The two-star model: exact solution in the sparse regime and condensation transition

A Annibale and O T Courtney

Department of Mathematics, King’s College London, The Strand, London WC2R 2LS, UK

E-mail: alessia.annibale@kcl.ac.uk

Received 29 April 2015, revised 15 July 2015
Accepted for publication 17 July 2015
Published 13 August 2015

Abstract

The two-star model is the simplest exponential random graph model that displays complex behavior, such as degeneracy and phase transition. Despite its importance, this model has been solved only in the regime of dense connectivity. In this work we solve the model in the finite connectivity regime, far more prevalent in real world networks. We show that the model undergoes a condensation transition from a liquid to a condensate phase along the critical line corresponding, in the ensemble parameters space, to the Erdős–Rényi (ER) graphs. In the fluid phase the model can produce graphs with a narrow degree statistics, ranging from regular to ER graphs, while in the condensed phase, the ‘excess’ degree heterogeneity condenses on a single site with degree $O(N)$. This shows the unsuitability of the two-star model, in its standard definition, to produce arbitrary finitely connected graphs with degree heterogeneity higher than ER graphs and suggests that non-pathological variants of this model may be attained by softening the global constraint on the two-stars, while keeping the number of links hardly constrained.

Keywords: complex networks, condensation, large deviation

(Some figures may appear in colour only in the online journal)

1. Introduction

Exponential random graph models (ERGM) are ensembles of random graphs where each graph configuration $c$ appears with a probability $p(c) \propto e^{-H(c)}$ given by the Gibbs–Boltzmann distribution, where $H(c)$ is the graph Hamiltonian, enclosing several properties of the networks in the ensemble. First introduced in the 1980s by Holland and Leinhardt [22], and further developed by Frank and Strauss [19] and in several later studies [3, 6, 7, 24, 27], ERGM soon became popular in social network analysis. Computational tools to analyze and
simulate networks based on ERGM are largely available on the web, as the ERGM and SIENA packages, and several paradigmatic models of random graphs can be written in the exponential form, for suitable choices of the graph Hamiltonian, including the Erdős–Rényi (ER) graph ensemble \[20\] and the ensemble of graphs with soft-constrained degree sequence \([4, 8, 16]\). However, ERGM have known drawbacks which limit their practical use as proxies or null models for real networks. In particular, they may display degeneracy behavior and may fail to produce graphs with properties within certain ranges, which are nevertheless observed in nature.

A useful model to understand degeneracy behavior and limitations of ERGM, is the two-star model, which is simple enough to be amenable of exact solutions, while exhibiting the interesting behavior of more complex ERGM. A well-known solution for the two-star model has been developed, in the limit of large system size \(N\), by Park and Newman \[12\], within a mean-field approach and expansions around it. However, this approach requires that the connectivity of the graph is \(O(N)\), a condition which is hardly met by real world networks, like social and biological networks, where the connectivity is \(O(N^0)\). In this work we solve the two-star model exactly, in the finite connectivity regime, for large system size, by using path integrals. We check our calculations against Monte-Carlo simulations and compare our results with those predicted by mean-field theories. Results show that the two-star model undergoes a condensation transition from a liquid to a condensate state, along the critical line corresponding, in the ensemble parameters space, to the ER graphs. In the liquid phase the model can only produce graphs within a narrow range of degree statistics, namely between regular and ER, whereas in the condensed phase a condensate of size \(O(\sqrt{N})\) emerges, residing on a single site.

The paper is structured as follows: in section 2 we define the model and review the mean-field solution, in section 3 we solve the two-star model in the finite connectivity regime and compare our results with the mean-field predictions and with Monte-Carlo simulations. In section 4 we show that the phase transition displayed by the system in the finite connectivity regime is a condensation, related to large fluctuations in the sums of random variables. Finally, we summarize our conclusions in section 5.

2. The model and the mean-field solution

The two-star model is a prototypical example of ERGM. In this section, we give a brief review of ERGM, the two-star model and its mean-field solution.

2.1. Background: the ERGM

For simple and undirected graphs of \(N\) nodes, ERGM are graph ensembles where each graph is defined by an adjacency matrix \(c\), with elements \(c_{ij} \in \{0, 1\}\), \(\forall i, j = 1, \ldots, N\), \(c_{ii} = 0\), \(\forall i, c_{ij} = c_{ji} \forall i < j\). Each graph \(c\) appears in the ensemble with a probability given by the Gibbs–Botzmann distribution

\[
p(c) = \frac{1}{Z} e^{-H(c)}
\]

with Hamiltonian \(H(c)\) and partition function \(Z = \sum_{c} e^{-H(c)}\). The Hamiltonian is given by

\[
H(c) = -\sum_{\mu=1}^{K} \lambda_{\mu} \Omega_{\mu}(c),
\]
where $\mathbf{W} = (\Omega_1, \ldots, \Omega_K)$ is a set of observables of the graph $c$ for which one has statistical estimates $\mathbf{W} = (\Omega_1, \ldots, \Omega_K)$, so that

$$\langle \Omega_\mu(c) \rangle = \Omega_\mu \quad \forall \mu = 1, \ldots, K \tag{3}$$

with $\langle \Omega_\mu(c) \rangle = \sum_c p(c) \Omega_\mu(c)$ and $\lambda = (\lambda_1, \ldots, \lambda_K)$ are ensembles parameters also called ‘conjugated’ observables, which have to be calculated from the equations for the constraints (3). The distribution (1) with Hamiltonian (2) maximizes the Shannon entropy

$$S(\Omega) = -\sum_c p(c) \ln p(c) \tag{4}$$

subject to the constraints (3) and the normalization $\sum_c p(c) = 1$. Hence, ERGM are maximum entropy ensembles conditioned on the imposed constraints.

For all but the simplest ERGM, exact solutions for the ensemble parameters are not available, so one has normally to take recourse to numerical methods. The latter have been the subject of intense investigation over the last few decades and range from pseudo-likelihood estimation [5, 7, 15, 21, 23, 24] to Markov Chain Monte Carlo maximum likelihood techniques [17, 23, 26, 28], to Bayesian inference [1, 2, 10].

Once the values of the parameters $\lambda$ are available, one can use the resulting probability distribution $p(c)$ to estimate the value of observables for which no estimate is available. We note that expectation values for the primary network observables $\Omega_\mu$ are given by

$$\langle \Omega_\mu(c) \rangle = \frac{1}{Z(\lambda)} \sum_c \Omega(c) e^{\sum_\mu \lambda_\mu \Omega_\mu(c)} = -\frac{\partial F(\lambda)}{\partial \lambda_\mu}, \tag{5}$$

where $F(\lambda) = -\ln Z(\lambda)$ is the free-energy, and their fluctuations are given by the second derivative of the free-energy $\langle \Omega_\mu^2 \rangle = \langle \Omega_\mu \rangle^2 = \frac{\partial^2 F}{\partial \lambda_\mu^2}$ which, in equilibrium, equates the so-called susceptibility, defined as $\chi_\mu = \partial \langle \Omega_\mu \rangle / \partial \lambda_\mu$, and measuring the deviation in $\langle \Omega_\mu \rangle$ when a change in the external variable $\lambda_\mu$ is applied. These relations suggest that in general we are able to calculate the ensemble parameters from the equations for the constraints, if we can calculate the free-energy of the ERGM. This is straightforward only for Hamiltonians which are linear in the adjacency matrix, where the graph distribution factorizes over the links $p(c) = \prod_{i<j} p(c_{ij})$. For nonlinear Hamiltonians, like the two-star model and Strauss model [7], solutions have been found within mean-field approximations and expansions about it [12, 14].

