ESAIM: PROCEEDINGS AND SURVEYS, **December 2014, Vol. 47, p. 97-115**
F. Abergel, M. Aiguier, D. Challet, P.-H. Cournède, G. Faÿ, P. Lafitte, Editors

RANDOM GRAPH ENSEMBLES WITH MANY SHORT LOOPS

ES Roberts[1, 2] and ACC Coolen[1, 3]

Abstract. Networks observed in the real world often have many short loops. This violates the tree-like assumption that underpins the majority of random graph models and most of the methods used for their analysis. In this paper we sketch possible research routes to be explored in order to make progress on networks with many short loops, involving old and new random graph models and ideas for novel mathematical methods. We do not present conclusive solutions of problems, but aim to encourage and stimulate new activity and in what we believe to be an important but under-exposed area of research. We discuss in more detail the Strauss model, which can be seen as the 'harmonic oscillator' of 'loopy' random graphs, and a recent exactly solvable immunological model that involves random graphs with extensively many cliques and short loops.

Résumé. Les réseaux observés dans la Nature ont souvent des cycles courts. Ceci contredit le postulat de hiérarchie sur lequel se base la majorité des modèles de réseaux aléatoires et la plupart des méthodes utilisées pour leur analyse. Dans cet article, nous esquissons des directions de recherches possibles, afin de progresser sur les réseaux contenant beaucoup de cycles courts, faisant appel à des modèles de réseaux aléatoires éprouvés ou nouveaux, et des idées pour de nouvelles méthodes mathématiques. Nous ne présentons pas de solutions définitives, mais notre but est d'encourager et de stimuler de nouveaux travaux dans ce que nous croyons être une direction de recherche importante, bien que insuffisamment explorée. Nous discutons en détail le modèle de Strauss, qui peut être considéré comme 'l'oscillateur harmonique' des réseaux aléatoires 'à boucles', ainsi qu'un modèle immunologique soluble exactement qui comporte des réseaux aléatoires avec de nombreux cliques et cycles courts.

1. Motivation and background

Tailored random graph ensembles, whose statistical features are sculpted to mimic those observed in a given application domain, provide a rational framework within which we can understand and quantify topological patterns observed in real life networks. Most analytical approaches for studying such networks, or for studying processes for which they provide the interaction infrastructure, assume explicitly or implicitly that they are locally tree-like. It permits, usually after further mathematical manipulations and in leading orders in the system size, factorisation across nodes and/or links. This in turn allows for the crucial combinatorial sums over all possible graphs with given constraints to be done analytically in the relevant calculations. However, real-world networks - for example protein-protein interaction networks (PPIN), immune networks, synthetic communication networks, or social networks - tend to have a significant number of short loops. It is widely

[1] Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, London SE1 1UL, United Kingdom

[2] Randall Division of Cell and Molecular Biophysics, King's College London, New Hunts House, London SE1 1UL, United Kingdom

[3] London Institute for Mathematical Sciences, 35a South St, Mayfair, London W1K 2XF, United Kingdom

Article published online by EDP Sciences and available at http://www.esaim-proc.org or http://dx.doi.org/10.1051/proc/201447006

accepted that the abundance of short loops in, for example, PPINs is intrinsic to the function of these networks. The authors of [1] suggested that the stability of a biological network is highly correlated with the relative abundance of motifs (e.g. triangles). The authors of [2] and [3] observed an apparent relationship between short cycles in gene-regulation networks and the system's response to stress and heat-shock. A highly cited paper [4] went as far as to propose that motifs (e.g. triangles) are the basic building blocks of most networks. Similarly, in many-particle physics we know that lattice models are difficult to solve, mainly because of the many short loops that exist between the interacting variables on the lattice vertices [5]; calculating the free energy of statistical mechanical models on tree-like lattices, in contrast, is relatively straightforward. It is evident that incorporating constraints relating to the statistics of short loops into the specifications of random graph ensembles is important for this branch of research to be able to more closely align with the needs of practitioners in the bioinformatics and network science communities.

In this paper we review the analytical techniques that are presently available to model and analyse ensembles of random graphs with extensive numbers of short loops, and we discuss possible future research routes and ideas. We will use the terms 'network' and 'graph' without distinction. We start with a description of tailored random graph ensembles, and argue why graphs with short loops should become the focus of research, to increase their applicability to real-world problems and for mathematical and methodological reasons. We then discuss the simplest network model with short loops that even after some forty years we still cannot solve satisfactorily: the Strauss model [6], which is the archetypical ensemble of finitely connected random graphs with controlled number of triangles. We describe some new results on the entropy of this ensemble, continuing the work of e.g. [7], as well as an as yet unexplored approach based on combining graph spectral analysis with the replica method. The next section discusses some recent results on the statistical mechanics of an immune network model [8]; in spite of its many short loops, this model could quite unexpectedly be solved analytically with the finite connectivity replica method [11–13], suggesting a possible and welcome new mechanism for analysing more general families of 'loopy' graphs. We then show how the model of [8] can indeed be used in other scenarios where short loops in networks play a functional role, and we work out in more detail its application in the context of factor graph representations of protein-protein interaction networks.

2. Modelling with random graphs – the importance of short loops

Networks are powerful conceptual tools in the modelling of real-world phenomena. In large systems of interacting variables they specify which pairs can interact, which leads to convenient visualisations and reduces the complexity of the problem. Random graphs serve as proxies for interaction networks that are (fully or partially) observed or built in biology, physics, economics, or engineering. Random graphs allow us to analytically solve statistical mechanical models of the processes for which the networks represent the infrastructure, by appropriate averaging of generating functions of observables over all 'typical' interaction networks. Or they can be used as 'null models' to quantify the statistical relevance of topological measurements which are taken from observed networks. In all cases it is vital that the random graphs actually resemble the true real-world networks in a quantitatively controllable way.

In this paper we limit ourselves, for simplicity, to nondirected graphs without self-links. Generalisation of the various models and arguments to directed and/or self-interacting graphs is usually straightforward. A nondirected simple N-node graph is characterised by its nodes $i \in \{1, \ldots, N\}$ and by the values of $\frac{1}{2}N(N-1)$ link variables $c_{ij} \in \{0, 1\}$. Here $c_{ij} = 1$ if the nodes (i, j) are connected by a link, and $c_{ij} = 0$ otherwise. We always have $c_{ij} = c_{ji}$ (since our graphs are nondirected) and $c_{ii} = 0$ (since our graphs are simple), for all $i, j \in \{1, \ldots, N\}$. We will denote the set of all such graphs as G. A random graph ensemble is defined by a probability measure $p(\boldsymbol{c})$ on G.

2.1. **Tailored random graph ensembles**

If we wish to use random graph ensembles to study real-world phenomena, it is vital that our measure $p(\boldsymbol{c})$ favours graphs that mimic those in our application domain. For instance, it makes no sense to use Erdös-Rényi

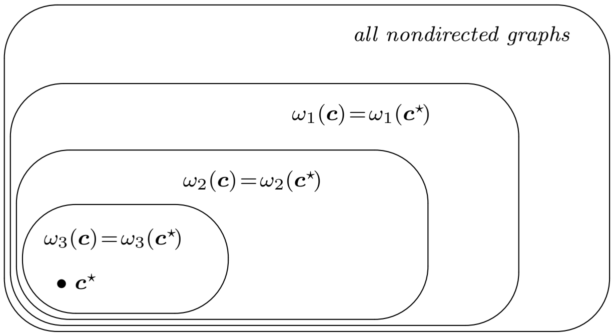
all nondirected graphs
$\omega_1(\boldsymbol{c})=\omega_1(\boldsymbol{c}^\star)$
$\omega_2(\boldsymbol{c})=\omega_2(\boldsymbol{c}^\star)$
$\omega_3(\boldsymbol{c})=\omega_3(\boldsymbol{c}^\star)$
$\bullet\ \boldsymbol{c}^\star$

Figure 1. Tailoring of random graph ensembles such that the generated graphs will mimic the features of an observed graph $\boldsymbol{c}^\star$, via successive imposition of the values of L chosen observables $\{\omega_1(\boldsymbol{c}),\ldots,\omega_L\}$. Subject to these imposed values, which can be built in as hard constraints (to be reproduced by *all* $\boldsymbol{c}$ with $p(\boldsymbol{c})>0$), or as soft constraints (to be reproduced on average), the measure $p(\boldsymbol{c})$ is defined by maximising the Shannon entropy $S[p]=-\sum_{\boldsymbol{c}\in G}p(\boldsymbol{c})\log p(\boldsymbol{c})$. The smaller the set of graphs that satisfy $\omega_\ell(\boldsymbol{c})=\omega_\ell(\boldsymbol{c}^\star)$ for all $\ell\leq L$, the more our random graphs are expected to resemble $\boldsymbol{c}^\star$.

graphs [14] as null models against which to test occurrence frequencies of motifs in biological networks: almost any measurement will come out as significant, simply because our yardstick is not realistic. Tailored random graph ensembles [15–18] involve measures $p(\boldsymbol{c})$ that are constructed such that specified topological features of the generated graphs $\boldsymbol{c}$ will systematically resemble those of a given real-world graph $\boldsymbol{c}^\star\in G$. To construct such measures one first defines the set of L observables $\{\omega_1(\boldsymbol{c}),\ldots,\omega_L(\boldsymbol{c})\}$ whose values the random graphs $\boldsymbol{c}$ is supposed to inherit from $\boldsymbol{c}^\star$. One then defines $p(\boldsymbol{c})$ as the maximum entropy ensemble on G, subject to the imposition of the values of $\{\omega_1(\boldsymbol{c}),\ldots,\omega_L(\boldsymbol{c})\}$, with the Shannon entropy [19] $S[p]=-\sum_{\boldsymbol{c}\in G}p(\boldsymbol{c})\log p(\boldsymbol{c})$. This can be done via hard constraints, where *each* $\boldsymbol{c}\in G$ with $p(\boldsymbol{c})>0$ must have the specified values, or via soft constraints, where our random graphs will be described by an exponential ensemble and exhibit the specified values of the L observables only *on average* (see also [42]). Based on a constrained maximum entropy calculation, we know that:

$$\text{hard constrained ensembles}: \qquad p_h(\boldsymbol{c})=Z_h^{-1}\prod_{\ell\leq L}\delta_{\omega_\ell(\boldsymbol{c}),\omega_\ell(\boldsymbol{c}^\star)} \tag{1}$$

$$\text{soft constrained ensembles}: \qquad p_s(\boldsymbol{c})=Z_s^{-1}\,\mathrm{e}^{\sum_{\ell=1}^L\hat{\omega}_\ell\omega_\ell(\boldsymbol{c})}, \qquad \sum_{\boldsymbol{c}\in G}p_s(\boldsymbol{c})\omega_\ell(\boldsymbol{c})=\omega_\ell(\boldsymbol{c}^\star)\quad\forall\ell\leq L \tag{2}$$

The maximum entropy formulation is essential to make sure one does not introduce any unwanted bias into our tailored graphs; we want to build in the features of $\boldsymbol{c}^\star$ and nothing else.

