=\\
&=\frac{1}{N}\sum_{ij}\left[\frac{d_id_j}{\alpha N}-\frac{1}{2}\left(\frac{d_id_j}{\alpha N}\right)^2+\frac{1}{6}\left(\frac{d_id_j}{\alpha N}\right)^3+\frac{1}{2}\frac{d_i^2d_j^2}{(\alpha N)^3}\right]\\
&=\frac{\langle d\rangle^2}{\alpha}+\mathcal{O}(N^{-1})=\langle k\rangle+\mathcal{O}(N^{-1})
\end{aligned}\tag{D9}$$

and

$$\begin{aligned}\langle k^2\rangle_a &= \frac{1}{N}\sum_{i\neq j\neq k}\langle a_{ij}a_{jk}\rangle = \frac{1}{N}\sum_{ij}\langle a_{ij}\rangle + \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle a_{ij}a_{jk}\rangle \\ &= \frac{\langle d\rangle^2}{\alpha} + \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle(1-\delta_{c_{ij},0})(1-\delta_{c_{jk},0})\rangle \\ &= \frac{\langle d\rangle^2}{\alpha} + \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}(1-2\langle\delta_{c_{ij},0}\rangle + \langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle) \\ &= \frac{\langle d\rangle^2}{\alpha} + (N-1)(N-2) - \frac{2}{N}\sum_{[ijk]}\left(1-\frac{d_id_j}{\alpha N}+\frac{1}{2}\left(\frac{d_id_j}{\alpha N}\right)^2\right) + \frac{1}{N}\sum_{[ijk]}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle \\ &= \frac{\langle d\rangle^2}{\alpha} + 2\frac{\langle d\rangle^2}{\alpha}N - 2\frac{\langle d\rangle^2}{\alpha} - \frac{\langle d^2\rangle^2}{\alpha^2} - 2N\frac{\langle d\rangle^2}{\alpha} + 2\frac{\langle d\rangle^2}{\alpha} + \frac{\langle d\rangle^3}{\alpha^2} + \frac{\langle d^2\rangle^2}{\alpha^2} - \frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2} \\ &= \frac{\langle d\rangle^2}{\alpha} + \frac{\langle d\rangle^3}{\alpha^2} + \frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2} \equiv \langle k^2\rangle, \end{aligned} \tag{D 10}$$

where we used

$$\begin{aligned}&\frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle = \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\int_{-\pi}^{\pi}\frac{d\omega d\omega'}{4\pi^2}\prod_{\mu}\langle e^{i\xi_j^{\mu}(\xi_i^{\mu}\omega+\xi_k^{\mu}\omega')}\rangle \\ &= \frac{1}{N}\sum_{[ijk]}\int_{-\pi}^{\pi}\frac{d\omega d\omega'}{4\pi^2}\prod_{\mu}\left\{1+\frac{d_j}{\alpha N}\left[\frac{d_i}{\alpha N}(e^{i\omega}-1)+\frac{d_k}{\alpha N}(e^{i\omega'}-1)+\frac{d_id_k}{(\alpha N)^2}(e^{i(\omega+\omega')}-e^{i\omega}-e^{i\omega'}+1)\right]\right\} \\ &= \frac{1}{N}\sum_{[ijk]}\int_{-\pi}^{\pi}\frac{d\omega d\omega'}{4\pi^2}\left\{1+\frac{d_j}{\alpha N}\left[d_i(e^{i\omega}-1)+d_k(e^{i\omega'}-1)+\frac{d_id_k}{\alpha N}(e^{i(\omega+\omega')}-e^{i\omega}-e^{i\omega'}+1)\right]\right. \\ &\quad\left.+\frac{1}{2}\left(\frac{d_j}{\alpha N}[d_i(e^{i\omega}-1)+d_k(e^{i\omega'}-1)]\right)^2-\frac{d_id_j^2d_k}{(\alpha N)^3}(d_i+d_k)\right\} \\ &= \frac{1}{N}\sum_{[ijk]}\left\{1+\frac{d_j}{\alpha N}\left[-d_i-d_k+\frac{d_id_k}{\alpha N}\right]+\frac{1}{2}\left(\frac{d_j}{\alpha N}\right)^2(d_i^2+d_k^2+2d_id_k)-\frac{d_i^2d_j^2d_k}{(\alpha N)^3}-\frac{d_id_j^2d_k^2}{(\alpha N)^3}\right\} \\ &= (N-1)(N-2)-2N\frac{\langle d\rangle^2}{\alpha}+2\frac{\langle d\rangle^2}{\alpha}+\frac{\langle d\rangle^3}{\alpha^2}+\frac{\langle d^2\rangle^2}{\alpha^2}+\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}. \end{aligned} \tag{D 11}$$

