Exactly Solvable Random Graph Ensemble with Extensively Many Short Cycles

Fabián Aguirre López[1,2], Paolo Barucca[3,4], Mathilde Fekom[5], and Anthony CC Coolen[1,2]

[1]Department of Mathematics, King's College London, The Strand, London WC2R 2LS, United Kingdom
[2]Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, London SE1 1UL, United Kingdom
[3]London Institute for Mathematical Sciences, 35a South St, Mayfair, London W1K 2XF, United Kingdom
[4]University of Zürich, Department of Banking and Finance, Zürich, ZH, Switzerland
[5]UFR de Physique, Université Paris Diderot (Paris 7), 5 Rue Thomas Mann, 75013 Paris, France

E-mail: fabian.aguirre_lopez@kcl.ac.uk, paolo.barucca@bf.uzh.ch, mathilde.fekom1@yahoo.fr, ton.coolen@kcl.ac.uk

Abstract. We introduce and analyse ensembles of 2-regular random graphs with a tuneable distribution of short cycles. The phenomenology of these graphs depends critically on the scaling of the ensembles' control parameters relative to the number of nodes. A phase diagram is presented, showing a second order phase transition from a connected to a disconnected phase. We study both the canonical formulation, where the size is large but fixed, and the grand canonical formulation, where the size is sampled from a discrete distribution, and show their equivalence in the thermodynamical limit. We also compute analytically the spectral density, which consists of a discrete set of isolated eigenvalues, representing short cycles, and a continuous part, representing cycles of diverging size.

PACS numbers: 64.60.aq, 02.10.Ox, 64.60.De

1. Introduction

Models with pairwise interacting elements are ubiquitous in physics and are sufficient to capture the phenomenology of many systems, ranging from condensed matter via biology to the social sciences and informatics. The properties of the network of interactions strongly affects the properties of the system under study, and hence the analysis of networks is central in modern physics. Testing statistically whether a specific network property influences the dynamics of a system requires sampling networks where such a property is controllable, and for which the probability measure is known. This is why random graph ensembles, especially maximum entropy ones from which it is possible to sample networks systematically with controlled properties, have gained increasing popularity. They range from degree configuration models [1–6], i.e. where the number of connections or nodes are fixed, to more complicated models where a block structure is given to the nodes in the network, as in stochastic block models [7], or other ensembles where clustering, i.e. the tendency of nodes with common neighbours to be connected, is enhanced [8–15]. However, controlling analytically and numerically second or higher-order properties of networks, i.e. node properties that not only depend on first neighbours, such as the density of cycles, is still a great mathematical and analytical challenge, whose range of applications continues to grow [10, 16–22].

So far, nearly all analytical results obtained for random graph ensembles rely on the assumption of the absence of short cycles, the *tree-like approximation*, and we have analytical solutions only for random graphs were clustering is absent or too weak (or improbable) to be relevant. One of the first random graph ensembles in literature to include short cycles was [23], where a term depending on the number of 3-cycles of the graph was included as a modification to the well known Erdös-Rényi model (ER). This was done in order to encourage this connection transitivity in the graph. However, as was found in simulations [23] and in a more rigorous way in [8, 11], unless the graph is particularly small, this approach does not allow for a tuneable number of triangles. Depending on the values and the scaling of the parameters, the model of [23] either stays in a phase very close to the ER model, with a very slight increase in triangles, or it collapses to a condensed phase, where the complete clique has probability one. This abrupt transition was found to be a generic feature of exponential random graph models. As was shown in [24, 25], this phenomenon will be observed not only in two-parameter models like the Strauss model, but in any exponential graph ensemble that is biased such as to induce a finite number of subgraph densities.

The natural way to prevent clique formation in the condensed phase is to study random graph ensembles with hard degree constraints. Here all graphs have exactly the same degree distribution, and this distribution is chosen such that the complete clique is not an allowed state. However, this constraint makes analytical solution intractable, leaving numerical sampling from the ensemble as the only route for investigation. Examples are the Poisonnian graphs studied numerically in [26], where it was found that a triangle bias induced finite size graphs to break down into small clusters to maximize the triangle density. Regular graphs with triangle bias were numerically explored in [6], and showed similar phenomenology. However, both Poissonnian and regular graph enembles with triangle bias have so far resisted analytical solution.