This leads to the question of which would be sensible choices for the observables $\{\omega_1(\boldsymbol{c}),\ldots,\omega_L\}$ to carry over from $\boldsymbol{c}^\star$ to our ensemble? Sensible choices are those for which we can do the relevant calculations, and for which the Shannon entropy of $p(\boldsymbol{c})$ would be smallest[1]. The calculations we might wish to do usually relate to stochastic processes for variables placed on the nodes of the graph, and the crucial question is whether the relevant combinatorial sums over all graphs generated from $p(\boldsymbol{c})$ can be carried out analytically. In equilibrium systems we would want to calculate the typical free energy per degree of freedom, averaged over the random graph ensemble, for Hamiltonians of the form $H(\boldsymbol{\sigma})=-\sum_{i<j}c_{ij}J_{ij}\sigma_i\sigma_j$. The replica method [9–13] then leads

[1] Since the effective number of graphs in an ensemble $p(\boldsymbol{c})$ is given by $\mathcal{N}[p]=\exp(S[p])$, the Shannon entropy can be interpreted as a measure of the size of the smallest box in Figure 1. The smaller this box, the more information on $\boldsymbol{c}^\star$ has been carried over to our graph ensemble.

us for hard-constrained ensembles of finitely connected graphs to a combinatorial problem of the following form:

$$\overline{\mathrm{e}^{-\beta\sum_{\alpha=1}^{n}H(\boldsymbol{\sigma}^{\alpha})}} = \frac{\sum_{\boldsymbol{c}\in G}\mathrm{e}^{\sum_{i<j}c_{ij}A_{ij}}\prod_{\ell\leq L}\delta_{\omega_\ell(\boldsymbol{c}),\omega_\ell(\boldsymbol{c}^\star)}}{\sum_{\boldsymbol{c}\in G}\prod_{\ell\leq L}\delta_{\omega_\ell(\boldsymbol{c}),\omega_\ell(\boldsymbol{c}^\star)}}, \qquad A_{ij}=\beta J_{ij}\sum_{\alpha=1}^{n}\sigma_i^\alpha\sigma_j^\alpha \tag{3}$$

where the overbar $\overline{\cdots}$ indicates averaging over the measure $p(\boldsymbol{c})$ of our random graph ensemble. Similarly, in dynamical studies based on generating functional analysis [20–22] we would be required to evaluate

$$\overline{\mathrm{e}^{-\mathrm{i}\sum_{it}\hat{h}_i(t)\sum_j c_{ij}J_{ij}\sigma_j(t)}} = \frac{\sum_{\boldsymbol{c}\in G}\mathrm{e}^{\sum_{i<j}c_{ij}A_{ij}}\prod_{\ell\leq L}\delta_{\omega_\ell(\boldsymbol{c}),\omega_\ell(\boldsymbol{c}^\star)}}{\sum_{\boldsymbol{c}\in G}\prod_{\ell\leq L}\delta_{\omega_\ell(\boldsymbol{c}),\omega_\ell(\boldsymbol{c}^\star)}}, \qquad A_{ij}=-\mathrm{i}J_{ij}\sum_t[\hat{h}_i(t)\sigma_j(t)+\hat{h}_j(t)\sigma_i(t)] \tag{4}$$

In both cases we see that our observables $\omega_\ell(\boldsymbol{c})$ should be chosen such that sums of the form $\sum_{\boldsymbol{c}\in G}\delta_{\boldsymbol{\omega},\boldsymbol{\omega}(\boldsymbol{c})}\mathrm{e}^{\sum_{i<j}c_{ij}A_{ij}}$ are analytically tractable. Setting $A_{ij}=0$ for all (i,j) gives us *en passant* the value of the Shannon entropy, which for hard-constrained ensembles (1) becomes $S[p]=\log\sum_{\boldsymbol{c}\in G}\prod_{\ell\leq L}\delta_{\omega_\ell(\boldsymbol{c}),\omega_\ell(\boldsymbol{c}^\star)}$. It turned out that these summations over all graphs are analytically feasible [15–18], in leading orders in N, for observables such as

$$\bar{k}(\boldsymbol{c})=\frac{1}{N}\sum_{ij}c_{ij}, \qquad p(k|\boldsymbol{c})=\frac{1}{N}\sum_i\delta_{k,\sum_j c_{ij}}, \qquad W(k,k'|\boldsymbol{c})=\frac{1}{\bar{k}N}\sum_{ij}c_{ij}\ \delta_{k,\sum_r c_{ir}}\delta_{k',\sum_r c_{jr}} \tag{5}$$

2.2. **The problem of short loops**

To quantify the extent to which real-world networks $\boldsymbol{c}^\star$ can be approximated by random graphs that share with $\boldsymbol{c}^\star$ the values of the average degree $\bar{k}(\boldsymbol{c}^\star)$, or the degree distribution $p(k|\boldsymbol{c}^\star)$, or the joint distribution $W(k,k'|\boldsymbol{c}^\star)$ of the degrees of connected node pairs, it is helpful to study systems for which alternative exact solutions or reliable simulation data are available. Using the methodology of [15–18], we can, for instance, calculate with the replica method the critical temperatures of Ising systems on random graphs with Hamiltonian $H(\boldsymbol{\sigma})=-\sum_{i<j}c_{ij}\sigma_i\sigma_j$. We choose our random graph ensembles to be increasingly constrained in the sense of Figure 1:

$$p_A(\boldsymbol{c}): \qquad \text{maximum entropy ensemble with imposed } \ \bar{k}(\boldsymbol{c}^\star)=\sum_k k\ p(k|\boldsymbol{c}^\star) \tag{6}$$

$$p_B(\boldsymbol{c}): \qquad \text{maximum entropy ensemble with imposed } \ p(k|\boldsymbol{c}^\star)\ \ \forall k\geq 0 \tag{7}$$

$$p_C(\boldsymbol{c}): \qquad \text{maximum entropy ensemble with imposed } \ p(k|\boldsymbol{c}^\star) \ \text{ and } \ W(k,k'|\boldsymbol{c}^\star)\ \ \forall k,k'\geq 0 \tag{8}$$

In Figure 2 we compare the critical temperatures for Ising systems on random graphs tailored according to (6,7,8) to the true critical temperature values of the approximated finitely connected graphs $\boldsymbol{c}^\star$, for cubic lattices and for so-called 'small world' lattices. While constraining only the average degree (A) is clearly insufficient, we see that constraining the degree distribution (B) brings us already closer to the true values of the transition temperatures. Adding the joint degree statistics of connected nodes (C) gets the critical temperatures nearly right for the small-world graphs, but fails to improve $T_c(d)$ for the regular lattices. Since in regular cubic lattices one simply has $W(k,k'|\boldsymbol{c}^\star)=kk'2^{-2d}p(k|\boldsymbol{c}^\star)p(k'|\boldsymbol{c}^\star)$, prescribing the values of $W(k,k'|\boldsymbol{c}^\star)$ here indeed gives no information that is not already contained in the degree distribution.

We conclude from Figure 2 that random graphs tailored on the basis of the observables (5) do capture valuable information, and give reasonable approximations of quantitative characteristics of stochastic processes that run on such graphs, but there is room for improvement. A hint at which would be an informative observable to add to $p(k|\boldsymbol{c})$ and $W(k,k'|\boldsymbol{c})$ in order to make our tailored random graphs more realistic approximations of $\boldsymbol{c}^\star$ is provided by comparison of the two case studies in Figure 2. A prominent difference between the topologies of cubic latices and small world graphs is the multiplicity of short loops. Cubic lattices have a finite number per node of loops of any even length, even in the limit $N\to\infty$, whereas in small world lattices the number of

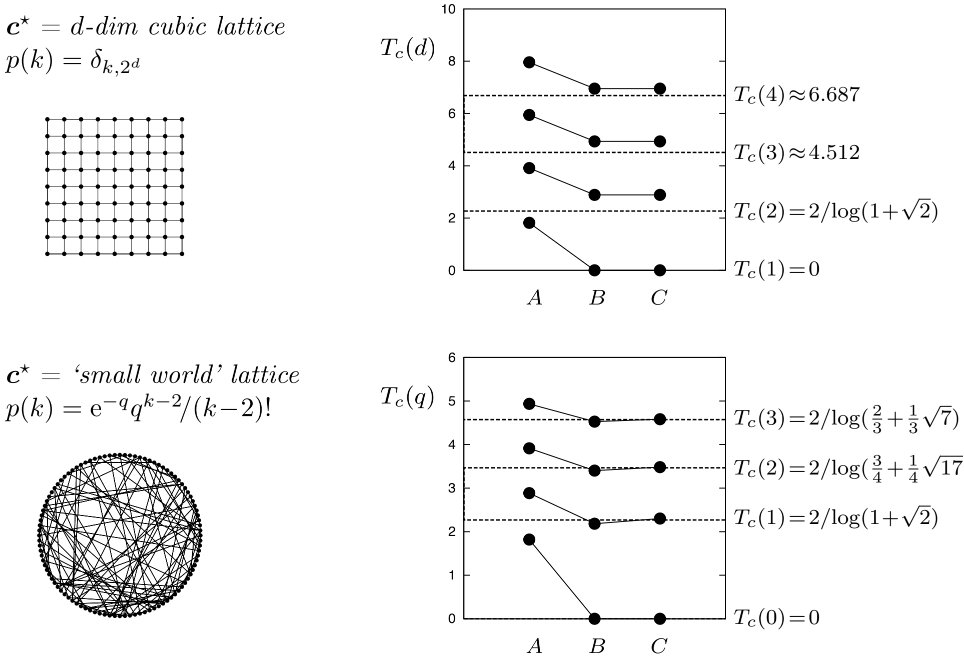

Figure 2. Critical temperatures of Ising systems on finitely connected lattices, calculated for random graph ensembles (6,7,8) which are tailored to either resemble d-dimensional cubic lattices $\boldsymbol{c}^\star$ (top, with $d = 1, 2, 3, 4$) or 'small world' lattices $\boldsymbol{c}^\star$ (bottom, Erdös-Rényi graph with average degree q superimposed upon a one-dimensional ring, with $q = 0, 1, 2, 3$). Connected markers: critical temperatures $T_c(d)$ and $T_c(q)$ calculated analytically for Ising models on the tailored random graphs. Dashed horizontal lines and corresponding values on the right: the true critical temperatures for Ising models on the lattices $\boldsymbol{c}^\star$. Transition temperatures are calculated analytically for the small world lattices [23] and for the cubic lattices with $d = 1, 2$ [5], and via numerical simulations [24] for cubic lattices with $d = 3, 4$.

short loops per node vanishes for $N \to \infty$. Explicit calculation shows that also the ensembles (6,7,8) typically generate locally tree-like graphs, and this explains why the critical temperatures $T_c(q)$ of the small world graphs are approximated very well, while those of the cubic lattices are not. By the same token, one can easily confirm that biological signalling networks, such as protein-protein interaction networks (PPIN) or gene regulation networks (GRN) have significantly more short loops than the typical graphs generated within the ensembles (6,7,8). See for example the data in Figure 3. The same is probably true for many social, economical and technological networks.