For loops of length 3, we have

$$\begin{aligned} m_3^a &= \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle a_{ij}a_{jk}a_{ki}\rangle = \frac{1}{N}\sum_{i\neq j\neq k(\neq i)}\langle(1-\delta_{c_{ij},0})(1-\delta_{c_{jk},0})(1-\delta_{c_{ki},0})\rangle \\ &= \frac{1}{N}\sum_{[ijk]}(1-3\langle\delta_{c_{ij},0}\rangle+3\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle-\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{ki},0}\rangle) \\ &= (N-1)(N-2)-3\frac{1}{N}\sum_{[ijk]}\left(1-\frac{d_id_j}{\alpha N}+\frac{1}{2}\left(\frac{d_id_j}{\alpha N}\right)^2\right) \\ &\quad+3\left[(N-1)(N-2)-2N\frac{\langle d\rangle^2}{\alpha}+2\frac{\langle d\rangle^2}{\alpha}+\frac{\langle d\rangle^3}{\alpha^2}+\frac{\langle d^2\rangle^2}{\alpha^2}-\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}\right] \\ &\quad-(N-1)(N-2)+3N\frac{\langle d\rangle^2}{\alpha}-3\frac{\langle d\rangle^2}{\alpha}-2\frac{\langle d\rangle^3}{\alpha^2}-\frac{3}{2}\frac{\langle d^2\rangle^2}{\alpha^2}-3\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2} \\ &= 3\frac{\langle d\rangle^2}{\alpha}N-3\frac{\langle d\rangle^2}{\alpha}-\frac{3}{2}\frac{\langle d^2\rangle^2}{\alpha^2} \\ &\quad+3\left[-2N\frac{\langle d\rangle^2}{\alpha}+2\frac{\langle d\rangle^2}{\alpha}+\frac{\langle d\rangle^3}{\alpha^2}+\frac{\langle d^2\rangle^2}{\alpha^2}+\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}\right] \\ &\quad+3N\frac{\langle d\rangle^2}{\alpha}-3\frac{\langle d\rangle^2}{\alpha}-2\frac{\langle d\rangle^3}{\alpha^2}-\frac{3}{2}\frac{\langle d^2\rangle^2}{\alpha^2}-3\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}=\frac{\langle d\rangle^3}{\alpha^2}\equiv m_3^c, \end{aligned} \tag{D 12}$$

where we used

$$\begin{aligned}
\frac{1}{N}\sum_{[ijk]}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{ki},0}\rangle &= \frac{1}{N}\sum_{[ijk]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\langle e^{i\xi_i^{\mu}(\xi_j^{\mu}\omega+\xi_k^{\mu}\omega'')+i\xi_j^{\mu}\xi_k^{\mu}\omega'}\rangle \\
&= \frac{1}{N}\sum_{[ijk]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\left\{\frac{d_i}{\alpha N}\langle e^{i(\xi_j^{\mu}\omega+\xi_k^{\mu}\omega''+\xi_j^{\mu}\xi_k^{\mu}\omega'}\rangle+\left(1-\frac{d_i}{\alpha N}\right)\langle e^{i\xi_j^{\mu}\xi_k^{\mu}\omega'}\rangle\right\} \\
&= \frac{1}{N}\sum_{[ijk]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\left(1+\frac{d_jd_k}{(\alpha N)^2}(e^{i\omega'}-1)+\frac{d_i}{\alpha N}\left\{-1-\frac{d_jd_k}{(\alpha N)^2}(e^{i\omega'}-1)\right.\right. \\
&\quad \left.\left.+\frac{d_j}{\alpha N}e^{i\omega}\left[1+\frac{d_k}{\alpha N}(e^{i(\omega''+\omega')}-1)\right]+\left(1-\frac{d_j}{\alpha N}\right)\left[1+\frac{d_k}{\alpha N}(e^{i\omega''}-1)\right]\right\}\right)^{\alpha N} \\
&= \frac{1}{N}\sum_{[ijk]}\left(1-\frac{d_jd_k}{(\alpha N)^2}+\frac{d_i}{\alpha N}\left\{-\frac{d_j}{\alpha N}-\frac{d_k}{\alpha N}+2\frac{d_jd_k}{(\alpha N)^2}\right\}\right)^{\alpha N} \\
&= \frac{1}{N}\sum_{[ijk]}\left[1-\frac{d_jd_k}{\alpha N}+\frac{d_i}{\alpha N}\left\{-d_j-d_k+2\frac{d_jd_k}{\alpha N}\right\}+\frac{1}{2}\left(\frac{d_i}{\alpha N}\right)^2(d_j^2+d_k^2+2d_jd_k)\right. \\
&\quad \left.+\frac{1}{2}\frac{d_j^2d_k^2}{(\alpha N)^2}+\frac{d_id_jd_k}{(\alpha N)^2}(d_j+d_k)\right] = (N-1)(N-2)-3N\frac{\langle d\rangle^2}{\alpha}+3\frac{\langle d\rangle^2}{\alpha}+2\frac{\langle d\rangle^3}{\alpha^2}+\frac{3}{2}\frac{\langle d^2\rangle^2}{\alpha^2}+3\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}. \qquad \text{(D 13)}
\end{aligned}$$