In this paper we introduce an exactly solvable ensemble of 2-regular random graphs, with an exponential measure that controls the presence of short cycles up to

any finite length. The imposition of 2-regularity removes the possibility of a complete clique forming, and forces the graph instead to be partitioned into a set of disconnected cycles of different lengths. This makes the ensemble analytically solvable and perfectly tuneable. The model displays a second-order transition, from a phase dominated by extensively long cycles, to a phase where only (extensively many) cycles of short lengths are present.

In section 2 we introduce and solve the model, in its canonical formulation. In section 3 we describe analytically the phases of the ensemble and the critical hypersurface in the space of parameters; from this result we also compute analytically the spectral density of the ensemble. In section 4 we demonstrate the equivalence of the canonical and grand canonical formulations of the model, and in section 5 we show the agreement of the analytical predictions with numerical experiments. In a final discussion section we summarize the results and delineate the future directions for this research, which are twofold. The first is to relax the 2-regularity constraint of the ensemble, in order to make it more directly comparable to realistic networks. The second is to understand better recent analytical approaches to random graph ensembles that involve constraints on the number of closed paths of all lengths, which is equivalent to constraining random graphs via their spectra [27].

2. Definitions

We define a random graph ensemble over the set of undirected simple regular graphs of degree 2, which we denote by $\mathcal{G}_N$. Any graph in $\mathcal{G}_N$ is necessarily a set of disjoint cycles. The probability assigned to each graph $\mathbf{A} \in \mathcal{G}_N$ is chosen proportional to the exponential of a weighted sum of the number of triangles, squares, pentagons, ... , K-cycles present in $\mathbf{A}$. We refer to this as biasing with respect of the number of short cycles. Thus

$$p(\mathbf{A}) = \frac{1}{Z_N(\boldsymbol{\alpha})} \exp\left(\sum_{\ell=3}^{K} \ell \alpha_\ell n_\ell(\mathbf{A})\right), \tag{1}$$

Here $n_\ell(\mathbf{A})$ denotes the number of length-ℓ cycles, i.e. closed paths of length ℓ without backtracking and without over-counting, and $\boldsymbol{\alpha} = (\alpha_3, \ldots, \alpha_K) \in \mathbb{R}^{K-2}$ is a vector of control parameters. Note that isolated nodes ($\ell = 1$) and dimers ($\ell = 2$) cannot occur due to the degree constraint. The factors ℓ in (1) are included for later convenience. We are effectively biasing with respect to the total number of ℓ-cycles starting at a given node through the introduction of the field α_ℓ.

The partition function $Z_N(\boldsymbol{\alpha})$ is given by

$$Z_N(\boldsymbol{\alpha}) = \sum_{\mathbf{A} \in \mathcal{G}_N} \exp\left(\sum_{\ell=3}^{K} \ell \alpha_\ell n_\ell(\mathbf{A})\right). \tag{2}$$

Expression (1) defines a maximum entropy random graph ensemble with respect to the $K-2$ observables $n_\ell(\mathbf{A})$, whose ensemble averages are controlled by varying the parameters $\boldsymbol{\alpha}$. We choose K to be a fixed number for all values of N. This exponential form is a particular version of the one presented in equation (1.1) of [24]. It is an ensemble where we are interested in controlling the expected values of a finite number of graph observables.

The average fraction of the N nodes that will be found in an ℓ-cycle is given by

$$m_\ell = \frac{\ell}{N} \langle n_\ell(\mathbf{A}) \rangle. \tag{3}$$

where $\langle f(\mathbf{A})\rangle = \sum_{\mathbf{A}} p(\mathbf{A}) f(\mathbf{A})$. Following the statistical mechanics route, we define a generating function $\phi_N(\boldsymbol{\alpha})$:

$$\phi_N(\boldsymbol{\alpha}) = N^{-1} \log[Z_N(\boldsymbol{\alpha})/N!]. \tag{4}$$

The main quantities of interest (3) for our graph ensemble (1) can be computed from (4) via

$$m_\ell = \partial \phi_N(\boldsymbol{\alpha})/\partial \alpha_\ell. \tag{5}$$

The generator $\phi_N(\boldsymbol{\alpha})$ is minus the free energy density, apart from a factor $N!$ which reflects (topologically irrelevant) node label permutations. Including this factor will ensure that the limit $\phi(\boldsymbol{\alpha}) = \lim_{N\to\infty} \phi_N(\boldsymbol{\alpha})$ exists.