It appears that the next natural observables to be constrained in order to make our tailored random graphs more realistic must involve the number of short loops per node. Moreover, the realistic scaling regime is for this number to be finite, even for $N \to \infty$. However, all available methods for analysing stochastic processes on graphs (replica methods, cavity methods, belief and survey propagation, generating functional analysis), or for calculating ensemble entropies, all require implicitly or explicitly that the underlying topologies are locally tree-like. Apart from correction methods to handle small deviations from the tree-like assumption [27–29], there appears to be as yet no systematic method for doing the relevant combinatorial sums over all graphs $\boldsymbol{c} \in G$ in expressions such as (3,4) analytically when short loops are prevalent.

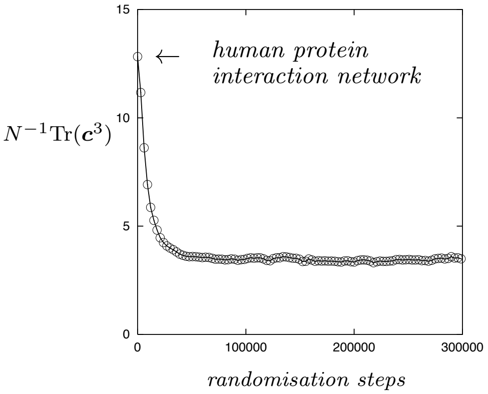

Figure 3. The effect on the number of triangles per node of randomising the human protein-protein interaction network (PPIN, taken from the HPRD database [25], with $N = 9463$ nodes) within the space of all graphs $\boldsymbol{c} \in G$ that have identical degrees and identical joint degree statistics $W(k,k|\boldsymbol{c})$ of connected nodes, generated using the method of [26]. Note that $N^{-1}\mathrm{Tr}(\boldsymbol{c}^3) = N^{-1}\sum_{ijk} c_{ij}c_{jk}c_{ki}$. Clearly, the human PPIN (the initial state, at steps=0) is atypical in this space, in that it has significantly more triangles per node than expected.

3. Strauss model – the ‘harmonic oscillator’ of loopy graphs

3.1. **Definitions**

We now turn to the simplest ensemble of finitely connected graphs with extensively many short loops. Since we have seen earlier that calculating graph ensemble entropies is usually a precursor to analysing processes on such graphs, we focus for now on how to determine Shannon entropies. Some tangential extensions have been explored in [41, 43, 44] and related papers. The Strauss model [6] is the maximum entropy soft-constrained random graph ensemble, with specified values of the average degree and the average number of triangles. It is defined via

$$p(\boldsymbol{c}) = Z^{-1}(u,g)\ \mathrm{e}^{u\sum_{ij}c_{ij}+g\sum_{ijk}c_{ij}c_{jk}c_{ki}}, \qquad Z(u,g) = \sum_{\boldsymbol{c}\in G}\mathrm{e}^{u\sum_{ij}c_{ij}+g\sum_{ijk}c_{ij}c_{jk}c_{ki}} \tag{9}$$

The ensemble parameters u and g are used to control the relevant ensemble averages, via the identities

$$\langle k\rangle \;=\; \sum_{\boldsymbol{c}\in G}p(\boldsymbol{c})\,\frac{1}{N}\sum_{ij}c_{ij} = \frac{\partial}{\partial u}\phi(u,g) \tag{10}$$

$$\langle m\rangle \;=\; \sum_{\boldsymbol{c}\in G}p(\boldsymbol{c})\,\frac{1}{N}\sum_{ij}c_{ij}c_{jk}c_{ki} = \frac{\partial}{\partial g}\phi(u,g) \tag{11}$$

with the following generating function whose evaluation requires that we do analytically the sum over all graphs $\boldsymbol{c} \in G$:

$$\phi(u,g) = \ \frac{1}{N}\log Z(u,g) = \frac{1}{N}\log\Big[\sum_{\boldsymbol{c}\in G}\mathrm{e}^{u\sum_{ij}c_{ij}+g\sum_{ijk}c_{ij}c_{jk}c_{ki}}\Big] \tag{12}$$

For this ensemble, we wish to calculate the Shannon entropy $S = -\sum_{\boldsymbol{c}\in G} p(\boldsymbol{c})\log p(\boldsymbol{c})$ in leading order in N. This follows directly from $\phi(u,g)$ since

$$\begin{aligned} S &= -\Big[\log Z(u,g) - \frac{1}{Z(u,g)}\sum_{\boldsymbol{c}} p(\boldsymbol{c})\log p(\boldsymbol{c})\Big] \\ &= \Big[1 - u\frac{\partial}{\partial u} - g\frac{\partial}{\partial g}\Big]\log Z(u,g) \;=\; N\left[\phi(u,g) - u\langle k\rangle - g\langle m\rangle\right] \end{aligned} \tag{13}$$

Upon setting g to be equal to zero, the ensemble (9) reduces to the Erdös-Rényi (ER) ensemble [14]

$$p_{\rm ER}(\boldsymbol{c}) = Z_{\rm ER}^{-1}(u)\ {\rm e}^{u\sum_{ij}c_{ij}}, \qquad \log Z_{\rm ER}(u) = \frac{N(N-1)}{2}\log({\rm e}^{2u}+1) \tag{14}$$

Differentiating and substituting in $u = -\frac{1}{2}\ln\left(\frac{1-p}{p}\right)$ then immediately leads us to

$$\bar{k} = \langle k\rangle\big|_{g=0} = \frac{1}{N}\frac{\rm d}{{\rm d}u}\log Z_{\rm ER}(u) = \frac{N-1}{1+{\rm e}^{-2u}} = (N-1)p \tag{15}$$

The parameter p can therefore be interpreted as the likelihood of having a link between any two nodes in the ER ensemble, so for finitely connected graphs $p = \mathcal{O}(N^{-1})$ and $u = -\frac{1}{2}\log N + \mathcal{O}(1)$. This connection with the ER ensemble suggests rewriting equation (12) as

$$\begin{aligned} \phi(u,g) &= \frac{1}{N}\log\sum_{\boldsymbol{c}\in G} p_{\rm ER}(\boldsymbol{c})\ {\rm e}^{g\sum_{ijk}c_{ij}c_{jk}c_{ki}} + \frac{1}{N}\log Z_{\rm ER}(u) \\ &= \frac{1}{2}(N-1)\log({\rm e}^{2u}+1) + \frac{1}{N}\log\sum_{r\geq 0} p(r|u){\rm e}^{gr} \end{aligned} \tag{16}$$

with

$$p(r|u) = \sum_{\boldsymbol{c}\in G} p_{\rm ER}(\boldsymbol{c})\ \delta_{r,\sum_{ijk}c_{ij}c_{jk}c_{ki}} \tag{17}$$

Hence the substance of the entropy calculation problem for the Strauss ensemble is mathematically equivalent to determining the moments of the distribution $p(r|u)$ of triangle counts in the Erdös-Rényi ensemble.

3.2. Simple approximation of the Shannon entropy of the Strauss model

We note that the average $\bar{r}(u)$ of the distribution (17) can be calculated easily and expressed in terms of the average degree $\bar{k}$ of the ER ensemble, giving

$$\bar{r}(u) = \sum_{r\geq 0} r p(r|u) = \sum_{\boldsymbol{c}\in G}\sum_{ijk} c_{ij}c_{jk}c_{ki} = N(N-1)(N-2)(1+{\rm e}^{-2u})^{-3} = \bar{k}^3 + \mathcal{O}(N^{-1}) \tag{18}$$

Here $\bar{k}$, which will differ from the average degree $\langle k\rangle$ of (9) as soon as $g > 0$, is related to the parameter u via the identity

$$u = -\frac{1}{2}\log[(N-1)/\bar{k} - 1] = \frac{1}{2}\log(\bar{k}/N) + \mathcal{O}(1/N) \tag{19}$$

As a first approximation, we can make the simple ansatz that (17) is a Poissonian distribution, which must then be given by $p(r|u) = {\rm e}^{-\bar{r}(u)}[\bar{r}(u)]^r/r!$. One does not expect this assumption to be valid exactly, but according

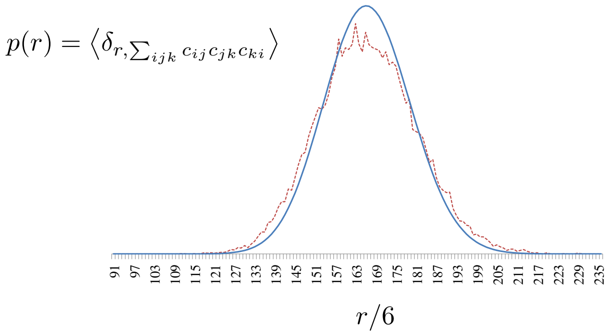

FIGURE 4. Red dotted line: distribution of triangle counts based on sampling 20,000 networks from an Erdös-Rényi (ER) ensemble with $N = 1000$ nodes and $\bar{k} = 10$. Dark blue solid line: a Poissonian distribution $p(r) = \mathrm{e}^{-\bar{r}}\bar{r}^r/r!$ with average number of triangles $\bar{r}$ identical to that measured in the simulated ER graphs. Although similar in shape, the observed triangle distribution $p(r)$ appears to decay to zero more slowly than the Poissonnian one, as r moves away from its average value.