Finally, for loops of length 4, we have

$$\begin{aligned}
m_4^a &= \frac{1}{N}\sum_{[i,j,k,\ell]}\langle a_{ij}a_{jk}a_{k\ell}a_{\ell i}\rangle = \frac{1}{N}\sum_{[i,j,k,\ell]}\langle(1-\delta_{c_{ij},0})(1-\delta_{c_{jk},0})(1-\delta_{c_{k\ell},0})(1-\delta_{c_{\ell i},0})\rangle \\
&= \frac{1}{N}\sum_{[i,j,k,\ell]}(1-4\langle\delta_{c_{ij},0}\rangle+4\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\rangle+2\langle\delta_{c_{ij},0}\rangle\langle\delta_{c_{k\ell},0}\rangle-4\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\rangle)+\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\delta_{c_{\ell i},0}\rangle) \\
&= (N-1)(N-2)(N-3)-4\left[(N-1)(N-2)(N-3)-N^2\frac{\langle d\rangle^2}{\alpha}+\frac{1}{2}N\frac{\langle d^2\rangle^2}{\alpha^2}--\frac{1}{2}\frac{\langle d^2\rangle^2}{\alpha^3}-\frac{1}{6}\frac{\langle d^3\rangle^2}{\alpha^3}\right] \\
&\quad +4\left[(N-1)(N-2)(N-3)-2N^2\frac{\langle d\rangle^2}{\alpha}+N\frac{\langle d\rangle^3}{\alpha^2}-\frac{\langle d^2\rangle^2}{\alpha^3}\right. \\
&\quad \left.-\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^3}+N\frac{\langle d^2\rangle^2}{\alpha^2}+N\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}-2\frac{\langle d^2\rangle^2\langle d\rangle}{\alpha^3}-\frac{1}{3}\frac{\langle d^3\rangle^2}{\alpha^3}-\frac{\langle d^3\rangle\langle d^2\rangle\langle d\rangle}{\alpha^3}\right] \\
&\quad +2\left[(N-1)(N-2)(N-3)-2N^2\frac{\langle d\rangle^2}{\alpha}+N\frac{\langle d^2\rangle^2}{\alpha^2}+N\frac{\langle d\rangle^4}{\alpha^2}-\frac{\langle d^2\rangle^2}{\alpha^3}-\frac{1}{3}\frac{\langle d^3\rangle^2}{\alpha^3}-\frac{\langle d^2\rangle^2\langle d\rangle^2}{\alpha^3}\right] \\
&\quad -4\left[(N-1)(N-2)(N-3)-3N^2\frac{\langle d\rangle^2}{\alpha}+2N\frac{\langle d\rangle^3}{\alpha^2}-\frac{3}{2}\frac{\langle d^2\rangle^2}{\alpha^3}\right. \\
&\quad -2\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^3}+\frac{3}{2}N\frac{\langle d^2\rangle^2}{\alpha^2}-\frac{\langle d\rangle^4}{\alpha^3}+2N\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}+N\frac{\langle d\rangle^4}{\alpha^2} \\
&\quad \left.-4\frac{\langle d^2\rangle^2\langle d\rangle}{\alpha^3}-2\frac{\langle d^2\rangle\langle d\rangle^3}{\alpha^3}-\frac{1}{2}\frac{\langle d^3\rangle^2}{\alpha^3}-2\frac{\langle d^3\rangle\langle d^2\rangle\langle d\rangle}{\alpha^3}-\frac{\langle d^2\rangle^2\langle d\rangle^2}{\alpha^3}\right] \\
&\quad +(N-1)(N-2)(N-3)-4N^2\frac{\langle d\rangle^2}{\alpha}+4N\frac{\langle d\rangle^3}{\alpha^2}-2\frac{\langle d^2\rangle^2}{\alpha^3} \\
&\quad -4\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^3}+2N\frac{\langle d^2\rangle^2}{\alpha^2}-3\frac{\langle d\rangle^4}{\alpha^3}+4N\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}+2N\frac{\langle d\rangle^4}{\alpha^2} \\
&\quad -8\frac{\langle d^2\rangle^2\langle d\rangle}{\alpha^3}-8\frac{\langle d^2\rangle\langle d\rangle^3}{\alpha^3}-\frac{2}{3}\frac{\langle d^3\rangle^2}{\alpha^3}-4\frac{\langle d^3\rangle\langle d^2\rangle\langle d\rangle}{\alpha^3}-2\frac{\langle d^2\rangle^2\langle d\rangle^2}{\alpha^3} \\
&= \frac{\langle d\rangle^4}{\alpha^3}\equiv m_4^c, \qquad \text{(D 14)}
\end{aligned}$$