### 2.2. The two-star model

In the two-star model, one chooses, as graph observables, the number of links

$$L(c) = \frac{1}{2} \sum_{i<j} c_{ij} \tag{6}$$

and the number of two-stars (i.e. paths of length two along the network links)

$$S(c) = \frac{1}{2} \sum_{i<j<k \neq (i,j)} c_{ij} c_{jk} \tag{7}$$

and assumes that the expectations $L = \langle L(c) \rangle$ and $S = \langle S(c) \rangle$ are known. This leads to the nonlinear network Hamiltonian
\[ H(\mathbf{c}) = -\theta_1 L(\mathbf{c}) - \theta_2 S(\mathbf{c}) = -\sum_{i<j} c_{ij} \left( \theta_1 + \theta_2 \sum_{k=i}^j c_{jk} \right) \]
\[ = -\sum_{i<j} c_{ij} \left( \theta_1 - \theta_2 \right) + \theta_2 \sum_{k=j}^i c_{jk} \]
\[ = -\sum_{i<j} c_{ij} \left( 2\alpha + 2\beta \sum_{k} c_{jk} \right). \] (8)

where we set \( \alpha = (\theta_1 - \theta_2)/2 \) and \( \beta = \theta_2/2 \). Alternatively, we can define the local degrees of graph \( \mathbf{c} \) as \( k_i(\mathbf{c}) = \sum_j c_{ij} \) \( \forall \ i \), and write the observables (6) and (7) as

\[ L(\mathbf{c}) = \frac{1}{2} \sum_i k_i(\mathbf{c}) \]
\[ S(\mathbf{c}) = \frac{1}{2} \left( \sum_j k_j^2(\mathbf{c}) - \sum_j k_j(\mathbf{c}) \right) \] (9)

so that the Hamiltonian reads

\[ H(\mathbf{c}) = -\alpha \sum_i k_i(\mathbf{c}) - \beta \sum_j k_j^2(\mathbf{c}). \] (10)

At this point it is useful to define the average connectivity \( \langle k \rangle = N^{-1} \sum_i k_i(\mathbf{c}) \) and the expected average connectivity in the ensemble \( \langle \langle k \rangle \rangle = \langle \langle k \rangle \rangle \), with \( \langle \cdots \rangle \) denoting the average over the ensemble probability \( p(\mathbf{c}) \). In the two-star model, we have \( L = N \langle k \rangle / 2 \) and \( S = N (\langle k^2 \rangle - \langle k \rangle ) / 2 \), hence we constrain \( \langle k \rangle \) and \( \langle k^2 \rangle \).

The two-star model has so far been solved for dense graphs, with average connectivity \( \langle k \rangle = O(N) \), for which mean-field approaches are exact in the thermodynamic limit. The mean-field solution (see section 2.3) shows that for large \( N \), and \( \beta N = 2J \), the two-star model undergoes, for any \( J \geq 1 \), a first-order phase transition, between a phase of low connectivity and one of high connectivity, separated by the critical line \( J = -\alpha \) in the space of ensemble parameters. The critical line terminates at the critical point \( J = 1 \), where a second order phase transition takes place from the symmetry-broken state with two phases, the one with high and the one with low connectivity respectively, to a symmetric state [9, 12]. In particular, for \( J \geq 1 \) the link density is discontinuous meaning that the model fails to produce intermediate connectivities by suitably tuning the ensemble parameters. Along the critical line \( J = -\alpha \) (and for \( J \geq 1 \)) one has degeneracy, meaning that for the same ensemble parameters the model produces either a sparse or a dense graph.

However, it is not clear a priori how well the mean-field scenario applies to the finite connectivity regime, where Gaussian fluctuations (around the mean-field solution) become dominant rather than a small perturbation around the leading order statistics. In addition, we are interested in establishing whether in the phase at low connectivity the model can produce arbitrary densities of stars for any given (finite) connectivity, by choosing suitably the ensemble parameters. More in general, we aim to establish whether the two-star model may serve as a plausible null-model for finitely connected networks, which are far more prevalent than dense networks in the real world. We answer these questions precisely in section 3 by solving the two-star model exactly, in the limit \( N \to \infty \), for finite connectivity \( \langle k \rangle = O(N^0) \).
2.3. The mean-field approach

In this section we briefly illustrate the mean-field approach, that is exact, in the thermodynamic limit, for dense graphs \([13]\), but it is expected to give inaccurate results for finite connectivity. By analogy with spin models, we can regard the expression in the brackets of (8) as the local field acting on the link \(c_{ij}\). The mean-field approximation replaces the local fields with their ensemble averages, i.e. \(c_{jk} \to \langle c_{jk} \rangle\). Doing so, all edges in the model become equivalent and the average probability to observe a link is the same for all links \(p = \langle c_{jk} \rangle \forall j, k\). Hence, the Hamiltonian becomes

\[
H(e) = -\lambda \sum_{i<j} c_{ij},
\]

where

\[
\lambda = 2\alpha + 2\beta (N-1)p \approx 2\alpha + 2\beta N p
\]

is now a function of the unknown probability \(p\) and the last approximation holds for \(N \gg 1\). The Hamiltonian (11) leads to the partition function

\[
Z = \sum_{e} \prod_{i<j} e^{x_{ij}} \prod_{i<j} \left( 1 + e^{x_{ij}} \right) = \left( 1 + e^{\lambda} \right)^{N(N-1)/2}
\]

which immediately gives the free-energy

\[
F = -\ln Z = -\frac{N(N-1)}{2} \ln\left( 1 + e^{\lambda} \right)
\]

that can be used to write the equation for the constraint

\[
L = -\frac{\partial F}{\partial \lambda} = \frac{N(N-1)}{2} \frac{e^{\lambda}}{1 + e^{\lambda}}.
\]

If links are all drawn randomly and independently with probability \(p\), we have

\[
L = \frac{N(N-1)}{2} p
\]

hence we get from (15)

\[
p = \frac{e^{\lambda}}{1 + e^{\lambda}} = \frac{1}{2} \left( 1 + \tanh \frac{\lambda}{2} \right).
\]

where in the last equality we used the identity \(e^x / \cosh x = (1 + \tanh x) / 2\). Now, however, \(\lambda\) is a function of \(p\), so the above gives a self-consistency equation for \(p\)

\[
p = \frac{1}{2} \left[ \tanh (\beta N p + \alpha) + 1 \right].
\]

This equation is identical to the one found by Park and Newman \([13]\) by solving the model exactly in the dense regime, i.e. for \(p = \mathcal{O}(1)\), and by setting \(\beta N = 2J\)

\[
p = \frac{1}{2} \left[ \tanh (2J p + \alpha) + 1 \right].
\]

As noted in \([13]\), a unique solution to (19) exists only for \(J < 1\). For \(J \geq 1\) and \(\alpha\) sufficiently close to \(-J\) there are three solutions, with only the outer two being stable, leading to a degeneracy in the solution.

Expansions around the mean-field solution and perturbation theories \([12, 14]\) give for the first two moments of the degree distribution.
where the second terms on the rhs, originate from the Gaussian fluctuations about the mean-field solution. These are subleading for large $N$, in the dense regime where $p = O(1)$ and $J = O(1)$. However, in the finite connectivity regime where $p = O(N^{-1})$ and $J = O(1)$ (see appendix A), Gaussian fluctuations are no longer small fluctuations about the leading order statistics. Upon setting $p = c/N$ and sending $N \rightarrow \infty$ at constant $\beta$ and $c$, we get

$$\langle k \rangle = c \left[ 1 + \frac{\beta}{(1 - 2\beta c)(1 - \beta c)} \right]$$

(22)

$$\langle k^2 \rangle = c \left[ 1 + \frac{1}{(1 - 2\beta c)(1 - \beta c)} \right]$$

(23)

showing that the mean-field approximation becomes inexact in the finite connectivity regime. For a full discussion of mean-field predictions in the finite connectivity regime see appendix A.

2.4. Upper and lower bounds on the total number of stars in the finite connectivity regime

Before solving the model in the finite connectivity regime, it is useful to derive expressions for the upper and lower physical bounds on the number of stars that the model can exhibit at a given finite connectivity. The expected number of two-stars is $S = N\langle k^2 \rangle - \langle k \rangle^2 / 2$. For a fixed number of edges $L$, the total number of stars is minimized by minimizing the second moments of the degree distribution while keeping the first moment fixed, i.e. by making the degree distribution regular, so that $\langle k^2 \rangle = \langle k \rangle^2$ attains its physical minimum. This corresponds to the minimum star density

$$\frac{S_{\text{max}}}{N} = \frac{1}{2} \langle k \rangle (\langle k \rangle - 1).$$

(24)

The total number of stars is maximized by maximizing the second moment while keeping the first moment fixed, resulting in a small number of vertices having very large degrees while all the others have low degrees. To see this, we consider the following iterative process. We pick two vertices $i$ and $j$ and assume that $k_i \geq k_j$. If $j$ has a neighbor which is not already connected to $i$ then this edge is rewired to increase $k_i$ by 1 and decrease $k_j$ by 1. This step is repeated until in every pair, the vertex with the lesser degree has no more ‘spare’ edges to rewire to the other vertex. In the case where $L \leq N - 1$ this always ends with a graph where there is one vertex with degree $L$, $L$ vertices connected to it with degree 1, and any left over vertices having degree 0 (because any other configuration would contain ‘spare’ edges). If we increase $L$ just beyond $N$, we can no longer increase the stars by rewiring to the original hub as that hub is already connected to every other vertex. This causes a new hub to form to take up the extra edges. This process of hubs being born and increasing in size until they have degree $N$ keeps going as $L$ is increased until the graph is complete.