to simulation data (see e.g. Figure 4), it is a reasonable initial step. It allows us to do the sum over $r \geq 0$ in (16), and find

$$\begin{aligned}\phi(u,g) &= \frac{1}{N}\bar{r}(u)(\mathrm{e}^g-1)+\frac{1}{2}(N-1)\log(\mathrm{e}^{2u}+1)\\ &= (N-1)(N-2)(\mathrm{e}^g-1)(1+\mathrm{e}^{-2u})^{-3}+\frac{1}{2}(N-1)\log(\mathrm{e}^{2u}+1) \qquad (20)\end{aligned}$$

We can now immediately work out (11) and (13):

$$\langle k\rangle = \frac{N-1}{1+\mathrm{e}^{-2u}}+6(N-1)(N-2)\mathrm{e}^{-2u}(\mathrm{e}^g-1)(1+\mathrm{e}^{-2u})^{-4} \qquad (21)$$

$$\langle m\rangle = \mathrm{e}^g(N-1)(N-2)(1+\mathrm{e}^{-2u})^{-3} \qquad (22)$$

$$S/N = (N-1)(N-2)(\mathrm{e}^g-1)(1+\mathrm{e}^{-2u})^{-3}+(N-1)u+\frac{1}{2}(N-1)\log(1+\mathrm{e}^{-2u})-u\langle k\rangle-g\langle m\rangle \qquad (23)$$

At $g=0$ we simply recover the equations of the ER model, with $\langle k\rangle=(N-1)/(1+\mathrm{e}^{-2u})$ and

$$g=0: \qquad \langle m\rangle=\langle k\rangle^3/N+\mathcal{O}(N^{-2}), \qquad S/N=\frac{1}{2}\langle k\rangle[\log(N/\langle k\rangle)+1]+\mathcal{O}(N^{-1}) \qquad (24)$$

For $g>0$ we need to inspect the solutions of our equations with finite positive values $\langle k\rangle$ and $\langle m\rangle$, by working out the different possible scalings of the parameter u with N. In view of what we know about the scaling with N of the correct solution for u at $g=0$, the natural ansatz to consider is $u\to-\infty$ as $N\to\infty$. We now find that

$$\langle k\rangle=\Big[\mathrm{e}^{2u}N+6\langle m\rangle-6(N\mathrm{e}^{2u})^2\mathrm{e}^{2u}\Big][1+\mathcal{O}(\mathrm{e}^{2u},N^{-1})], \qquad \mathrm{e}^{-g}=\frac{N^2\mathrm{e}^{6u}}{\langle m\rangle}[1+\mathcal{O}(\mathrm{e}^{2u},N^{-1})] \qquad (25)$$

Solutions with finite $\langle k\rangle$ and $\langle m\rangle$ for $N\to\infty$ seem to require that $\mathrm{e}^{2u}N=\mathcal{O}(1)$, giving

$$u=\frac{1}{2}\log[\langle k\rangle-6\langle m\rangle]-\frac{1}{2}\log N+\mathcal{O}(N^{-1}), \qquad g=-3\log[\langle k\rangle-6\langle m\rangle]+\log(N\langle m\rangle)+\mathcal{O}(N^{-1}) \qquad (26)$$

This solution clearly exists only if $\langle m\rangle < \frac{1}{6}\langle k\rangle$. The corresponding entropy expression is found to be the following, which indeed reduces correctly to the ER entropy for $\langle m\rangle \to 0$:

$$S/N = \frac{1}{2}[\langle k\rangle - 6\langle m\rangle]\Big(1-\log[\langle k\rangle - 6\langle m\rangle]\Big) + [\frac{1}{2}\langle k\rangle - \langle m\rangle]\log N + \langle m\rangle[1-\log\langle m\rangle] + \mathcal{O}(N^{-1}) \quad (27)$$

At the point where $\langle m\rangle \uparrow \frac{1}{6}\langle k\rangle$, we see that u would have to become even more negative. There is no entropy crisis since

$$\lim_{\langle m\rangle \uparrow \langle k\rangle/6} S/N = \frac{1}{3}\langle k\rangle \log N + \frac{1}{6}\langle k\rangle\left[1-\log\left(\frac{1}{6}\langle k\rangle\right)\right] + \mathcal{O}(N^{-1}) \quad (28)$$

Hence there is no evidence for bifurcation to an alternative solution at $\langle m\rangle = \frac{1}{6}\langle k\rangle$. The failure of this simple route to lead to solutions in the regime $\langle m\rangle \geq \frac{1}{6}\langle k\rangle$ must therefore be due to the invalidity of the Poissonian assumption for $p(r|u)$.

3.3. The diagrammatic approach of Burda et al.

A drawback of the Strauss model [6], is that it has a condensed phase, where the typical networks have a tendency to form complete cliques, which does not reflect the topology of real networks. Burda et al [7] refined our understanding of this phenomenon, using a diagrammatic approach to study the series expansion of the factor $\mathrm{e}^{g\sum c_{ij}c_{jk}c_{ki}}$ in (9), with g taken to be the small parameter. They showed that the clustered phase occurred above certain critical values of the parameters, and evaluated the free energy and the expectation value of $\langle m\rangle$, in their notation, as

$$\log Z(G,\gamma) = \gamma(\mathrm{e}^G - 1) + \log Z_{\mathrm{ER}}, \qquad \langle m\rangle = N^{-1}\bar{k}^3\mathrm{e}^G \quad (29)$$

where the identification of the different parameter conventions follows from $\gamma = \bar{k}^3/6$ and $G = 6g$. We know also from [7] that if we make the substitution $G^\star \log N + \alpha = G$, in which $G^\star$ and α are functions of $\bar{k}$ but without N dependency, then the perturbation series will break down for a value of $G^\star$ that is strictly less than 1 (the actual breakdown value was found numerically to be about 0.7). Hence, we can see that within this range $N^{G^\star - 1} \to 0 \quad \text{as} \quad N \to \infty$. This means that the number of triangles per node tends to zero in the large N limit, throughout the regime where the perturbation series of [7] converges. One can also show that, in leading order in N, the average degree remains unchanged in the regime where the perturbation series converges, i.e. $\langle k\rangle = \bar{k} + \mathcal{O}(N^{-1})$.

If we write $p = \bar{k}/N$, we can write the ensemble entropy according to the expansion of [7] as

$$\begin{aligned} S - S_{\mathrm{ER}} &= \left[1 + p(1-p)\ln\left(\frac{1}{p}-1\right)\frac{\partial}{\partial p} - G\frac{\partial}{\partial G}\right]\frac{\bar{k}^3}{6}(\mathrm{e}^G - 1) \\ &= \frac{\bar{k}^3}{6}(\mathrm{e}^G-1) + \frac{1}{2}\ln\left(\frac{1}{p}-1\right)\bar{k}^3(1-p)(\mathrm{e}^G-1) - \frac{G}{6}\bar{k}^3\mathrm{e}^G \end{aligned} \quad (30)$$

Upon expanding the logarithm, and eliminating the parameter G with equation (29), it then follows that

$$S = \frac{\bar{k}\,(N-1)}{2}\left[1+\ln\left(\frac{N}{\bar{k}}\right)\right] + \frac{\langle T\rangle}{6}\left[1+\ln\left(\frac{N^3}{\langle T\rangle}\right)\right] - \frac{\bar{k}^3}{6}\left[1+\ln\left(\frac{N^3}{\bar{k}^3}\right)\right] - \frac{3\bar{k}^2}{4} + \mathcal{O}(\epsilon_N) \quad (31)$$

in which $\langle T\rangle = \langle m\rangle N$ is the average number of triangles in typical graphs from the Strauss ensemble, and $\lim_{N\to\infty}\epsilon_N = 0$. This form has similarities with previously derived results, e.g. [15, 17]. If $\langle T\rangle = \bar{k}^3$ then the entropy reduces to the Erdös-Rényi entropy, as expected. Direct comparisons with expressions obtained via the Poissonnian ansatz are not valid, because the expressions refer to different scalings with N of $\langle m\rangle$. The next

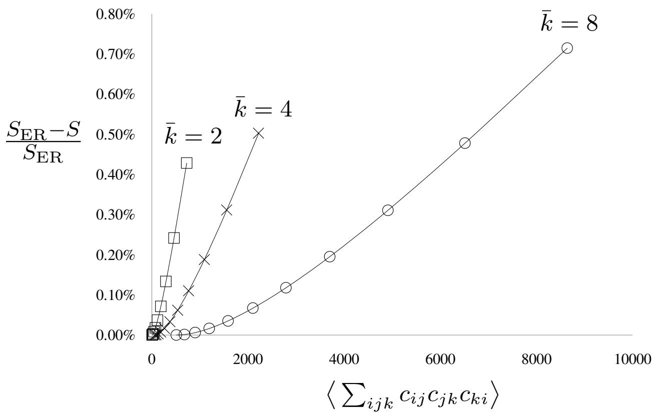

FIGURE 5. Relative complexity (i.e. relative entropy reduction) versus the expected number of triangles in the Strauss model with $N = 10,000$, calculated from the perturbation theory of [7] for values of the coupling constant G below the transition point to the clustered phase. Here S_{ER} is calculated with the actual average degree $\langle k \rangle$ of the Strauss model, rather than the implied $\bar{k}$, in order to remove the trivial effect that even a more dense uncorrelated network will automatically have more triangles.