where we used

$$\begin{aligned}
&\frac{1}{N}\sum_{[ijk\ell]}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\rangle = \frac{1}{N}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\langle e^{i\xi_j^\mu(\xi_i^\mu\omega+\xi_k^\mu\omega')+\xi_k^\mu\xi_\ell^\mu\omega''}\rangle\\
&= \frac{1}{N}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\prod_{\mu}\left\{\frac{d_j}{\alpha N}\langle e^{i(\xi_i^\mu\omega+\xi_k^\mu\omega')+i\xi_k^\mu\xi_\ell^\mu\omega''}\rangle + \left(1-\frac{d_j}{\alpha N}\right)\langle e^{i\xi_k^\mu\xi_\ell^\mu\omega''}\rangle\right\}\\
&= \frac{1}{N}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega''}{8\pi^3}\left(1+\frac{d_id_j}{(\alpha N)^2}(e^{i\omega}-1)+\frac{d_jd_k}{(\alpha N)^2}(e^{i\omega'}-1)+\frac{d_kd_\ell}{(\alpha N)^2}(e^{i\omega''}-1)\right.\\
&\quad+\frac{d_id_jd_k}{(\alpha N)^3}(e^{i(\omega+\omega')}-e^{i\omega}-e^{i\omega'}+1)+\frac{d_jd_kd_\ell}{(\alpha N)^3}(e^{i(\omega'+\omega'')}-e^{i\omega'}-e^{i\omega''}+1)\\
&\quad\left.+\frac{d_id_jd_kd_\ell}{(\alpha N)^4}(e^{i(\omega+\omega'+\omega'')}-e^{i(\omega+\omega')}-e^{i(\omega+\omega'')}+e^{i\omega'})\right)^{\alpha N}\\
&= \frac{1}{N}\sum_{[ijk\ell]}\left\{1-\frac{d_id_j}{\alpha N}-\frac{d_jd_k}{\alpha N}-\frac{d_kd_\ell}{\alpha N}+\frac{d_id_jd_k}{(\alpha N)^2}+\frac{d_jd_kd_\ell}{(\alpha N)^2}\right.\\
&\quad-\frac{1}{2}\left[\frac{d_i^2d_j^2}{(\alpha N)^3}+\frac{d_j^2d_k^2}{(\alpha N)^3}+\frac{d_k^2d_\ell^2}{(\alpha N)^3}+2\frac{d_id_j^2d_k}{(\alpha N)^3}+2\frac{d_id_jd_kd_\ell}{(\alpha N)^3}+2\frac{d_jd_k^2d_\ell}{(\alpha N)^3}\right]\\
&\quad+\frac{1}{2}\left[\frac{d_i^2d_j^2}{(\alpha N)^2}+\frac{d_j^2d_k^2}{(\alpha N)^2}+\frac{d_k^2d_\ell^2}{(\alpha N)^2}+2\frac{d_id_j^2d_k}{(\alpha N)^2}+2\frac{d_id_jd_kd_\ell}{(\alpha N)^2}+2\frac{d_jd_k^2d_\ell}{(\alpha N)^2}-2\frac{d_i^2d_j^2d_k}{(\alpha N)^3}\right.\\
&\quad-2\frac{d_i^2d_jd_kd_\ell}{(\alpha N)^3}-2\frac{d_id_j^2d_k^2}{(\alpha N)^3}-2\frac{d_id_jd_k^2d_\ell}{(\alpha N)^3}-2\frac{d_j^2d_k^2d_\ell}{(\alpha N)^3}-2\frac{d_jd_k^2d_\ell^2}{(\alpha N)^3}\Big]-\frac{1}{6}\left[\frac{d_i^3d_j^3}{(\alpha N)^3}+\frac{d_j^3d_k^3}{(\alpha N)^3}\right.\\
&\quad\left.\left.+\frac{d_k^3d_\ell^3}{(\alpha N)^3}+3\frac{d_i^2d_j^3d_k}{(\alpha N)^3}+3\frac{d_id_j^3d_k^2}{(\alpha N)^3}+3\frac{d_i^2d_j^2d_kd_\ell}{(\alpha N)^3}+3\frac{d_id_jd_k^2d_\ell^2}{(\alpha N)^3}+3\frac{d_j^2d_k^3d_\ell}{(\alpha N)^3}+3\frac{d_jd_k^3d_\ell^2}{(\alpha N)^3}\right]\right\}\\
&= (N-1)(N-2)(N-3)-3N^2\frac{\langle d\rangle^2}{\alpha}+2N\frac{\langle d\rangle^3}{\alpha^2}-\frac{3}{2}\frac{\langle d^2\rangle^2}{\alpha^3}-2\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^3}-\frac{\langle d\rangle^4}{\alpha^3}\\
&\quad\frac{3}{2}N\frac{\langle d^2\rangle^2}{\alpha^2}+2N\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}+N\frac{\langle d\rangle^4}{\alpha^2}-4\frac{\langle d^2\rangle^2\langle d\rangle}{\alpha^3}-2\frac{\langle d^2\rangle\langle d\rangle^3}{\alpha^3}-\frac{1}{2}\frac{\langle d^3\rangle^2}{\alpha^3}-2\frac{\langle d^2\rangle\langle d^2\rangle\langle d\rangle}{\alpha^3}-\frac{\langle d^2\rangle^2\langle d\rangle^2}{\alpha^3}
\end{aligned}\tag{D 15}$$