3. Analytical solution

3.1. Summation over graphs

To evaluate the partition function (2) we need to perform a sum over graphs. Such sums are usually not analytically tractable, especially when the ensemble definition involves cycles, as is the case in (1). Here we are able to perform the summation by rewriting it as

$$Z_N(\boldsymbol{\alpha}) = \sum_{\boldsymbol{n}} D(\boldsymbol{n}) \mathrm{e}^{\sum_{\ell=3}^K \ell \alpha_\ell n_\ell}, \tag{6}$$

with $\boldsymbol{n} = (n_3, \ldots, n_N) \in \mathbb{N}^{N-2}$. This decomposition reflects the fact that, in the particular case of $\mathcal{G}_N$, we are fortunate that each graph has to be a collection of cycles, and can therefore be identified fully by a sequence $\boldsymbol{n} = (n_3, \ldots, n_N)$ that specifies the number of cycles of each possible length up to N, and a labelling of the nodes. The sum over graphs is then performed by summing over all possible sequences $\boldsymbol{n}$, keeping track of the multiplicity of each sequence via an associated density of states $D(\boldsymbol{n})$:

$$D(\boldsymbol{n}) = \sum_{\mathbf{A}\in\mathcal{G}_N} \prod_{\ell=3}^N \delta_{n_\ell, n_\ell(\mathbf{A})} = \frac{N!\ \delta_{N,\sum_{\ell=3}^N \ell n_\ell}}{\prod_{\ell=3}^N [(2\ell)^{n_\ell} n_\ell!]}. \tag{7}$$

Apart from the condition $N = \sum_{\ell=3}^N \ell n_\ell$, this density is proportional to $N!$ but corrected for over-counting due to the indistinguishability of different length-ℓ cycles, giving a divisor $n_\ell!$, and due to the different ways one can number the nodes in each ℓ-cycle without altering the graph (ℓ cyclic permutations, plus ℓ anti-cyclic permutations), giving a further divisor $(2\ell)^{n_\ell}$. Using the integral form of the Kronecker delta $\delta_{nm} = \int_{-\pi}^{\pi} (\mathrm{d}\omega/2\pi) \mathrm{e}^{\mathrm{i}\omega(n-m)}$, we can thus write the partition function as

$$\begin{aligned} Z_N(\boldsymbol{\alpha}) &= \sum_{\boldsymbol{n}} \frac{N!}{\prod_{\ell=3}^N [(2\ell)^{n_\ell} n_\ell!]} \left(\prod_{\ell=3}^K \mathrm{e}^{\ell \alpha_\ell n_\ell} \right) \int_{-\pi}^{\pi} \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{\mathrm{i}\omega\left(N - \sum_{\ell=3}^N \ell n_\ell\right)} \\ &= \frac{N!}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\omega\ \mathrm{e}^{\mathrm{i}\omega N} \prod_{\ell=3}^K \left(\sum_{n_\ell \geq 0} \frac{\mathrm{e}^{(\alpha_\ell - \mathrm{i}\omega)\ell n_\ell}}{(2\ell)^{n_\ell} n_\ell!} \right) \prod_{\ell=K+1}^N \left(\sum_{n_\ell \geq 0} \frac{\mathrm{e}^{-\mathrm{i}\omega \ell n_\ell}}{(2\ell)^{n_\ell} n_\ell!} \right) \\ &= \frac{N!}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\omega\ \exp\left(\mathrm{i}\omega N + \sum_{\ell=3}^K \frac{\mathrm{e}^{(\alpha_\ell - \mathrm{i}\omega)\ell}}{2\ell} + \sum_{\ell=K+1}^N \frac{\mathrm{e}^{-\mathrm{i}\omega\ell}}{2\ell} \right). \end{aligned} \tag{8}$$

From this, in combination with (4), we infer that

$$\phi_N(\boldsymbol{\alpha}) = \frac{1}{N}\log\int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{2\pi}\ \mathrm{e}^{Nf_N(\omega,\boldsymbol{\alpha})} \tag{9}$$

with

$$f_N(\omega,\boldsymbol{\alpha}) = \mathrm{i}\omega + \sum_{\ell=3}^{K}\frac{\mathrm{e}^{(\alpha_\ell-\mathrm{i}\omega)\ell}}{2\ell N} + \sum_{\ell=K+1}^{N}\frac{\mathrm{e}^{-\mathrm{i}\omega\ell}}{2\ell N}. \tag{10}$$

An exact expression for (9), valid for any finite N, would require to perform the integral in it. Instead, we proceed in the usual way as in statistical physics. We look at the thermodynamic limit, focusing then on $\phi(\boldsymbol{\alpha}) = \lim_{N\to\infty}\phi_N(\boldsymbol{\alpha})$. This will allow us to calculate the asymptotic expressions for (3), which should differ from the finite size values by $\mathcal{O}(1/N)$ corrections.