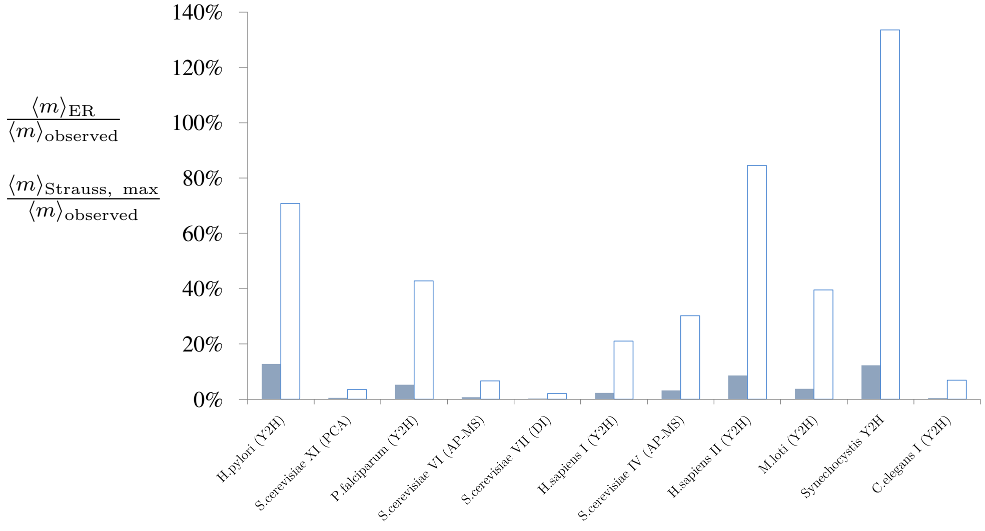

FIGURE 6. Dark bars: number of triangles in an ER ensemble divided by the number of triangles in protein-protein interaction networks (PPIN) of different species, where both have the same average degree. Light bars: maximum number of triangles in the Strauss ensemble (before the clumping transition) divided by the number of triangles in the observed protein-protein interaction networks, where both have the same average degree. The species are ordered from left to right in terms of increasing network sizes (which are on average around 2,000 nodes). References for the datasets are found in [16].

step would be to eliminate $\bar{k}$ in favour of the observable $\langle k \rangle$. This is not simple, as it effectively requires the solution of a fourth order equation.

The authors of [7] numerically deduced the critical values for the coupling parameter G for networks with average degrees of 2, 4 and 8. This gives a region within which we know that it is valid to apply formula (31), and reasonable to use a model with such parameters to model real networks. We evaluated (31) and related quantities for values of the parameter G up to roughly the critical point. The implied parameter $\bar{k}$ is found numerically. The results are shown in Figure 5. For the (realistic) values of average degree and network size considered, we observe that by the time the coupling constant G reaches the critical value, the number of triangles in the network is predicted to increase more than tenfold compared to an ER ensemble with the same average degree. However, the complexity is low when viewed as a proportion of the overall entropy of the ensemble. Constraining only the average number of loops, and remaining within a phase with relatively low clustering, apparently still leaves significant topological freedom for the networks in the ensemble.

Figure 6 compares the number of triangles observed in several biological networks, with average degree around $\langle k\rangle = 4$, with the maximum number of triangles that a Strauss ensemble graph could be expected to generate, before it goes into its degenerate clustered phase. This shows that, while the Strauss model is a substantial improvement on the Erdös-Rényi model from the point of view of the number of loops, it usually collapses into its clustered phase before it reaches biologically realistic values for its parameters. If we wish to extend the perturbation analysis beyond the critical point, we need to look at different scalings of the parameter g. However, since the un-physical behaviour of the Strauss model above this point has already been shown, such a result would be of limited application. The authors of [7] have extended their analysis in [30] to general uncorrelated degree distributions - but the agreement with simulations was less precise.

3.4. Spectrally parametrised loopy ensembles

Since the Strauss model 'clumps' above a certain critical point before realistic numbers of short loops are achieved, with the imposed triangle numbers being realised in dense cliques, one would like to define more versatile graph ensembles by including additional observables to tailor the graphs further in terms of short loops and penalise the formation of large cliques. Since the two constrained observables in the Strauss model are both seen to be specific traces of the matrix $\boldsymbol{c}$, i.e.

$$\sum_{ij} c_{ij} = \sum_{ij} c_{ij}c_{ji} = \mathrm{Tr}(\boldsymbol{c}^2), \qquad \sum_{ijk} c_{ij}c_{jk}c_{ki} = \mathrm{Tr}(\boldsymbol{c}^3) \tag{32}$$

one in effect constrains in this ensemble the second and third moment of the eigenvalue spectrum $\varrho(\mu|\boldsymbol{c})$. A natural generalisation of the Strauss ensemble would therefore be obtained by prescribing more general spectral features, or even the full eigenvalue spectrum itself. For a soft-constrained maximum entropy ensemble this would involve, rather than the scalar pair (u, g), a functional Lagrange parameter $\hat{\varrho}(u)$. Thus we obtain

$$p(\boldsymbol{c}) = Z^{-1}[\hat{\varrho}]\ \mathrm{e}^{N\int \mathrm{d}\mu\ \hat{\varrho}(\mu)\varrho(\mu|\boldsymbol{c})}, \qquad \varrho(\mu|\boldsymbol{c}) = \frac{1}{N}\sum_i \delta\left[\mu - \mu_i(\boldsymbol{c})\right] \tag{33}$$

where $\mu_i(\boldsymbol{c})$ denotes the i-th eigenvalue of $\boldsymbol{c}$. For the choice $\hat{\varrho}(\mu) = u\mu^2 + v\mu^3$ one recovers from (33) the Strauss ensemble. Including higher order terms, e.g. via $\hat{\varrho}(\mu) = \sum_{\ell=2}^{L} v_\ell \mu^\ell$ for some $L \geq 4$, would give better control over the clumping of the original Strauss model. If we define our ensemble by a full imposed spectrum $\varrho(\mu)$, which is equivalent to sending $L \to \infty$, we would have to solve the function $\hat{\varrho}(\mu)$ from

$$\forall \mu \in \mathrm{I\!R}: \qquad \sum_{\boldsymbol{c}\in G} p(\boldsymbol{c})\varrho(\mu|\boldsymbol{c}) = \varrho(\mu) \tag{34}$$

Although the early steps of the argument in [7] would still apply, it is not clear how to analytically re-sum the contributing terms in the expansion for these more general ensembles. Here we need new analytical tools.

Below we discuss a potential new route to tackle the relevant combinatorial sums. Determining the entropy of the generalised spectrally constrained graph ensemble (33) would require the evaluation of the functional

$$\phi[\hat{\varrho}] = \frac{1}{N}\log\sum_{\boldsymbol{c}\in G} \mathrm{e}^{N\int \mathrm{d}\mu\ \hat{\varrho}(\mu)\varrho(\mu|\boldsymbol{c})} \tag{35}$$

Calculating the sum over all graphs $\boldsymbol{c}\in G$ in (35) directly is not feasible. However, we can rewrite $\phi[\hat{\varrho}]$ using the standard spectrum formula of [31]:

$$\varrho(\mu|\boldsymbol{c}) = \frac{2}{N\pi}\lim_{\varepsilon\downarrow 0}\mathrm{Im}\frac{\partial}{\partial\mu}\log Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c}), \qquad Z(\mu|\boldsymbol{c}) = \int_{\mathrm{I\!R}^N}\mathrm{d}\boldsymbol{\psi}\ \mathrm{e}^{-\frac{1}{2}\mathrm{i}\boldsymbol{\psi}\cdot[\boldsymbol{c}-\mu\mathrm{1\!I}]\boldsymbol{\psi}} \tag{36}$$

This gives us, after integration by parts in the exponent,

$$\phi[\hat{\varrho}] \ = \ \lim_{\varepsilon\downarrow 0}\frac{1}{N}\log\sum_{\boldsymbol{c}\in G}\mathrm{e}^{-\frac{2}{\pi}\mathrm{Im}\int\mathrm{d}\mu\ \ \log Z(\mu+\mathrm{i}\varepsilon|\boldsymbol{c})\ \partial\hat{\varrho}(\mu)/\partial\mu} \tag{37}$$

We can now discretize the integral via $\int\mathrm{d}\mu\to\Delta\sum_\mu$, and use the identity $\mathrm{e}^{-2\mathrm{Im}\log z} = z^{\mathrm{i}}/\overline{z}^{\mathrm{i}}$ to write

$$\begin{aligned}
\phi[\hat{\varrho}] \ &= \ \lim_{\varepsilon,\Delta\downarrow 0}\frac{1}{N}\log\sum_{\boldsymbol{c}\in G}\prod_{\mu}\mathrm{e}^{-\frac{2\Delta}{\pi}\ (\partial\hat{\varrho}(\mu)/\partial\mu)\ \mathrm{Im}\log Z(\mu+\mathrm{i}\varepsilon|\boldsymbol{c})} \\
&= \ \lim_{\varepsilon,\Delta\downarrow 0}\frac{1}{N}\log\sum_{\boldsymbol{c}}\prod_{\mu}\Big[Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c})^{\mathrm{i}}\ \overline{Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c})}^{\,-\mathrm{i}}\Big]^{\frac{\Delta}{\pi}\ \partial\hat{\varrho}(\mu)/\partial\mu} \\
&= \ \lim_{\varepsilon,\Delta\downarrow 0}\ \lim_{n_\mu\to\frac{\Delta\mathrm{i}}{\pi}\partial\hat{\varrho}(\mu)/\partial\mu}\ \lim_{m_\mu\to -n_\mu}\ \Phi[\{n_\mu,m_\mu\}]
\end{aligned} \tag{38}$$

in which

$$\Phi[\{n_\mu,m_\mu\}] \ = \ \frac{1}{N}\log\sum_{\boldsymbol{c}\in G}\prod_{\mu}\Big[Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c})^{n_\mu}\ \overline{Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c})}^{\,m_\mu}\Big] \tag{39}$$

Expression (39) is reminiscent of formulae encountered in replica analyses of heterogeneous many-variable systems, which suggests a strategy for proceeding with the calculation. We can carry out the sum over all graphs $\boldsymbol{c}\in G$, which is the core obstacle in the problem, by evaluating equation (39) for positive integer values of $\{n_\mu, m_\mu\}$ (where the powers of $Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c})$ and $\overline{Z(\mu{+}\mathrm{i}\varepsilon|\boldsymbol{c})}$ simply become multiple *replicated* Gaussian integrals in which the entries $\{c_{ij}\}$ appear in factorised form). The full expression could then be determined by taking the limits in (38) via analytical continuation.