and

$$\begin{aligned}
&\frac{1}{N}\sum_{[ijk\ell]}\langle\delta_{c_{ij},0}\delta_{c_{jk},0}\delta_{c_{k\ell},0}\delta_{c_{\ell i},0}\rangle\\
&= \frac{1}{N}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4}\prod_{\mu}\langle e^{i\xi_i^\mu(\xi_j^\mu\omega+\xi_\ell^\mu\omega''')+\xi_k^\mu(\xi_j^\mu\omega'+\xi_\ell^\mu\omega'')}\rangle\\
&= \frac{1}{N}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4}\prod_{\mu}\left\{\frac{d_i}{\alpha N}\langle e^{i(\xi_j^\mu\omega+\xi_\ell^\mu\omega''')+i\xi_k^\mu(\xi_j^\mu\omega'+\xi_\ell^\mu\omega'')}\rangle + \left(1-\frac{d_i}{\alpha N}\right)\langle e^{i\xi_k^\mu(\xi_j^\mu\omega'+\xi_\ell^\mu\omega'')}\rangle\right\}\\
&= \frac{1}{N}\sum_{[ijk\ell]}\int_{-\pi}^{\pi}\frac{d\omega d\omega' d\omega'' d\omega'''}{16\pi^4}\left(1+\frac{d_id_j}{(\alpha N)^2}(e^{i\omega}-1)+\frac{d_jd_k}{(\alpha N)^2}(e^{i\omega'}-1)+\frac{d_kd_\ell}{(\alpha N)^2}(e^{i\omega''}-1)\right.\\
&\quad+\frac{d_id_\ell}{(\alpha N)^2}(e^{i\omega'''}-1)+\frac{d_id_jd_k}{(\alpha N)^3}(e^{i(\omega+\omega')}-e^{i\omega}-e^{i\omega'}+1)+\frac{d_id_kd_\ell}{(\alpha N)^3}(e^{i(\omega''+\omega''')}-e^{i\omega''}-e^{i\omega'''}+1)\\
&\quad+\frac{d_id_jd_\ell}{(\alpha N)^3}(e^{i(\omega+\omega''')}-e^{i\omega}-e^{i\omega'''}+1)+\frac{d_jd_kd_\ell}{(\alpha N)^3}(e^{i(\omega'+\omega'')}-e^{i\omega'}-e^{i\omega''}+1)\\
&\quad+\frac{d_id_jd_kd_\ell}{(\alpha N)^4}(e^{i(\omega+\omega'+\omega''+\omega''')}-e^{i(\omega+\omega')}-e^{i(\omega''+\omega''')}-e^{i(\omega+\omega''')}-e^{i(\omega'+\omega'')}\\
&\quad\left.+e^{i\omega}+e^{i\omega'}+e^{i\omega''}+e^{i\omega'''}-1)\right)^{\alpha N}\\
&= \frac{1}{N}\sum_{[ijk\ell]}\left\{1-\frac{d_id_j}{\alpha N}-\frac{d_jd_k}{\alpha N}-\frac{d_kd_\ell}{\alpha N}-\frac{d_id_\ell}{\alpha N}+\frac{d_id_jd_k}{(\alpha N)^2}+\frac{d_id_kd_\ell}{(\alpha N)^2}+\frac{d_id_jd_\ell}{(\alpha N)^2}+\frac{d_jd_kd_\ell}{(\alpha N)^2}\right.\\
&\quad-\frac{d_id_jd_kd_\ell}{(\alpha N)^3}-\frac{1}{2}\left[\frac{d_i^2d_j^2}{(\alpha N)^3}+\frac{d_j^2d_k^2}{(\alpha N)^3}+\frac{d_k^2d_\ell^2}{(\alpha N)^3}+\frac{d_i^2d_\ell^2}{(\alpha N)^3}+2\frac{d_id_j^2d_k}{(\alpha N)^3}+2\frac{d_id_jd_kd_\ell}{(\alpha N)^3}+2\frac{d_i^2d_jd_\ell}{(\alpha N)^3}\right.
\end{aligned}$$