In this work we consider sparse graphs, where the total number of edges $L$ is $O(N)$, so that the maximum star configuration has an $O(1)$ number of vertices with degree $N$, and an $O(N)$ number of vertices with degree $O(1)$. The vertices with degree $N$ give each a contribution to the number of stars $N(N-1)/2$ showing that the maximum density of stars in a finitely
connected graph is
\[ \frac{S_{\text{max}}}{N} \sim O(N), \]  
(25)

Expansions about mean-field solutions (22), (23) show that both moments \( \langle k \rangle, \langle k^2 \rangle \) diverge at the same parameter values, suggesting that it is impossible to achieve the maximum expected star density for any finite value of connectivity. However, higher order (non-Gaussian) fluctuations may become dominant in the finite connectivity regime and the expansions (22) and (23) may get very inaccurate. We will test their accuracy against MCMC simulations and results of the exact calculation in the next section.

3. Exact solution in the finite connectivity regime

In this section, we solve the two-star model exactly, in the thermodynamic limit \( N \to \infty \), in the sparse regime, where \( \langle c_{ij} \rangle = O(N^{-1}) \) \( \forall i, j \), and mean-field solutions are expected to become inexact. First, we rewrite the graph Hamiltonian by performing the sum over \( k \) in (8) and using symmetry of \( c_{ij} = c_{ji} \)

\[ H(\mathbf{c}) = -2\alpha \sum_{i < j} c_{ij} - \beta \sum_{i < j} \left( k_i(\mathbf{c}) + k_j(\mathbf{c}) \right), \]  
(26)

leading to the network distribution

\[ p(\mathbf{c}) = \frac{1}{Z} e^{-H(\mathbf{c})} = \frac{1}{Z} \prod_{i < j} \left[ e^{2\alpha + \beta[k_i(\mathbf{c}) + k_j(\mathbf{c})]} \delta_{c_{ij},1} + \delta_{c_{ij},0} \right], \]  
(27)

where \( \delta_{x,y} \) is the Kronecker delta, taking value 1 for \( x = y \) and 0 otherwise. Our objective is to calculate the normalized logarithm of the partition function

\[ Z = \sum_{\mathbf{c}} e^{-H(\mathbf{c})} = \sum_{\mathbf{c}} \prod_{i < j} \left[ e^{2\alpha + \beta[k_i(\mathbf{c}) + k_j(\mathbf{c})]} \delta_{c_{ij},1} + \delta_{c_{ij},0} \right], \]  
(28)

giving the free-energy density \( f = -\lim_{N \to \infty} N^{-1} \log Z \), in the thermodynamic limit. In order to carry out the sum over graph configurations \( \mathbf{c} \), we insert in the partition function, the following representation of unit \( \delta_{x,y} \) with \( \delta_{\mathbf{k},\mathbf{k}}(\mathbf{c}) = \prod_i \delta_{k_i, k_i(c)} \), and use the Fourier representation of the Kronecher delta

\[ \delta_{x,y} = \int_{-\pi}^{\pi} \frac{d\Omega}{2\pi} e^{-i\Omega(x-y)} \]  
(29)

to transform \( \delta \)-functions in integrals. This leads to

\[ Z = \sum_{\mathbf{k}} \int_{-\pi}^{\pi} \frac{d\Omega}{2\pi} e^{i\Omega \mathbf{k}} \prod_{i < j} \left[ e^{2\alpha + \beta(k_i + k_j)} - i(\Omega_i + \Omega_j) \delta_{c_{ij},1} + \delta_{c_{ij},0} \right], \]  
(30)

where we have denoted \( \Omega = (\Omega_1, ..., \Omega_N), \mathbf{k} = (k_1, ..., k_N) \) and used the symmetry of \( \mathbf{c} \) to manipulate \( \sum_{ij} \Omega_k(c) = \sum_{ij} \Omega_i c_{ij} = \sum_{ij} (\Omega_i + \Omega_j) c_{ij} \). At this stage it is useful to note that the second term in the square brackets of (30) is order \( O(N^{-1}) \) in the finite connectivity regime. To see this, we consider the likelihood \( \langle c_{ij} \rangle = \sum_{\mathbf{c}} c_{ij} p(\mathbf{c}) \) for two nodes \( i, j \) to be connected, where
\[ p(c) = \frac{1}{Z} \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega} \prod_{i<j} \left[ e^{2\alpha + \beta (k_i + k_j - i(\Omega_i + \Omega_j))} \delta_{c,1} + \delta_{c,0} \right]. \] (31)

We get
\[ \langle c_{ij} \rangle = \frac{1}{Z} \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega} \prod_{k<l} \left[ 1 + e^{2\alpha + \beta (k_i + k_j - i(\Omega_i + \Omega_j))} \right] \]
\[ = \left\{ \left( \frac{e^{2\alpha + \beta (k_i + k_j - i(\Omega_i + \Omega_j))}}{1 + e^{2\alpha + \beta (k_i + k_j - i(\Omega_i + \Omega_j))}} \right) \right\}_{k,\Omega}. \] (32)

where the angular brackets denote the average
\[ \langle \cdot \rangle_{k,\Omega} = \frac{1}{Z} \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega} \prod_{k<l} \left[ 1 + e^{2\alpha + \beta (k_i + k_j - i(\Omega_i + \Omega_j))} \right]. \] (33)

Since \( \langle c_{ij} \rangle = O(N^{-1}) \), in the finite connectivity regime, it is convenient to transform \( \alpha \to \tilde{\alpha} = -\frac{1}{2} \log(N/c) \), with \( c = O(N^0) \), to make (32) explicitly \( O(N^{-1}) \)
\[ \langle c_{ij} \rangle \approx \frac{c}{N} \langle e^{2\tilde{\alpha} + \beta (k_i + k_j - i(\Omega_i + \Omega_j))} \rangle_{k,\Omega}. \] (34)

This leads to the partition function
\[ Z = \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega} \prod_{i<j} \left[ 1 + \frac{c}{N} e^{2\tilde{\alpha} + \beta (k_i + k_j - i(\Omega_i + \Omega_j))} \right], \]
\[ = \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega} \exp \left[ \frac{c}{N} \sum_{k<l} e^{2\tilde{\alpha} + \beta (k_i + k_j - i(\Omega_i + \Omega_j))} \right], \] (35)

where in the last equality we used \( 1 + x/N = \exp(x/N) + O(N^{-2}) \). To make progress, we need to achieve a factorization, with respect to site index \( i \), of the sum over the variables \( \{k_i\} \) and the integral over the variables \( \{\Omega_i\} \). To this purpose, we introduce for \( \Omega \in [-\pi, \pi] \) and \( k \in \mathbb{N} \) the functions
\[ P(k, \Omega | k, \Omega) = \frac{1}{N} \sum_{r=1}^{N} h_{k,r} \delta(\Omega - \Omega_r) \] (36)
giving the likelihood to observe the pair \((k, \Omega)\) in a network site \( i \). In the limit of large \( N \), one expects that the physical behavior of the system is determined by the distribution of the \( \{k_i, \Omega_i\} \), rather than their sequence, hence (36) is intuitively the natural observable to consider, in order to probe the system’s behavior. We note, however, that site factorization can be achieved by other functions, like, for example, \( P(k | \Omega, \Omega) = N^{-1} \sum_{r=1}^{N} h_{k,r} e^{-\alpha \Omega} \), although the latter is not directly interpretable as a distribution. Next we introduce, for each \((k, \Omega)\), a \( \delta \)-function enforcing the definition (36) and insert a factor 1 into (35) in the form of an integral over this \( \delta \)-function, written in Fourier representation
\[ 1 = \int dP(k, \Omega) \delta \left[ P(k, \Omega) - P(k, \Omega | k, \Omega) \right] \]
\[ = \int \frac{dP(k, \Omega) d\hat{P}(k, \Omega)}{2\pi/N} e^{\alpha i N \hat{P}(k, \Omega)} \prod_{r=1}^{N} h_{k,r} \delta(\Omega - \Omega_r) \hat{P}(k, \Omega), \] (37)