The sum over graphs has thereby been tamed, and the previous combinatorial difficulties converted into the intricacies of an unusual replica limit. In the original papers that launched and used the replica method, the (real-valued) replica dimension n had to be taken to zero, reflecting 'frozen' heterogeneity in the micro-parameters of stochastic processes [9–13]. Later statistical mechanical studies have found how real-valued but nonzero replica dimensions emerge in a natural way to describing nested processes that equilibrate at distinct temperatures and timescales, e.g. slowly evolving heterogeneity in the micro-parameters of 'fast' stochastic physical or biological processes [32–35]. To our knowledge there have not yet been calculations in which purely *imaginary* replica dimensions emerge, as in the calculation above.

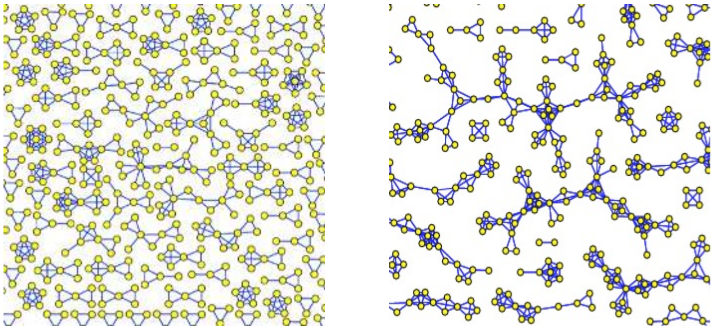

FIGURE 7. Snapshots of the finitely connected immune network of [8], describing effective interactions between T-clones, that are obtained by integrating out the B-clone variables, for different choices of the model's global control parameters. It is immediately clear from these images that the graphs of [8] have a finite number of short loops per node, and are certainly not locally tree-like.

4. Solvable immune models on loopy networks

In a recently published statistical mechanical study of the interaction between T and B-clones in the adaptive immune system [8] the authors succeeded in obtaining a full analytical solution, leading to phase diagrams and testable predictions for observables which agreed perfectly with numerical simulations. Following a preceding paper [36] they mapped the problem to a new but equivalent spin system describing only T-T interactions, by integrating out the degrees of freedom that represented the B-clones. The new effective model for interacting T-clones could then be solved using replica methods. What is intriguing in the context of this paper is that in this new model for T-clones the spins are positioned on the nodes of a finitely connected interaction graph with an extensive number of short loops, see e.g. Figure 7. Given the arguments in the previous sections, one would not have expected analytical solution to be possible.

4.1. The model of Agliari et al.

In the models of [8, 36] one studies the interactions between B-clones $b_\mu \in \mathbb{R}$ ($\mu = 1 \ldots N_B$), T-clones $\sigma_i \in \{-1, 1\}$ ($i = 1 \ldots N_T$), and external triggers of the immune system (the so-called 'antigens'). Each B-clone can recognise and attack one specific antigen species, and the T-clones are responsible for coordinating the B-clones via chemical signals (cytokines), but do so in a somewhat promiscuous manner. The collective system is described as a statistical mechanical process in equilibrium, characterised by the following Hamiltonian

$$H = \frac{1}{2\sqrt{\beta}} \sum_{\mu=1}^{N_B} b_\mu^2 - \sum_{\mu=1}^{N_B} b_\mu h_\mu, \qquad h_\mu = \sum_{i=1}^{N_T} \xi_i^\mu \sigma_i + \lambda_\mu a_\mu \tag{40}$$

Here a_μ represents the log-concentration of antigen type μ, λ_μ is the sensitivity of the μ-th B-clone to its allocated antigen, and $\xi_i^\mu \in \{-1, 0, 1\}$ represents the cytokine interaction between T-clone σ_i and B-clone b_μ. The 'field' h_μ acts as the combined input to B-clone μ. If h_μ is positive clone μ will expand, if it is negative clone μ will contract. The ξ_i^μ can be excitatory ($\xi_i^\mu = 1$), inhibitory ($\xi_i^\mu = -1$), or absent ($\xi_i^\mu = 0$), and are drawn randomly and independently from

$$p(\xi_i^\mu) = \frac{c}{2N_T} \Big[\delta_{\xi_i^\mu,1} + \delta_{\xi_i^\mu,-1}\Big] + (1 - \frac{c}{N_T})\delta_{\xi_i^\mu,0} \tag{41}$$

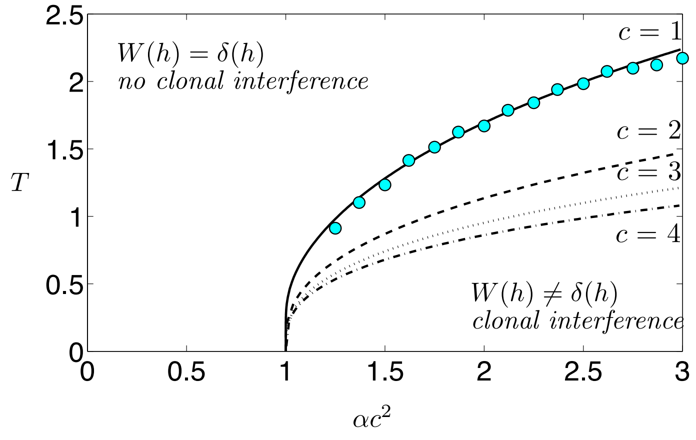

FIGURE 8. Transition lines in the $(\alpha c^2, T)$ plane for different values of c, with $T = \beta^{-1}$. The distribution $W(h)$ gives the statistics of clonal interference fields, caused by increased connectivity in the graph. Circles: values calculated via numerical solution of (45) for $c = 1$ (diagram reproduced from [8]).

Thus the original system is described by a weighted bi-partite interaction graph. The parameter $c \geq 0$ controls the degree of promiscuity of the B-T interactions. Realistic clone numbers would be $N_B \sim 10^8$ and $N_T \sim 10^8$, so statistical mechanical approaches are valid. The authors of [8] study this system in the regime where $N_T = N$ and $N_B = \alpha N$, with finite $\alpha > 0$ and $N \to \infty$. They 'integrate out' the B-clones in the system's partition function (which requires only a simple Gaussian integral), and are left with a system of interacting Ising spins, described by the effective Hamiltonian

$$H_{\text{eff}}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i \sigma_j J_{ij} - \sum_{i=1}^{N} \sigma_i \sum_{\mu=1}^{\alpha N} \psi_\mu \xi_i^\mu, \qquad J_{ij} = \sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu, \qquad \psi_\mu = \lambda_\mu a_\mu \tag{42}$$

This Hamiltonian is reminiscent of the one found in attractor neural network models [37, 38], with a so-called Hebbian interaction matrix J_{ij} coupling the spins. However, due to the scaling with N of the probabilities in (41), the present (weighted) interaction matrix is finitely connected. Moreover, unlike finitely connected neural network models [13], where one stores and recalls a finite number of binary patterns with *extensive* information content each, here one seeks to store and recall an *extensive* number of patterns with a *finite* number of bits each. This distinction is not only vital in the immunological context, since an organism has to defend itself against extensive simultaneous invasions to survive, but it also generates fundamental mathematical differences. Finitely connected attractor models like [13] operate on graphs that are locally tree-like, by construction, whereas the model (42) typically involves 'loopy' interaction graphs; see Figure 7.

4.2. Statistical mechanical analysis

The details of the statistical mechanical analysis of (42) can be found in [8], here we only discuss their results. In view of the extensive number of 'stored patterns' in this model, compared to finitely connected attractor neural networks, the conventional analysis route (involving sub-lattice magnetisations as order parameters) can no longer be used. Instead the appropriate order parameter is

$$\mathscr{P}(M, \psi) = \frac{1}{\alpha N}\sum_{\mu=1}^{\alpha N} \delta_{M, M_\mu(\boldsymbol{\sigma})} \delta(\psi - \psi_\mu) \qquad M_\mu(\boldsymbol{\sigma}) = \sum_{i=1}^{N} \xi_i^\mu \sigma_i \tag{43}$$

Each of the (extensively many) state overlaps $M_\mu(\boldsymbol{\sigma})=\sum_{i=1}^N \xi_i^\mu\sigma_i$ represent the combined activation/repression signal coming from the T-cells and acting upon B-clone μ, so the conditional distribution $\mathscr{P}(M|\psi)$ quantifies the strength and specificity of the response of the adaptive immune system to a typical antigen attack. Given this order parameter, the calculation of the disorder-averaged free energy involves path integral techniques combined with replica methods. The end result, within the replica-symmetric (RS) ansatz, is the following self-consistent equation for a field distribution $W(h)$:

$$
\begin{aligned}
W(h) &= \mathrm{e}^{-c}\sum_{k\geq 0}\frac{c^k}{k!}\mathrm{e}^{-\alpha ck}\sum_{r\geq 0}\frac{(\alpha c)^r}{r!}\int_{-\infty}^{\infty}\Big[\prod_{s\leq r}\mathrm{d}h_s W(h_s)\Big]\sum_{\ell_1\ldots\ell_r\leq k}\int\!\mathrm{d}\psi\;P(\psi) && (44)\\
&\quad\times\sum_{\tau=\pm 1}\delta\left[h-\tau\psi-\frac{1}{2\beta}\log\left(\frac{\sum_{\sigma_1\ldots\sigma_k=\pm 1}\mathrm{e}^{\beta(\sum_{\ell\leq k}\sigma_\ell)^2/2c+\beta(\sum_{\ell\leq k}\sigma_\ell)(\psi+\tau/c)+\beta\sum_{s\leq r}h_s\sigma_{\ell_s}}}{\sum_{\sigma_1\ldots\sigma_k=\pm 1}\mathrm{e}^{\beta(\sum_{\ell\leq k}\sigma_\ell)^2/2c+\beta(\sum_{\ell\leq k}\sigma_\ell)(\psi-\tau/c)+\beta\sum_{s\leq r}h_s\sigma_{\ell_s}}}\right)\right] && (45)
\end{aligned}
$$

in which β is the inverse temperature (i.e. inverse noise level) of the system. The distribution $W(h)$ turns out to describe the distribution of clonal interference fields, i.e. the unwanted signalling cross-talk between clones. Equation (45) always has the trivial solution $W(h)=\delta(h)$, which represents interference-free operation, but exhibits bifurcations away from this state at parameter combinations (α,c,β) such that

$$
1=\alpha c^2\sum_{k\geq 0}\mathrm{e}^{-c}\frac{c^k}{k!}\left\{\frac{\int\!\mathrm{d}z\;\mathrm{e}^{-\frac{1}{2}z^2}\tanh(z\sqrt{\beta/c}+\beta/c)\cosh^{k+1}(z\sqrt{\beta/c}+\beta/c)}{\int\!\mathrm{d}z\;\mathrm{e}^{-\frac{1}{2}z^2}\cosh^{k+1}(z\sqrt{\beta/c}+\beta/c)}\right\}^2 \tag{46}
$$

This equation therefore defines a phase transition that separates a regime of interference-free operation of the adaptive immune system from one that exhibits deteriorating performance due to interference between clones. Cross-sections of this transition surface are shown in Figure 8.