$$+2\frac{d_i d_k^2 d_\ell}{(\alpha N)^3}+2\frac{d_i d_j d_k d_\ell}{(\alpha N)^3}+2\frac{d_i d_k d_\ell^2}{(\alpha N)^3}\Bigg]+\frac{1}{2}\Bigg[\frac{d_i^2 d_j^2}{(\alpha N)^2}+\frac{d_j^2 d_k^2}{(\alpha N)^2}+\frac{d_k^2 d_\ell^2}{(\alpha N)^2}+\frac{d_i^2 d_\ell^2}{(\alpha N)^2}+2\frac{d_i d_j^2 d_k}{(\alpha N)^2}$$
$$+2\frac{d_i d_j d_k d_\ell}{(\alpha N)^2}+2\frac{d_i^2 d_j d_\ell}{(\alpha N)^2}+2\frac{d_j d_k^2 d_\ell}{(\alpha N)^2}+2\frac{d_i d_j d_k d_\ell}{(\alpha N)^2}+2\frac{d_i d_k d_\ell^2}{(\alpha N)^2}-2\frac{d_i^2 d_j^2 d_k}{(\alpha N)^3}-2\frac{d_i^2 d_j^2 d_\ell}{(\alpha N)^3}$$
$$-2\frac{d_i^2 d_j d_k d_\ell}{(\alpha N)^3}-2\frac{d_i d_j^2 d_k d_\ell}{(\alpha N)^3}-2\frac{d_i d_j^2 d_k^2}{(\alpha N)^3}-2\frac{d_i d_j d_k^2 d_\ell}{(\alpha N)^3}-2\frac{d_i d_j^2 d_k d_\ell}{(\alpha N)^3}-2\frac{d_j^2 d_k^2 d_\ell}{(\alpha N)^3}-2\frac{d_i d_j d_k^2 d_\ell}{(\alpha N)^3}$$
$$-2\frac{d_i d_k^2 d_\ell^2}{(\alpha N)^3}-2\frac{d_i d_j d_k d_\ell^2}{(\alpha N)^3}-2\frac{d_j d_k^2 d_\ell^2}{(\alpha N)^3}-2\frac{d_i^2 d_j d_k d_\ell}{(\alpha N)^3}-2\frac{d_i^2 d_k d_\ell^2}{(\alpha N)^3}-2\frac{d_i^2 d_j d_\ell^2}{(\alpha N)^3}-2\frac{d_i d_j d_k d_\ell^2}{(\alpha N)^3}\Bigg]$$
$$-\frac{1}{6}\Bigg[\frac{d_i^3 d_j^3}{(\alpha N)^3}+\frac{d_j^3 d_k^3}{(\alpha N)^3}+\frac{d_k^3 d_\ell^3}{(\alpha N)^3}+\frac{d_i^3 d_\ell^3}{(\alpha N)^3}+3\frac{d_i^2 d_j^3 d_k}{(\alpha N)^3}+3\frac{d_i d_j^3 d_k^2}{(\alpha N)^3}+3\frac{d_i^2 d_j^2 d_k d_\ell}{(\alpha N)^3}$$
$$+3\frac{d_i d_j d_k^2 d_\ell^2}{(\alpha N)^3}+3\frac{d_i^3 d_j^2 d_\ell}{(\alpha N)^3}+3\frac{d_i^3 d_j d_\ell^2}{(\alpha N)^3}+3\frac{d_i^2 d_j d_k^3}{(\alpha N)^3}+3\frac{d_j d_k^3 d_\ell^2}{(\alpha N)^3}+3\frac{d_i d_j^2 d_k^2 d_\ell}{(\alpha N)^3}+3\frac{d_i^2 d_j d_k d_\ell^2}{(\alpha N)^3}+3\frac{d_i d_k^2 d_\ell^3}{(\alpha N)^3}+3\frac{d_i^2 d_k d_\ell^3}{(\alpha N)^3}\Bigg]\Bigg\}$$
$$=(N-1)(N-2)(N-3)-4N^2\frac{\langle d\rangle^2}{\alpha}+4N\frac{\langle d\rangle^3}{\alpha^2}-2\frac{\langle d^2\rangle^2}{\alpha^3}-\frac{\langle d\rangle^4}{\alpha^3}-4\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^3}+2N\frac{\langle d^2\rangle^2}{\alpha^2}-2\frac{\langle d\rangle^4}{\alpha^3}+4N\frac{\langle d^2\rangle\langle d\rangle^2}{\alpha^2}+2N\frac{\langle d\rangle^4}{\alpha^2}$$
$$-8\frac{\langle d^2\rangle^2\langle d\rangle}{\alpha^3}-8\frac{\langle d^2\rangle\langle d\rangle^3}{\alpha^3}-\frac{2}{3}\frac{\langle d^3\rangle^2}{\alpha^3}-4\frac{\langle d^3\rangle\langle d^2\rangle\langle d\rangle}{\alpha^3}-2\frac{\langle d^2\rangle^2\langle d\rangle^2}{\alpha^3}. \quad \text{(D 16)}$$