Discretizing \( \Omega \) in steps of size \( \Delta \) which is eventually sent to zero, we can then write the free-energy as the following path integral, with the short-hand \( dPd\hat{P} = \prod_{k,\Omega} [dP(k, \Omega)d\hat{P}(k, \Omega)] \) and sums over \( \Omega \) transformed into integrals:
\[ f = \lim_{N \to -\infty} \frac{1}{N} \log Z = \lim_{N \to -\infty} \frac{1}{N} \log \int \left\{ d\mathcal{P} \right\} e^{-N\Phi[P, \hat{\varphi}]} \]  

where \n
\[ \Phi[P, \hat{\varphi}] = -i \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega P(k, \Omega) \hat{\varphi}(k, \Omega) \]
\[ \frac{c}{2} \sum_{k, k' \geq 0} \int_{-\pi}^{\pi} d\Omega d\Omega' P(k, \Omega) P(k', \Omega') e^{2\alpha + \beta(k + k') - i(\Omega + \Omega')} \]
\[ - \log \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{\alpha |k|^2 - i\hat{\varphi}(k, \Omega)}. \]

For large \( N \), we can evaluate the path integral in (38) by the saddle-point method

\[ f = \min_{P, \hat{\varphi}} \Phi[P, \hat{\varphi}] \]

giving (see appendix B)

\[ f = -\frac{c}{2} - \log \sum_{k \geq 0} e^{\alpha k + \beta |k|^2} g(k), \]

where

\[ g(k) = \left( \frac{c \langle k \rangle}{k!} \right)^{1 - (k + \epsilon)/2}. \]

and \( \alpha \) and \( \beta \) are to be found from the equations for the constraints, which read

\[ \langle k \rangle = \sum_{k \geq 0} k g(k) e^{\alpha k + \beta |k|^2} \]
\n\[ \sum_{k \geq 0} g(k) e^{\alpha k + \beta |k|^2}, \]

\[ \langle k^2 \rangle = \sum_{k \geq 0} k^2 g(k) e^{\alpha k + \beta |k|^2} \]
\n\[ \sum_{k \geq 0} g(k) e^{\alpha k + \beta |k|^2}. \]

We note that for the choice \( c = \langle k \rangle \), \( g(k) \) becomes a Poissonian distribution with parameter \( \langle k \rangle \).

In conclusion, we have that for the finitely-connected two-star model, with constrained average connectivity \( \langle k \rangle \) and degree variance \( \langle k^2 \rangle \), the network distribution is

\[ p(e) = \frac{1}{Z} \prod_{i<j} \left[ \frac{c}{N} e^{2\alpha + \beta (\langle k \rangle + \langle k \rangle^2)} \delta_{e_{ij}^1} + \delta_{e_{ij}^0} \right]. \]

where \( \alpha \) and \( \beta \) are determined from \( \langle k \rangle = \sum_{k \geq 0} k p(k) \) and \( \langle k^2 \rangle = \sum_{k \geq 0} k^2 p(k) \) with

\[ p(k) = g(k) \frac{e^{\alpha k + \beta |k|^2}}{\sum_{k \geq 0} g(k) e^{\alpha k + \beta |k|^2}}. \]

To test the validity of our equations, we check that for \( \beta = 0 \), we retrieve the results for ER graphs. For \( \beta = 0 \), (43) gives \( \langle k \rangle = ce^{2\alpha} \) and substituting in (45) we get
\[ p(e) = \frac{1}{Z} \prod_{i<j} \left[ \frac{\langle k \rangle}{N} \delta_{e_{ij},1} + \delta_{e_{ij},0} \right] \]  \hspace{1cm} (47)

with

\[ Z = \sum \prod_{i<j} \left[ \frac{\langle k \rangle}{N} \delta_{e_{ij},1} + \delta_{e_{ij},0} \right] = \prod_{i<j} \left[ 1 + \frac{\langle k \rangle}{N} \right]. \]

For large \( N \), this yields

\[ p(e) = \prod_{i<j} \left[ \frac{\langle k \rangle}{N} \delta_{e_{ij},1} + \left( 1 - \frac{\langle k \rangle}{N} \right) \delta_{e_{ij},0} \right] \]  \hspace{1cm} (48)

thus recovering the ER ensemble.

### 3.1. Numerical results

For \( \beta = 0 \), we need to solve equations (43), (44) numerically. We note, however, that for positive values of \( \beta \), the sums on the rhs of these equations do not converge, hence only values \( \beta \leq 0 \) are admissible. The parameter \( c \) can be chosen arbitrarily, so we will set it to unit, without loss of generality.

In figure 1 we compare theoretical results from (43), (44) (orange symbols) with MCMC simulations for networks of \( N = 3000 \) nodes (blue symbols) and predictions from mean-field theory (22) and (23) (green symbols). Plots show the logarithm of the link and of the star densities normalized with the logarithm of the system size, as functions of \( \beta \), for fixed values of \( \alpha = -0.5, 4 \), corresponding to low and high connectivity respectively. MCMC simulations show a divergence in the links and stars densities for \( \beta > 0 \), consistently with the fact that equations (43), (44) do not converge in this regime, and show excellent agreement with theoretical predictions at \( \beta \leq 0 \). In particular, simulations data are on top of theoretical ones at low connectivity (top panel) and deviations stay within finite size effects at high connectivities (bottom panel). In contrast, mean-field predictions are seen to perform well at high connectivity but, as expected, get very inaccurate for small connectivity. In figure 2, we show three dimensional plots of the average connectivity and average density of stars, as functions of \( \alpha \) and \( \beta \). One has that for fixed values of \( \alpha \), both connectivity and star density are at their highest for \( \beta = 0 \) and they decay quickly for \( \beta < 0 \). Notably, the star density is always finite while the connectivity is finite, hence the model fails to produce an arbitrary number of stars for any given finite connectivity. In particular, we find that the star density is always close to its physical minimum. This is better understood by looking at contour plots of constant average connectivity and constant average star density in figure 3. These show that for \( \beta \) fixed the connectivity increases with \( \alpha \) quicker than the star density, so that the star density decreases as we move along the contours of constant connectivity in the direction of decreasing \( \beta \). Hence, the maximum number of stars is obtained for \( \beta = 0 \), which corresponds to ER graphs, satisfying \( \langle k^2 \rangle - \langle k \rangle = \langle k \rangle^2 \), so that the contour of constant connectivity and star density coincide. The reason for this behavior can be understood by looking at equation (46), showing that for \( \beta \leq 0 \), \( p(k) \) is Poissonian \((\beta = 0)\) or narrower \((\beta < 0)\), whereas for \( \beta > 0 \), \( p(k) \) is not normalizable, with an asymptotic behavior for large-\( k \) given by \( p(k) \sim e^{\xi (k - \log k)} \), independent of \( \alpha \).

This shows that for \( \langle k^2 \rangle > \langle k \rangle^2 + \langle k \rangle \), equations (43) and (44) have no solution. We will see in the next section that in this range of the imposed constraints, the partition function can no longer be calculated by the saddle-point method illustrated above and a different approach is needed. This leads to a phase diagram, in the \( \langle k \rangle - \langle k^2 \rangle \) plane, with two phases which display different behaviors of the partition function \( Z \) in the large \( N \)-limit. The critical line
separating the two phases is found by solving (43) and (44) for \( \beta = 0 \), which gives
\[
\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle.
\]
For \( \langle k^2 \rangle < \langle k \rangle^2 + \langle k \rangle \)
\[
Z \sim e^{-Nf(\langle k \rangle, \langle k^2 \rangle)},
\]
where \( f(\langle k \rangle, \langle k^2 \rangle) \) is given by (41), whereas for \( \langle k^2 \rangle > \langle k \rangle^2 + \langle k \rangle \) the asymptotic behavior of \( Z \) is given by (55). We will show in the next section that the phase transition occurring at \( \langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle \) is a condensation from a liquid to a condensed phase, due to a large deviation of sums of random variables. The asymptotic behavior of \( p(k) \) suggests that condensation may be avoided in the scaling regime \( \beta = \mathcal{O}(\log N/N) \), that is the same where Strauss model is known to display a non-trivial behavior [11].