4.3. Stepping back - why is the 'loopy' Agliari et al model solvable?

In retrospect it is clear why the interacting spin model (42) is solvable, in spite of the extensively many short loops in its interaction graph $\boldsymbol{J}=\{J_{ij}\}$. The reason is that the $N\times N$ interaction matrix has the form $\boldsymbol{J}=\boldsymbol{\xi}^\dagger\boldsymbol{\xi}$, with a $p\times N$ matrix $\boldsymbol{\xi}$ with entries $\{\xi_i^\mu\}$ that itself represents a locally tree-like graph. Mathematically this allows one to introduce a Hubbard-Stratonovich-type [5] transformation in the partition function, to a new but equivalent model that has a locally tree-like structure, at the cost of introducing p new Gaussian degrees of freedom:

$$
J_{ij}=\sum_{\mu=1}^p\xi_i^\mu\xi_j^\mu:\qquad \sum_{\boldsymbol{\sigma}\in\{-1,1\}^N}\mathrm{e}^{\beta\sum_{i<j}J_{ij}\sigma_i\sigma_j}=\int_{\mathrm{I\!R}^p}\frac{\mathrm{d}\boldsymbol{z}}{(2\pi)^{p/2}}\sum_{\boldsymbol{\sigma}\in\{-1,1\}^N}\mathrm{e}^{\sqrt{\beta}\sum_{\mu i}z_\mu\xi_i^\mu\sigma_i-\frac{1}{2}\sum_\mu z_\mu^2} \tag{47}
$$

This construction indeed reflects exactly the origin of the model in [8], with the B-cells acting as auxiliary Gaussian variables. The more general question would now be for which other interaction networks with extensively many short loops one could write (or approximate sensibly) the interaction matrix in a form $\boldsymbol{J}=\boldsymbol{\xi}^\dagger\boldsymbol{\xi}$ for some p and some $p\times N$ bi-partite but locally tree-like graph $\boldsymbol{\xi}$. Models on such networks could then be mapped similarly onto an equivalent model where the short loops have been traded in for extra degree of freedom. This would obviously only be possible for graphs with non-negative spectra.

Another curious aspect of the Agliari et al model is that it defies another expectation. Normally the entropy calculation for a random graph ensemble is easier than solving an interacting spin model on the ensemble's graphs. Here the situation is reversed. The entropy calculation is harder. For the present nondirected weighted

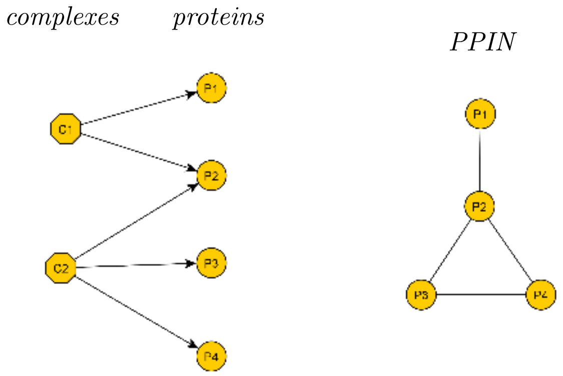

FIGURE 9. In the bipartite graph on the left proteins and complexes are both represented as nodes, and a link indicates that a given protein is a constituent of the complex. In the graph on the right (the standard representation of PPIN databases) there is a link between two proteins if they are jointly part of one or more protein complexes. The latter graph is weighted if the value of the link c_{ij} is defined as the *number* of complexes in which proteins P_i and P_j participate simultaneously.

graph ensemble we have

$$\boldsymbol{c}\in \mathbb{N}^{\frac{1}{2}N(N-1)} = G', \qquad p(\boldsymbol{c}) = \Big\langle \prod_{i<j} \delta_{c_{ij},\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu}\Big\rangle_{\boldsymbol{\xi}} \tag{48}$$

in which $\langle\ldots\rangle_{\boldsymbol{\xi}}$ denotes averaging over (41). This ensemble is not a maximum entropy ensemble with constraints in the sense of (1) or (2), and the methods used for the latter ensembles are no longer applicable. Using the replica identity $\langle \log Z\rangle = \lim_{n\to 0} n^{-1}\log\langle Z^n\rangle$ we can, however, write the Shannon entropy per node of the ensemble (48) as

$$\begin{aligned} S &= \sum_{\boldsymbol{c}\in G'} p(\boldsymbol{c})\log p(\boldsymbol{c}) = \lim_{n\to 0}\frac{1}{n}\log\sum_{\boldsymbol{c}\in G'} p^{n+1}(\boldsymbol{c}) \\ &= \lim_{n\to 0}\frac{1}{n}\log\sum_{\boldsymbol{c}\in G'}\Big\langle\cdots\Big\langle\prod_{i<j}\Big[\prod_{\alpha=1}^{n+1}\delta_{c_{ij},\sum_{\mu=1}^{p}\xi_{i,\alpha}^{\mu}\xi_{j,\alpha}^{\mu}}\Big]\Big\rangle_{\boldsymbol{\xi}_1}\cdots\Big\rangle_{\boldsymbol{\xi}_{n+1}} \\ &= \lim_{n\to 0}\frac{1}{n}\log\Big\langle\cdots\Big\langle\prod_{i<j}\Big\{\sum_{\ell\geq 0}\Big[\prod_{\alpha=1}^{n+1}\delta_{\ell,\sum_{\mu=1}^{p}\xi_{i,\alpha}^{\mu}\xi_{j,\alpha}^{\mu}}\Big]\Big\}\Big\rangle_{\boldsymbol{\xi}_1}\cdots\Big\rangle_{\boldsymbol{\xi}_{n+1}} \end{aligned} \tag{49}$$

Working out the leading orders in N of this remaining combinatorial problem looks feasible but has not yet been fully solved, and is the subject of ongoing work.

4.4. The connection with models of proteins and their complexes

Many systems can be modelled with an interaction graph similar to the one of [8, 36]. The author of [39] showed how a similar model could be used to describe a social network, where the probability of two nodes being connected is driven by whether or not they are both members of the same social group. A less explored application is that although most protein-protein interaction network (PPIN) data repositories (such as [25]) report only binary interactions between pairs of protein species, the more natural description is in fact that of a bipartite graph, with one set of αN nodes representing protein complexes and another set representing

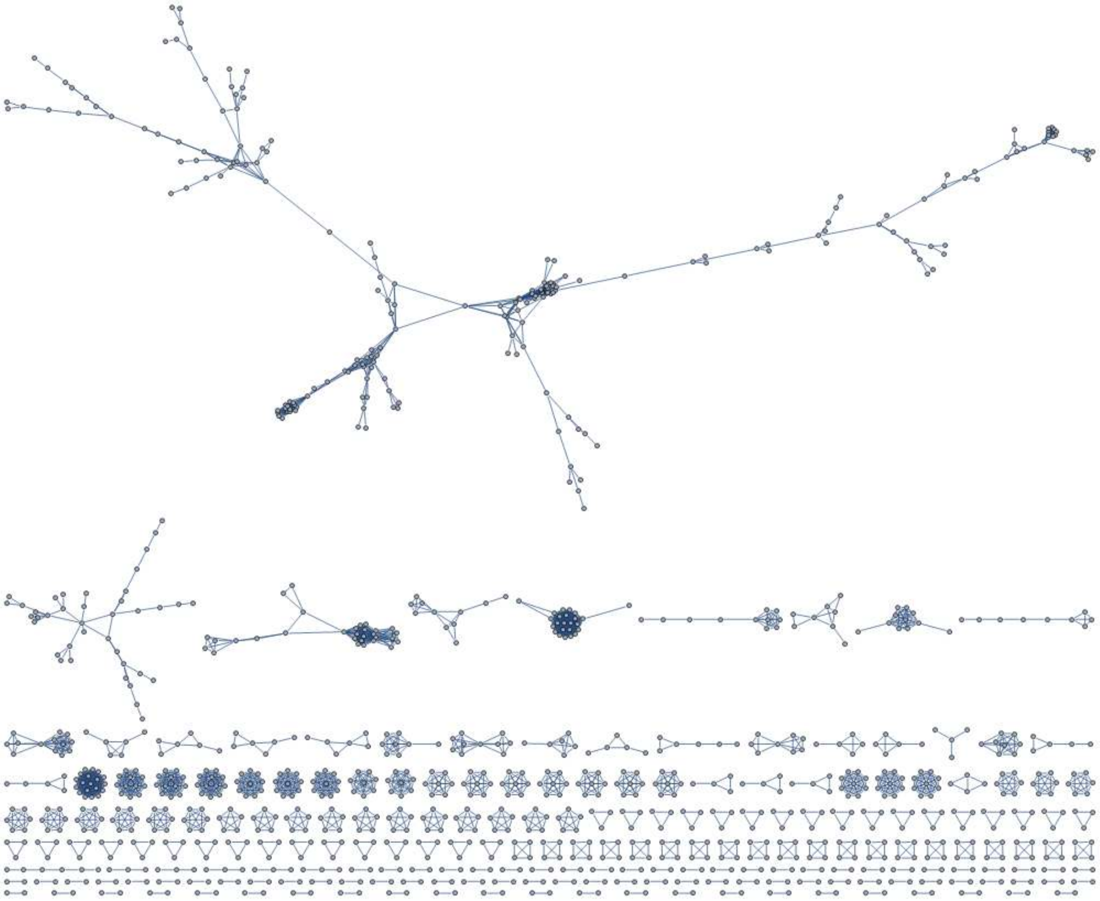

FIGURE 10. Projection onto the protein space, in the sense of Figure 9, of the protein complexes and their core constituents identified in *Sac. cerevisiae* [40] via mass-spectroscopy. The resulting network shows distinct cliques, but only the beginning of the dense core that is generally seen in protein-protein interaction networks. One would expect the dense core to emerge if the data included also non-core proteins, as well as protein reactions which may not necessarily correspond to named complexes.