References

1. Hakes L, Pinney JW, Robertson DL, Lovel SC. 2008 Protein–protein interaction networks and biology—what's the connection? *Nat. Biotechnol.* **26** 69–72. (doi:10.1038/nbt0108-69)
2. Han JDJ, Dupuy D, Bertin N, Cusick ME, Vidal M. 2005 Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* **23**, 839–844. (doi:10.1038/nbt1116)
3. De Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, Wiuf C, Stumpf MPH. 2006 The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol.* **4**, 39. (doi:10.1186/1741-7007-4-39)
4. Fernandes LP, Annibale A, Kleinjung J, Coolen ACC, 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)
5. Lee SH, Kim PJ, Jeong H. 2006 Statistical properties of sampled networks. *Phys. Rev. E* **73**, 016102. (doi:10.1103/PhysRevE.73.016102)
6. Stumpf MPH, Wiuf C. 2005 Sampling properties of random graphs: the degree distribution. *Phys. Rev. E* **72**, 036118. (doi:10.1103/PhysRevE.72.036118)
7. Stumpf MPH, Wiuf C, May RM. 2005 Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224. (doi:10.1073/pnas.0501179102)
8. Viger F, Barrat A, Dall' Asta L, Zhang CH, Kolaczyk ED. 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 IM, Eliazar II. 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 ACC. 2011 What you see is not what you get: how sampling affects macroscopic features of biological networks. *Interface Focus* **1**, 836–856. (doi:10.1098/rsfs.2011.0050)
11. Chang X, Xu T, Li Y, Wang K. 2013 Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Sci. Rep.* **3**, 1691. (doi:10.1038/srep01691)
12. Newman MEJ, Strogatz SH, Watts DJ. 2001 Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118. (doi:10.1103/PhysRevE.64.026118)
13. Ivanic J, Wallqvist A, Reifman J. 2008 Probing the extent of randomness in protein interaction networks. *PLoS Comput. Biol.* **4**, e1000114. (doi:10.1371/journal.pcbi.1000114)
14. Talavera D, Williams SG, Norris MGS, Robertson DL, Lovell SC. 2012 Evolvability of yeast protein–protein interaction interfaces. *J. Mol. Biol.* **419**, 387–396. (doi:10.1016/j.jmb.2012.03.021)
15. Simonis N *et al.* 2008 Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. *Nat. Methods* **6**, 47–54. (doi:10.1038/nmeth.1279)
16. Parrish JR *et al.* 2007 A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol.* **8**, R130. (doi:10.1186/gb-2007-8-7-r130)
17. Arifuzzaman M *et al.* 2006 Large-scale identification of protein–protein interaction of *Escherichia coli* K-12. *Genome Res.* **16**, 686–691. (doi:10.1101/gr.4527806)
18. Rain JC *et al.* 2001 The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215. (doi:10.1038/35051615)
19. Rual J-FF *et al.* 2005 Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178. (doi:10.1038/nature04209)
20. Stelzl U *et al.* 2005 A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968. (doi:10.1016/j.cell.2005.08.029)
21. Ewing RM *et al.* 2007 Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89. (doi:10.1038/msb4100134)
22. Shimoda Y, Shinpo S, Kohara M, Nakamura Y, Tabata S, Sato S. 2008 A large-scale analysis of protein–protein interactions in the nitrogen-fixing bacterium *Mesorhizobium loti*. *DNA Res.* **15**, 13–23. (doi:10.1093/dnares/dsm028)
23. Lacount DJ *et al.* 2005 A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **438**, 103–107. (doi:10.1038/nature04104)
24. Uetz P *et al.* 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627. (doi:10.1038/35001009)
25. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574. (doi:10.1073/pnas.061034498)
26. Ho Y *et al.* 2002 Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183. (doi:10.1038/415180a)