Figure 1. Plot of \( \log \langle k \rangle / \log N \) (left panels) and \( \log \langle k^2 \rangle - \langle k \rangle / \log N \) (right panels) as functions of the ensemble parameter \( \beta \), for \( \alpha = -0.5 \) (top panels) and \( \alpha = 4 \) (bottom panels) and \( \kappa = 1 \). MCMC denote Monte Carlo simulations for \( N = 3000 \), FC denotes exact results from formulae (43), (44) and PN denotes predictions from expansions about mean-field theory (22), (23).

Figure 2. Plot of the average connectivity (left panel) and the average density of stars (right panel) as functions of the ensemble parameters \( \alpha, \beta \).
4. Condensation as large deviation of sums of random degrees

In this section we show that the condensation transition occurring at \( \langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle \) is related to a large deviation of sums of random degrees and we provide the asymptotics of \( Z \) in the condensate phase \( \langle k^2 \rangle > \langle k \rangle^2 + \langle k \rangle \). To this purpose, it is convenient to use an alternative but equivalent definition of the two-star model, where the constraints on the links and stars are implemented directly in the network distribution

\[
\frac{1}{Z(N(k), N(k^2))} \prod_{i<j} \left[ \frac{c}{N} \delta_{\epsilon,1} + \left( 1 - \frac{c}{N} \right) \delta_{\epsilon,0} \right] 
\times \begin{vmatrix}
\sum_i k_i(c) - N(k) \\
\sum_i k_i^2(c) - N(k^2)
\end{vmatrix}
\]

(50)

via Kronecker deltas and links are drawn otherwise randomly and independently with likelihood \( c/N \). By using path integrals and the saddle point method used in section 3 we find that in the fluid phase \( \langle k^2 \rangle < \langle k \rangle^2 + \langle k \rangle \) the partition function \( Z(N(k), N(k^2)) \) has the large deviation behavior (49) with rate function given by the free-energy density (see appendix C for full details)

\[
f(\langle k \rangle, \langle k^2 \rangle) = \alpha(\langle k \rangle) + \beta(\langle k^2 \rangle) - \log \sum_k e^{\epsilon k + \epsilon k^2} g(k),
\]

(51)

where \( \alpha, \beta \) are determined from equations (43), (44) and \( g(k) \) is defined in (42). Up to additive constants, the free-energy density (51) is identical to (41), hence the definition (50) of the two-star model is thermodynamically equivalent to the standard definition (45). In the fluid phase, the marginal distribution \( p(k) \), calculated in appendix C, is given by (46), showing that all random degrees contribute to the sums \( \sum_i k_i, \sum_i k_i^2 \) with small values (as \( \beta < 0 \) for \( \langle k^2 \rangle < \langle k \rangle^2 + \langle k \rangle \)).

Notably, the free-energy (51) and the marginal distribution (46) are identical to those of the system of independently and identically distributed random variables \( k = \{k_1, \ldots, k_N\} \) with ‘bare’ distribution \( \tilde{g}(k) = g(k)/\sum_k g(k) \) and global constraints on the average and the variance, introduced in [25]
The connection between the two models is transparent: a graph drawn from (50) will have a degree sequence \( \mathbf{k} \) where the degrees are random variables drawn independently from the Poissonian distribution \( \hat{g}(k) \) with average \( \sqrt{c \langle k \rangle} \), subject to the two global constraints. In particular, for the choice \( c = \langle k \rangle \) each degree will be a Poissonian variable with average \( \langle k \rangle \), subject to the global constraint of \( \langle k^2 \rangle \).

Systems with factorized steady states (52) are known to display condensation for non-heavy-tailed distributions \( g(k) \sim e^{-k^\gamma} \) with \( 1 \leq \gamma < 2 \), when conditioned on large deviations of their linear statistics, i.e. for \( \langle k^2 \rangle \) larger than a critical value \( \langle k \rangle \), which is found by solving (43) and (44) for \( \beta = 0 \) [25], i.e. from

\[
\sigma(\langle k \rangle) = \frac{\sum_k k^2 g(k)e^{\alpha k}}{\sum_k g(k)e^{\alpha k}},
\]

with \( \alpha \) solving

\[
\langle k \rangle = \frac{\sum_k kg(k)e^{\alpha k}}{\sum g(k)e^{\alpha k}}.
\]  

For \( g(k) \) defined in (42), the critical value \( \sigma(\langle k \rangle) \) can be calculated exactly: equation (54) gives \( \alpha = \ln \sqrt{\langle k \rangle/c} \), and substituting in (53) we obtain \( \sigma(\langle k \rangle) = \langle k^2 \rangle + \langle k \rangle \).

For \( \langle k^2 \rangle > \sigma(\langle k \rangle) \), the method developed in [25] yields for function (42) with asymptotics \( g(k) \sim e^{-k \log k} \), the following asymptotic behavior for the partition function

\[
Z(N\langle k \rangle, N\langle k^2 \rangle) \sim e^{-N\langle k \rangle} - [N\langle k^2 \rangle - N\sigma(\langle k \rangle)]^\gamma \log [N\langle k^2 \rangle - N\sigma(\langle k \rangle)]^2
\]

with

\[
I(\langle k \rangle) = \alpha \langle k \rangle - \log \sum_{k \geq 0} \hat{g}(k)e^{\alpha k}
\]

and \( \alpha \) determined from (54). In this regime, the marginal \( p(k) \) shows a bump at \( k = O(\sqrt{N}) \), meaning that there is a condensate of size \( O(\sqrt{N}) \) residing on a single site. Sampling networks from the two-star distribution (50) will then lead to graphs which have a bulk of homogeneous degrees and one single site accounting for all the degree heterogeneity in the network, making this model unsuitable as a null model for finitely connected random graphs with constrained average degree and variance. In addition, although Markov processes generating random variables \( \mathbf{k} \) under the two global constraints can be constructed as in [25], the definition of algorithms to sample networks \( \mathbf{c} \) from the measure (50) with the two hard constraints on degree average and variance, in an efficient and unbiased way, poses a challenge in its own right. We note that removing the hard constraint on the variance we obtain the system with factorized states studied in [18]
which is known to exhibit condensation [29] for heavy-tailed bare distributions \( \hat{g}(k) \sim Ak^{-\gamma} \), with \( \gamma > 2 \), for \( \langle k \rangle < \sum_{k>0} k \hat{g}(k) \). For these models, the "dressed" marginal distribution is found to be [18]

\[
p(k) = \frac{\hat{g}(k)e^{\alpha k}}{\sum \hat{g}(k)e^{\alpha k}}
\]

with \( \alpha \) solving \( \langle k \rangle = \sum_k kp(k) \). This suggests that condensation transitions in the two-star model may be avoided by removing the hard constraint on the variance and choosing the bare distribution \( \hat{g}(k) \) in such a way that the 'dressed' marginal displays the desired variance \( \langle k^2 \rangle = \sum_k p(k)k^2 \) while satisfying \( \sum_{k>0} k \hat{g}(k) < \langle k \rangle \).

5. Conclusion

In this work we analyzed the finitely connected two-star model. This model has been solved analytically within mean-field approximations, which predict a second-order phase transition between a symmetric and a symmetry-broken phase and a first-order transition in the link density occurring along the critical line separating the symmetry-broken phases, in the ensemble parameter space. In this work we solved the two-star model exactly, in the thermodynamic limit, in the finite connectivity regime, where mean-field approximations are shown to become inexact. Our results show that in the thermodynamic limit the system undergoes a condensation transition, from a liquid to a condensed phase, related to large deviations in the sum of the random degrees, induced by the global constraints on their linear statistics. We showed that in the liquid phase, the degree statistics exhibited by the model is always in between regular and ER graphs. Our results are in excellent agreement with MCMC simulations and are compared with existing results from mean-field theory and expansions around it. The latter become inaccurate for small connectivity, albeit providing the correct location of the critical line. When the global constraints imposed on the linear statistics insist on a degree statistics more heterogeneous than ER graphs, the model undergoes a condensation transition, whereby a condensate residing on a single site appears in the system, which accounts for all the degree heterogeneity of the network, while the bulk of the network have homogeneous degrees.