the N individual proteins. A link from a given protein to a given complex then indicates that the protein participates in the complex; see Figure 9. If from such data we construct a new weighted graph, with nodes that represent the proteins only, and links between the nodes that give the number of complexes in which they jointly participate, we obtain once more graphs $\boldsymbol{c} \in G'$ from the graph ensemble (48). The only difference with the immune models is that now the variables ξ_i^μ take their values randomly and independently from the set $\{0,1\}$:

$$p(\boldsymbol{c}) = \Big\langle \prod_{i<j} \delta_{c_{ij},\sum_{\mu=1}^p \xi_i^\mu \xi_j^\mu} \Big\rangle_{\boldsymbol{\xi}}, \qquad p(\xi_i^\mu) = \frac{q}{N}\delta_{\xi_i^\mu,1} + \Big(1-\frac{q}{N}\Big)\delta_{\xi_i^\mu,0} \tag{50}$$

This graph ensemble was studied in [39], in terms of its topological properties (percolation transition, path lengths, and so on), but not in the context of protein interactions. Several observables of the projected network

$\boldsymbol{c} \in G'$ can be immediately deduced from the parameters of the bipartite graph. For instance:

$$\langle c_{ij} \rangle = \alpha N \langle \xi_i \xi_j \rangle = \alpha q^2/N \tag{51}$$

$$p(c_{ij}) = \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega c_{ij}} \left\langle \mathrm{e}^{\mathrm{i}\omega \xi_i \xi_j} \right\rangle_{\{\xi\}}^{\alpha N} = \binom{\alpha N}{c_{ij}} \left[\frac{q^2}{N^2}\right]^{c_{ij}} \left[1 - \frac{q^2}{N^2}\right]^{\alpha N - c_{ij}} \tag{52}$$

from which it also follows that $\bar{k} = N^{-1} \sum_{ij} [1 - p(c_{ij} = 0)] = \alpha q^2$. As with the immune model, one obtains loops in the PPIN graph $\boldsymbol{c}$ again from an underlying model which is tree-like; here this underlying model is the description of the system in terms of the proteins and their complexes. In Figure 10 we show the result of using the data from [40] on protein complexes found in yeast in order to construct a protein-protein interaction network according to the definition in (50), i.e. via $c_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$, with ξ_i^μ indicating whether ($\xi_i^\mu = 1$) or not ($\xi_i^\mu = 0$) protein i participates in complex μ. It will be clear that, using the technology of [8], one should now be able to solve analytically models of protein reaction processes on 'loopy' protein-protein interaction networks, provided they are built on these along the lines of (50).

5. Conclusion

In this paper we have discussed some conceptual and mathematical issues that emerge as soon as one studies (processes on) graphs and networks that are not locally tree-like, but exhibit extensive numbers of short loops. Stochastic processes on graphs with many short loops have quantitative characteristics that are very different from those that run on tree-like structures. This is why Ising models on tree-like lattices are trivially solved, but Ising models on finite-dimensional lattices are not. Yet our currently available arsenal of mathematical tools for analysing 'loopy' graphs is rather limited, because the tree-like assumption is at the very core of most of our techniques, whether we are analysing processes on graphs or calculating entropies of graph ensembles. Even asymptotic results for infinitely large 'loopy' graphs are largely absent. However, we believe that this area will become increasingly important in the coming years, since the fact is that many (if not most) of the networks that we can observe in the real-world do have significant numbers of loops.

We have tried to propose and explain some new ideas and possible routes forward, aimed at making progress in the analytical study of 'loopy' random graph ensembles. Some built on our own ideas (including e.g. simple approximations and replica techniques) and some built on work of others (such as the papers by Burda et al, that are based on diagrammatic expansions). Most of the material relates to ongoing work, and is still only beginning to be explored. We discussed in more detail two specific random graph ensembles: the Strauss model, because it is the simplest possible random graph ensemble in which one can induce arbitrary numbers of short loops (in this case triangles), and an ensemble that emerged in a recent immunological model, because this model could be solved analytically in spite of its 'loopy' nature. In this paper we have focussed on static random graph models. Dynamical or growing models are a distinct area of study - readers may be interested to refer to [45–47] and related papers.

Acknowledgements

Both authors gratefully acknowledge financial support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom, and would like to Franca Fraternali, Alessia Annibale and Eric Blanc for discussions and Sun Sook Chung and Alessandro Pandini for the triangle count data used as the basis for Figure 6.

References

[1] RJ Prill, PA Iglesias and A Levchenko 2005 *PLoS biology* **11**, e343.

[2] J Jeong and P Berman 2008 *BMC systems biology* **2**, 12.

[3] HJCM El-Samad, H Kurata, JC Doyle, CA Gross and M Khammash 2005 *Proc. Natl. Acad. Sci. USA* 102, 2736–2741.

[4] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii and U Alon 2002 *Science* **298**, 824–827.

[5] RJ Baxter 1982 *Exactly Solved Models in Statistical Mechanics* (Academic Press)

[6] D Strauss 1986 *SIAM Review* **28**, 513-527.

[7] Z Burda, J Jurkiewicz and A Krzywicki 2004 *Phys. Rev. E* **69**, 026106.

[8] E Agliari, A Annibale, A Barra, ACC Coolen and D Tantari 2013 *J. Phys. A: Math. Theor.* **46**, 415003.

[9] D Sherrington and S Kirkpatrick 1975 *Phys. Rev. Lett.* **35**, 1792-1796.

[10] M Mézard, G Parisi and MA Virasoro 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[11] L Viana L AJ and Bray 1985 *J. Phys. C* **18**, 3037.

[12] R Monasson 1998 *J. Phys. A: Math. Gen.* **31**, 513 -529.

[13] B Wemmenhove and ACC Coolen 2003 *J. Phys. A: Math. Gen.* **36**, 9617-9633.

[14] P Erdös and A Rényi A 1959 *Publ. Math.* **6**, 290–297.

[15] A Annibale, ACC Coolen, LP Fernandes, F Fraternali and J Kleinjung 2009 *J. Phys. A: Math. Theor.* **42**, 485001.

[16] LP Fernandes, A Annibale, J Kleinjung, ACC Coolen and F Fraternali 2010 *PLoS ONE* **5**, e12083.

[17] ES Roberts, T Schlitt and ACC Coolen 2011 *J. Phys. A: Math. Theor.* **44**, 275002.

[18] ACC Coolen, F Fraternali, A Annibale, LP Fernandes and J Kleinjung 2011 in *Handbook of Statistical Systems Biology* (Wiley; M Stumpf, DJ Balding and M Girolami, Eds), 309-330.

[19] TM Cover and JA Thomas 1991 *Elements of Information Theory* (New York: Wiley)

[20] JPL Hatchett, B Wemmenhove, I Pérez-Castillo, T Nikoletopoulos, NS Skantzos and ACC Coolen 2004 *J. Phys. A: Math. Theor.* **37**, 6201-6220.

[21] A Mozeika and ACC Coolen 2009 *J. Phys. A: Math. Theor.* **42**, 195006.

[22] K Mimura and ACC Coolen 2009 *J. Phys. A: Math. Theor.* **42**, 415001.

[23] T Nikoletopoulos, ACC Coolen, I Pérez-Castillo, NS Skantzos, JPL Hatchett and B Wemmenhove 2004 *J. Phys. A: Math. Theor.* **37**, 6455.

[24] T Preis, P Virnau, W Paul and JJ Schneider 2009 *J. Comp. Phys.* **228**, 4468.

[25] TSK Prasad, R Goel, K Kandasamy, S Keerthikumar, S Kumar *et al.* 2009 *Nucleic Acids Res.* **37**, D767.

[26] ACC Coolen, A De Martino and A Annibale 2009 *J. Stat. Phys.* **136**, 1035-1067.

[27] A Montanari and T Rizzo T 2005 *J. Stat. Mech.*, P10011.

[28] M Chertkov and VY Chernyak 2006 *J. Stat. Mech.*, P06009.

[29] V Gomez V, JM Mooij and HJ Kappen HJ 2007 *J. Machine Learning Res.* **8**, 1987-2016.

[30] Z Burda, J Jurkiewicz and A Krzywicki, A 2004 *Phys. Rev. E* **70**, 026106.

[31] SF Edwards and RC Jones 1976 *J. Phys. A: Math. Gen.* **9**, 1595-1603.

[32] ACC Coolen, RW Penney and D Sherrington 1993 *Phys. Rev. B* **48**, 16116-16118.

[33] V Dotsenko, S Franz and M Mézard 1994 *J. Phys. A: Math. Gen.* **27**, 2351-2365.

[34] T Uezu and ACC Coolen 2002 *J. Phys. A: Math. Gen.* **35**, 2761-2809.

[35] S Rabello, ACC Coolen, CJ Pérez-Vicente and F Fraternali 2008 *J. Phys. A: Math. Theor.* **41**, 285004.

[36] E Agliari, A Annibale, A Barra, ACC Coolen and D Tantari 2013 *J. Phys. A: Math. Theor.* **46**, 335101.

[37] DJ Amit 1989 *Modeling Brain Function* (Cambridge: University Press)

[38] ACC Coolen, R Kühn and P Sollich 2005 *Theory of Neural Information Processing Systems* (Oxford: University Press)

[39] MEJ Newman 2003 *Phys. Rev. E* **68**, 026121.

[40] A-C Gavin, P Aloy, P Grandi, R Krause, M Boesche, M Marzioch, C Rau, LJ Jensen, S Bastuck et al. 2006 *Nature* **440**, 631-636.

[41] S Horvát, É Czabarka, and Z Toroczkai 2014 *arXiv:1407.0991*

[42] K Anand and G Bianconi 2009 *Phys. Rev. E* **80**, 045102.

[43] DV Foster, JG Foster, P Grassberger and M Paczuski 2011 *Phys. Rev. E* **84** 066117.

[44] P Colomer-de-Simón and M Boguna 2014 *arXiv:1401.8176.*

[45] P Holme and JK Beom 2002 *Phys. Rev. E* **65** 026107.

[46] M Marsili F Vega-Redondo and F Slanina 2052 *PNAS* **101** 1439-1442.

[47] A Vázquez 2003 *Phys. Rev. E* **67** 056104.