27. Gavin AC *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147. (doi:10.1038/415141a)
28. Gavin ACC *et al.* 2006 Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636. (doi:10.1038/nature04532)
29. Krogan NJ *et al.* 2006 Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643. (doi:10.1038/nature04670)
30. Collins SR *et al.* 2007 Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450. (doi:10.1074/mcp.M600381-MCP200)
31. Sato S, Shimoda Y, Muraki A, Kohara M, Nakamura Y, Tabata S. 2007 A large-scale protein–protein interaction analysis in *Synechocystis* sp. PCC6803. *DNA Res.* **14**, 207–216. (doi:10.1093/dnares/dsm021)
32. Titz B *et al.* 2008 The binary protein interactome of treponema pallidum—the syphilis spirochete. *PLoS ONE* **3**, e2292. (doi:10.1371/journal.pone.0002292)
33. Albert R, Barabassi AL. 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97. (doi:10.1103/RevModPhys.74.47)
34. Barabasi AL, Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
35. Dorogovtsev SN, Mendes JF. 2003 *Evolution of networks*. Oxford, UK: Oxford University Press.
36. Junker BH, Schreiber F. 2008 *Analysis of biological networks*. New York, NY: Wiley.
37. Agliari E, Annibale A, Barra A, Coolen ACC, Tantari D. 2013 Immune networks: multitasking capabilities near saturation. *J. Phys. A: Math. Theor.* **46**, 415003. (doi:10.1088/1751-8113/46/41/415003)
38. Sollich P, Tantari D, Annibale A, Barra A. 2014 Extensive parallel processing on scale-free networks. *Phys. Rev. Lett.* **113**, 238106. (doi:10.1103/PhysRevLett.113.238106)
39. Roberts ES, Coolen ACC. 2014 Random graph ensembles with many short loops. *ESAIM: Proc. Surv.* **47**, 97–115. (doi:10.1051/proc/201447006)
40. Abramowitz M, Stegun IA. 1972 *Handbook of mathematical functions*. New York, NY: Dover.
41. Woodsmith J, Stelzl U. 2014 Studying post-translational modifications with protein interaction networks. *Curr. Opin. Struct. Biol.* **24**, 34–44. (doi:10.1016/j.sbi.2013.11.009)
42. Kelly WP, Stumpf MP. 2102 Assessing coverage of protein interaction data using capture–recapture models. *Bull. Math. Biol.* **74**, 356–374. (doi:10.1007/s11538-011-9680-2)
43. Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE. 2011 A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* **4**, rs8. (doi:10.1126/scisignal.2001699)
44. Havuginama PC *et al.* 2012 A census of human soluble protein complexes. *Cell* **150**, 1068–1081. (doi:10.1016/j.cell.2012.08.011)