We conclude that the finitely connected two-star model is unsuitable, in its standard definition, as a null-model for real networks, with prescribed connectivity and link density. We suggest possible modifications of the model which may lead to non-trivial degree statistics in the finite connectivity regime, which either entail a different scaling of the Lagrange multiplier \( \beta \) constraining the stars or the use of a soft constraint for the star density, while the link density is hardly constrained.

Acknowledgments

A A wishes to thank Peter Sollich for useful discussions.

Appendix A. Mean-field predictions in the finitely connected regime

In the low connectivity regime we demand that \( p = c/N \) with \( c = \mathcal{O}(N^0) \). To ensure that, we see from (19) that we have to set \( \beta = \mathcal{O}(1) \) and scale the ensemble parameter \( \alpha \) as \( \alpha \rightarrow \tilde{\alpha} = 1/2 \log N \) so that the rhs of (19) becomes
where the last equality holds for large $N$ and yields $p = \mathcal{O}(N^{-1})$ as required. Equating this to the LHS of (19) we get the self-consistency equation

$$c = e^{2/\kappa + 2\delta}.$$  

(A.2)

A graphical analysis of (A.2) reveals that for $\beta \leq 0$ there is only one solution, whereas for $\beta > 0$ there can be one, two or no solutions depending on the choice of $\hat{\alpha}$ and $\beta$. To identify the regions of the parameters for which solutions exist, note that the rhs of (A.2) is convex and strictly increasing in $c$. Let $c^*$ be the value of $c$ for which $\frac{\partial}{\partial c}\text{rhs} = 2\beta e^{2/\kappa + 2\delta} = 1$. If the value of the rhs at $c^*$ is greater than $c^*$ then there can be no solutions of the self-consistency equation, while if it is less than $c^*$ there are two, and there is only one solution if the rhs equals $c^*$. We find that $c^*$ is given by

$$2\beta e^{2/\kappa + 2\delta} = 1.$$  

(A.3)

We have that there exists only one solution when $c^* = e^{2/\kappa + 2\delta}$, which using the defining equation of $c^*$ (A.3) simplifies to $c^* = 1/(2\beta)$. Inserting in (A.3) we find that there is a unique solution for $\beta = \frac{1}{2} e^{-1-2\delta}$. We have no solutions when $c^* < e^{2/\kappa + 2\delta} = 1/(2\beta)$ which inserted in (A.3) gives $1 < 2\beta e^{1+2\delta}$ hence $\beta > \frac{1}{2} e^{-1-2\delta}$.

Finally we have two solutions for $c^* > e^{2/\kappa + 2\delta}$, that is in the region $\beta < \frac{1}{2} e^{-1-2\delta}$. This implies that at $\beta = e^{-1-2\delta}/2$ the only solutions of (A.2) is $c = e^{1+2\delta}$, whereas for $\beta < e^{-1-2\delta}/2$ there are two solutions, one smaller and one greater than $e^{1+2\delta}$. As $\beta$ approaches zero the greater solution tends to infinity while the lower solution tends to $e^{2\delta}$. From (17) we have $e^\lambda = 1/(1-p)$ which substituted in (14) gives for the free-energy density $f = F/N = \ln(1-p)(N-1)/2 \simeq -c/2$, for $p = c/N$ and $N \gg c$, showing that the free-energy decreases as the connectivity increases. Hence, the stable solution for $0 \leq \beta \leq e^{-1-2\delta}/2$ will be the one with higher connectivity. In conclusion, the mean-field theory in the finite connectivity regime predicts a critical line at $\beta = 0$ where the average connectivity jumps from $\mathcal{O}(1)$ to $\mathcal{O}(N)$ values. Although the mean-field approximation is invalid in the finite connectivity regime, it picks up the correct location of the critical line $\beta = 0$ found from the exact analysis.

Appendix B. Saddle-point calculation of the free-energy density

In this section we calculate the free-energy density (40). Extremizing the action $\Phi$ over $P$, $\hat{P}$ leads to the saddle-point equations
The equation above suggests to define
\[
P(k, \Omega) = \frac{e^{i\Omega k - i\tilde{P}(k, \Omega)}}{\sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\Omega k - i\tilde{P}(k, \Omega)}}
\]
and
\[
\gamma(\hat{\alpha}, \beta) = \sum_{k \geq 0} e^{\hat{\alpha} + ik} P(k).
\]
Hence,
\[
-i\tilde{P}(k, \Omega) = ce^{\hat{\alpha} + ik} \gamma(\hat{\alpha}, \beta)
\]
and
\[
P(k, \Omega) = \frac{e^{i\Omega k + c\gamma e^{\hat{\alpha} + ik} - i\tilde{P}}}{\sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\Omega k + c\gamma e^{\hat{\alpha} + ik} - i\tilde{P}}}
\]
Using the identity
\[
\int_{-\pi}^{\pi} d\Omega e^{i\Omega k + x(\hat{\alpha}, \beta)e^{-i\Omega}} = \begin{cases} 
2\pi \left[ x(\hat{\alpha}, \beta) \right]^k / k! & \text{if } k \geq 0 \\
0 & \text{if } k < 0
\end{cases}
\]
we get
\[
P(k) = \left( c\gamma e^{\hat{\alpha} + ik} \right)^k / (k-1)! \left[ \sum_{k \geq 0} \left( c\gamma e^{\hat{\alpha} + ik} \right)^k / k! \right]^{-1} \theta \left( k - \frac{1}{2} \right)
\]
where \( \theta(x) \) is the Heaviside step function, taking value 1 for \( x > 0 \) and 0 for \( x < 0 \). Note that \( P(k) \) is not a distribution because it is not normalized to one, instead we have \( \sum_{k \geq 0} P(k) = \langle e^{i\tilde{P}} \rangle \gamma \) with
\[
\langle \cdot \rangle_\gamma = \sum_{k \geq 0} \frac{\left( c\gamma e^{\hat{\alpha} + ik} \right)^k / k!}{\sum_{k \geq 0} \left( c\gamma e^{\hat{\alpha} + ik} \right)^k / k!}.
\]
The resulting free-energy density is
\[
f(\hat{\alpha}, \beta) = \frac{c\gamma^2 (\hat{\alpha}, \beta)}{2} - \log \sum_{k \geq 0} \left[ e^{\hat{\alpha} + ik} c\gamma (\hat{\alpha}, \beta) \right]^k / k!,
\]
where \( \gamma \) solves the self-consistency equation
\[
c\gamma^2 (\hat{\alpha}, \beta) = \langle k \rangle_\gamma.
\]
Finally, the ensemble parameters \( \hat{\alpha}, \beta \), have to be determined from the equations for the constraints which are found by taking the derivatives of \( f \) with respect to \( \hat{\alpha}, \beta \) as
\[
\langle k \rangle = - \frac{\partial f}{\partial \alpha} = - \frac{\partial f}{\partial \beta} = -c^2 \frac{\partial \gamma}{\partial \alpha} + \langle k \rangle \gamma + \frac{1}{\gamma} \langle k \rangle \gamma \frac{\partial \gamma}{\partial \beta} = \langle k \rangle \gamma \quad (B.11)
\]

and
\[
\langle k^2 \rangle = - \frac{\partial f}{\partial \beta} = -c^2 \frac{\partial \gamma}{\partial \beta} + \langle k^2 \rangle \gamma + \frac{1}{\gamma} \langle k \rangle \gamma \frac{\partial \gamma}{\partial \beta} = \langle k^2 \rangle \gamma, \quad (B.12)
\]

where the last equality in (B.11) and (B.12) follows from the saddle-point equation (B.10). Combining (B.11) and (B.10) we have \( \gamma = \sqrt{\langle k \rangle / c} \), yielding the below equations for the ensemble parameters
\[
\langle k \rangle = \sum_{k > 0} k \left( \sqrt{c \langle k \rangle} e^{\alpha + \beta} \right)^k / k! \quad (B.13)
\]
\[
\langle k^2 \rangle = \sum_{k > 0} k^2 \left( \sqrt{c \langle k \rangle} e^{\alpha + \beta} \right)^k / k! \quad (B.14)
\]

and leading to the free-energy given in (41).

Appendix C. Full derivation of the equivalent model in the liquid phase

In this section we calculate the normalizing constant \( Z(\langle k \rangle, N \langle k^2 \rangle) \) of the distribution (50), which provides an equivalent definition of the two-star model (45), as well as the marginal distribution
\[
p(k | e) = \frac{1}{N} \sum_i \delta_{k, k_i(e)}. \quad (C.1)
\]

First we focus on \( Z(\langle k \rangle, N \langle k^2 \rangle) \). Using the identity \( 1 = \sum_k \delta_{k, k_i(e)} \)
\[
Z \left( \langle k \rangle, N \langle k^2 \rangle \right) = \sum_k \sum_{\epsilon} e \delta_{k, k_i(e)} \prod_{i, j} \left[ \frac{c}{N} \delta_{\epsilon_{ij}, 1} + \left( 1 - \frac{c}{N} \right) \right]
\]
\[
\times \left( \sum_i \delta_{k, k_i(e)} - N \langle k \rangle \right) \left( \sum_i \delta_{k^2, k_i(e)} - N \langle k^2 \rangle \right) \quad (C.2)
\]

and the Fourier representation of the Kronecker deltas, we get
\[
Z \left( \langle k \rangle, N \langle k^2 \rangle \right) = \int_{-\pi}^\pi \frac{\text{d}^N \omega}{(2\pi)^N} e^{i\omega N \langle k \rangle + i\omega N \langle k^2 \rangle} \sum_k \int_{-\pi}^\pi \frac{\text{d} \Omega}{(2\pi)^N} e^{i\Omega k} e^{-i\omega \sum_i k_i(e) \delta_{\epsilon_{ij}, 1}} \left( \frac{c}{N} \delta_{\epsilon_{ij}, 1} + \left( 1 - \frac{c}{N} \right) \right) \quad (C.3)
\]

We next introduce the following order parameters to achieve site factorization
\[
P(\Omega | \Omega) = \frac{1}{N} \sum_{r=1}^N \delta (\Omega - \Omega_r) \quad (C.4)
\]
and insert into (C.3) for each $\Omega$ the following integral:

$$1 = \int dP(\Omega) \delta[P(\Omega) - P(\Omega|\Omega)]$$

$$= \int dP(\Omega) d\hat{P}(\Omega) \frac{e^{\alpha N P(\Omega) P(\Omega) - i \sum_{\subseteq} \delta(\Omega - \Omega')}}{2\pi N^{2}}.$$  \hfill (C.5)

Discretizing $\Omega$ in steps of size $\Delta$ which is eventually sent to zero, and proceeding as in section 3, we can then write the free-energy density as the path integral

$$f = - \lim_{N \to \infty} \frac{1}{N} \log \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{iN(\omega N(k) + \omega N(k'))}$$

$$\times \int \{ dP d\hat{P} \} e^{i\int d\Omega P(\Omega) \hat{P}(\Omega)} + \frac{\alpha}{2} \left[ \int d\Omega d\Omega' P(\Omega) P(\Omega') e^{-i(\Omega + \Omega')} \right]$$

$$\times \prod_{i} \sum_{k_i} \int \frac{d\Omega}{2\pi} e^{i\hat{P}(\omega - \omega') = - \lim_{N \to \infty} \frac{1}{N} \log \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \int \{ dP d\hat{P} \} e^{-N\Phi[P,\hat{P},\omega,\omega'],}$$  \hfill (C.6)

where

$$\Phi[P, \hat{P}, \omega, \omega'] = - i\omega \langle k \rangle$$

$$- i\omega' \langle k^2 \rangle$$

$$- i \int_{-\pi}^{\pi} d\Omega P(\Omega) \hat{P}(\Omega)$$

$$- \frac{\alpha}{2} \int d\Omega d\Omega' P(\Omega) P(\Omega') e^{-i(\Omega + \Omega')}$$

$$- \log \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{i\hat{P}(\omega - \omega') = - \lim_{N \to \infty} \frac{1}{N} \log \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} \int \{ dP d\hat{P} \} e^{-N\Phi[P,\hat{P},\omega,\omega'],}$$  \hfill (C.7)

For large $N$, we can evaluate the integral by steepest descent

$$f = \min_{P, \hat{P}, \omega, \omega'} \Phi[P, \hat{P}, \omega, \omega'].$$  \hfill (C.8)

Extremizing the action $\Phi$ over $P, \hat{P}$ we obtain, proceeding as in section 3,

$$c\gamma^2 = \langle k \rangle \gamma$$  \hfill (C.9)

with

$$\langle \cdot \rangle = \sum_{k \geq 0} \frac{(c\gamma)^k e^{-i\omega k - i\omega' k^2} / k!}{\sum_{k \geq 0} (c\gamma)^k e^{-i\omega k - i\omega' k^2} / k!}.$$  \hfill (C.10)

Extremizing $\Phi$ over $\omega, \omega'$ we obtain

$$\langle k \rangle = \langle k \rangle \gamma, \quad \langle k^2 \rangle = \langle k^2 \rangle \gamma.$$  \hfill (C.11)

Substituting the saddle point equations in the free-energy we finally get

$$f = - i\omega \langle k \rangle - i\omega' \langle k^2 \rangle - \log \sum_{k \geq 0} \int_{-\pi}^{\pi} d\Omega e^{-i\omega k - i\omega' k^2} \langle g(k) \rangle$$  \hfill (C.12)
with
\[ g(k) = \frac{\left( \sqrt{c(k)} \right)^k e^{-(c(k)+k)/2}}{k!} \]  
(C.13)
and \( \omega, \omega' \) determined from
\[ \langle k \rangle = \frac{\sum_{k \geq 0} kg(k) e^{-i\omega k - i\omega'^2 k}}{\sum_{k \geq 0} g(k) e^{-i\omega k - i\omega'^2 k}}, \]  
(C.14)
\[ \langle k^2 \rangle = \frac{\sum_{k \geq 0} k^2 g(k) e^{-i\omega k - i\omega'^2 k}}{\sum_{k \geq 0} g(k) e^{-i\omega k - i\omega'^2 k}}. \]  
(C.15)
Setting \(-i\omega = \alpha, -i\omega' = \beta\) yields (51).

Next, we turn to the marginal distribution (C.1). In the limit of large \( N \), we expect this quantity to converge to its ensemble average

\[ p(k) = \langle p(k|e) \rangle = \int \frac{dx}{2\pi} e^{ikx} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N\langle k \rangle + i\omega' N\langle k^2 \rangle} \]
\[ \times \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega - i\omega \sum_k k_l - i\omega' \sum_k k_l^2} \]
\[ \times \sum_{k, l} \prod_{k < l} \left[ \frac{C}{N} e^{-i(\Omega_k + \Omega_l) - i\xi (\delta_{kl} + \delta_{lk})} \delta_{kl, 1} + \left( 1 - \frac{C}{N} \right) \delta_{kl, 0} \right] \]
\[ = \int \frac{dx}{2\pi} e^{ikx} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N\langle k \rangle + i\omega' N\langle k^2 \rangle} \]
\[ \times \sum_k \int_{-\pi}^{\pi} \frac{d\Omega}{(2\pi)^N} e^{ik\Omega - i\omega \sum_k k_l - i\omega' \sum_k k_l^2} \]
\[ \times \exp \left\{ \sum_{k < l} \frac{C}{N} \left[ e^{-i(\Omega_k + \Omega_l) - i\xi (\delta_{kl} + \delta_{lk})} - 1 \right] \right\}. \]  
(C.16)

First, we work out the curly brackets as
\[ \sum_{k < l} \frac{C}{N} \left[ e^{-i(\Omega_k + \Omega_l) - i\xi (\delta_{kl} + \delta_{lk})} - 1 \right] = \sum_{k < l} \frac{C}{N} \left[ e^{-i(\Omega_k + \Omega_l)} \left( e^{-i\xi} - 1 \right) \left( \delta_{kl} + \delta_{lk} \right) + 1 \right] - 1. \]  
(C.17)

Inserting the order parameter (C.4) via path integrals as done in (C.6) we arrive at
\[ p(k) = \int \frac{dx}{2\pi} e^{ikx} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N\langle k \rangle + i\omega' N\langle k^2 \rangle} \]
\[ \times \left\{ \int \left( d\Omega \hat{P}(\Omega) \right)^N e^{i\Omega \sum_{k \geq 0} k_l - i\omega \sum_k k_l^2 - i\xi \sum_{k \geq 0} k_l^2 - i\omega' \sum_k k_l^2} \hat{P}(\Omega) \right\} \]
\[ \times \frac{1}{N} \sum_{k \geq 0} \sum_{l \geq 0} \int \frac{d\Omega}{(2\pi)^N} e^{ik\Omega - i\omega \sum_k k_l - i\omega' \sum_k k_l^2} \hat{P}(\Omega) \]
\[ = \int \frac{dx}{2\pi} e^{ikx} \int_{-\pi}^{\pi} \frac{d\omega d\omega'}{4\pi^2} e^{i\omega N\langle k \rangle + i\omega' N\langle k^2 \rangle} \]
\[ \times \int \{ dP \hat{d} \hat{p} \} e^{iN \int d\Omega \psi(\Omega) \hat{p} + \frac{i}{N} \sum_{\ell} \int d\Omega_{\ell} e^{-i\ell l_{\ell} - wh_{\ell} h_{\ell}^{2} - i\hat{p}(\Omega) + \frac{\delta}{\Omega}(e^{\alpha} - 1) \int d\Omega P(\Omega)e^{-\alpha}} \]

\[ \times \frac{1}{N} \sum_{\ell} \int d\Omega_{\ell} e^{-i\ell l_{\ell} - wh_{\ell} h_{\ell}^{2} - i\hat{p}(\Omega)} \]

\[ \times \prod_{i} \int d\Omega e^{-i\ell l_{i} - w_{i} h_{i}^{2} - i\hat{p}(\Omega)} \]

\[ = \int \frac{dx}{2\pi} e^{	ext{i}ax} \int_{-\pi}^{\pi} \frac{dx' dx''}{4\pi^{2}} \int \{ dP \hat{d} \hat{p} \} e^{-N\Phi(P, \hat{p}, \omega')} \]

\[ \times \sum_{\ell} \int d\Omega e^{i\ell l_{\ell} - wh_{\ell} h_{\ell}^{2} - i\hat{p}(\Omega)} \]

\[ \frac{\delta}{\Omega}(e^{\alpha} - 1) \int d\Omega P(\Omega)e^{-\alpha}}, \quad \text{(C.18)} \]

where \( \Phi(P, \hat{p}, \omega, \omega') \) is as in (C.7). We next calculate the integrals over \( \omega, \omega', P, \hat{p} \) by steepest descent. At the saddle point we have

\[ -i\hat{p}(\Omega) = e^{-i\Omega} \sqrt{e(k)} \]

and we obtain

\[ p(k) = \int \frac{dx}{2\pi} e^{iak} \sum_{q} \int d\Omega e^{i\Omega q - i\omega q^{2} + \sqrt{e(k)} e^{-i\omega}} \]

\[ \sum_{q} \int d\Omega e^{i\Omega q - i\omega q^{2} + \sqrt{e(k)} e^{-i\omega}}, \quad \text{(C.20)} \]

where \( \omega, \omega' \) solve (C.14), (C.15). Performing the \( \Omega \)-integrals we have

\[ p(k) = \int \frac{dx}{2\pi} e^{iak} \sum_{q} e^{-i\omega q - i\omega q^{2} / q!} e^{-i\omega q^{2} / q!} \]

\[ \sum_{q} e^{-i\omega q - i\omega q^{2} / q!} \]

\[ \text{(C.21)} \]

and carrying out the integration over \( x \) finally gives

\[ p(k) = e^{-iak - i\omega k^{2}} \left( \frac{\sqrt{e(k)}}{k!} \right) \]

\[ \sum_{q} e^{-i\omega q - i\omega q^{2}} \left( \frac{\sqrt{e(k)}}{q!} \right) \]

\[ \text{(C.22)} \]

with \( g(k) \) defined in (42). Setting \( -i\omega = \alpha, -i\omega' = \beta \) finally yields (46).

References

[1] Caimo A and Friel N 2011 Bayesian inference for exponential random graph models Soc. Netw. 33 41–55
[2] Caimo A and Friel N 2013 Bayesian model selection for exponential random graph models Soc. Netw. 35 11–24
[3] Sanil A, Banks D and Carley K 1995 Models for evolving fixed node networks: model fitting and model testing Soc. Netw. 17 65–81
[4] Bollobás B 1980 A probabilistic proof of an asymptotic formula for the number of labeled regular graphs Eur. J. Comb. 1 311–6
[5] Anderson C J, Wasserman S and Croach B A 2002 p* primer: logit models for social networks Soc. Netw. 21 37–66
[6] Banks D and Carley K M 1996 Models of social network evolution J. Math. Sociol. 21 173–96
[7] Strauss D and Ikeda M 1990 Pseudolikelihood estimation for social networks J. Am. Stat. Assoc. 85 204–12
[8] Bender E and Canfield E 1978 The asymptotic number of labelled graphs with given degree sequences J. Comb. Theory A 24 296
[9] Bizhani G, Paczuski M and Grassberger P 2012 Statistical mechanics of networks Phys. Rev. E 86 011128
[10] Austad H and Friel N 2010 Deterministic bayesian inference for the $p^*$ model AISTATS, JMLR 9 41–48
[11] Jonasson J 1999 The random triangle model J. Appl. Probab. 36 852–67
[12] Park J and Newman M E J 2004 Solution of the 2-star model of a network Phys. Rev. E 70 066146
[13] Park J and Newman M E J 2004 Statistical mechanics of networks Phys. Rev. E 70 066117
[14] Park J and Newman M E J 2005 Solution for the properties of a clustered network Phys. Rev. E 72 026136
[15] Besag J E 1975 Statistical analysis of non-lattice data Statistician 24 179–95
[16] Molloy M and Reed B 1995 A critical point for random graphs with a given degree sequence Random Struct. Algorithms 6 161
[17] Handcock M S 2003 Statistical Models for Social Networks: Degeneracy and Inference (Washington, DC: Academic) pp 229–40
[18] Evans M R, Majumdar S N and Zia R K P 2006 Canonical analysis of condensation in factorised steady states J. Stat. Phys. 123 357–90
[19] Frank O and Strauss D 1986 Markov graphs J. Am. Stat. Assoc. 81 832–42
[20] Erdős P and Rényi A 1959 On random graphs i Publ. Math. 6 290
[21] Pattison E P and Robins G L 2002 Neighbourhood-based models for social network Sociological Methodol. 32 301–37
[22] Holland P W and Leinhardt S 1981 An exponential family of probability distributions for directed graphs J. Am. Stat. Assoc. 76 33–50
[23] Wasserman S and Robins G L 2005 An introduction to random graphs, dependence graphs, and $p^*$ Models and Methods in Social Network Analysis (Cambridge: Cambridge University Press) pp 401–25
[24] Wasserman S and Pattison P 1996 Logit models and logistic regression for social networks: and introduction to markov graphs and $p^*$ Psychometrika 61 401–25
[25] Szavits-Nossan J, Evans M R and Majumdar S N 2014 Condensation transition in joint large deviations of linear statistics J. Phys. A: Math. Theor 47 455004
[26] Pattison T A B 2006 Snijders amd P E, Robins G L, and Handcock M S. New specifications for exponential random graph models Sociological Methodol. 36 99–153
[27] Snijders T A B 2001 The statistical evaluation of social network dynamics In Sociological Methodology (Boston: Blackwell)
[28] Snijders T A B 2002 Markov chain monte carlo estimation of exponential random graph model J. Soc. Struct. 2 1–40
[29] Dorogovtsev S N, Mendes J F F and Samukhin A N 2003 Principles of statistical mechanics of uncorrelated random networks Nucl. Phys. B 666 396–416

21