The limit $N\to\infty$ of (9) can now be obtained by evaluating the integral over ω in (8) via steepest descent:

$$\phi(\boldsymbol{\alpha}) = \lim_{N\to\infty}\mathrm{extr}_\omega f_N(\omega,\boldsymbol{\alpha}). \tag{11}$$

The extremum is found by solving $\partial f(\omega,\boldsymbol{\alpha})/\partial\omega = 0$.

3.2. Scaling with N of control parameters

We observe that for finite $\{\alpha_\ell\}$ our model cannot exhibit nonzero cycle densities m_ℓ in the infinite size limit, since the $\boldsymbol{\alpha}$-dependent term in (10) vanishes for $N\to\infty$. We are therefore led to redefining the parameters $\boldsymbol{\alpha}$ with a size dependent shift,

$$\alpha_\ell = \tilde{\alpha}_\ell + \frac{1}{\ell}\log(N), \tag{12}$$

where $\tilde{\alpha}_\ell = \mathcal{O}(1)$. An intuitive explanation for this scaling is presented in section 4. We denote the vector of shifted $\mathcal{O}(1)$ control parameters by $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_3,\ldots,\tilde{\alpha}_K)$, and we define $\phi_N(\boldsymbol{\alpha}) = \varphi_N(\tilde{\boldsymbol{\alpha}})$. This implies that for $N\to\infty$ we will have $m_\ell = \partial\varphi(\tilde{\boldsymbol{\alpha}})/\partial\tilde{\alpha}_\ell$, in which now

$$\varphi(\tilde{\boldsymbol{\alpha}}) = \lim_{N\to\infty}\mathrm{extr}_\omega\Big\{\mathrm{i}\omega + \sum_{\ell=3}^{K}\frac{\mathrm{e}^{(\tilde{\alpha}_\ell-\mathrm{i}\omega)\ell}}{2\ell} + \sum_{\ell=K+1}^{N}\frac{\mathrm{e}^{-\mathrm{i}\omega\ell}}{2\ell N}\Big\}. \tag{13}$$

Differentiation of this latter expression reveals that the value ω_N at the extremum is to be solved from

$$1 = \frac{1}{2}\sum_{\ell=3}^{K}\mathrm{e}^{(\tilde{\alpha}_\ell-\mathrm{i}\omega_N)\ell} + \frac{1}{2N}\sum_{\ell=K+1}^{N}\mathrm{e}^{-\mathrm{i}\omega_N\ell}, \tag{14}$$

and that the asymptotic values of the observables m_ℓ are subsequently given by

$$m_\ell = \frac{1}{2}\mathrm{e}^{(\tilde{\alpha}_\ell-\mathrm{i}\omega_N)\ell}. \tag{15}$$

This last identity, in combination with (14), prompts us to introduce $m_\infty = 1 - \sum_{\ell\le K} m_\ell \in [0,1]$, which gives the fraction of the nodes that are *not* in cycles of length K or less. It is for $N\to\infty$ apparently given by

$$m_\infty = \lim_{N\to\infty}\frac{1}{2N}\sum_{\ell=K+1}^{N}\mathrm{e}^{-\mathrm{i}\omega_N\ell}. \tag{16}$$

It follows from (15), that the physical saddle point ω, after contour deformation, must be purely imaginary. We switch accordingly to the new variable $x = \mathrm{e}^{-\mathrm{i}\omega} \in \mathbb{R}_0^+$, in terms of which our equations become:

$$1 = \frac{1}{2}\sum_{\ell=3}^{K} x_N^\ell \mathrm{e}^{\ell\tilde{\alpha}_\ell} + \frac{1}{2N}\sum_{\ell=K+1}^{N} x_N^\ell, \tag{17}$$

$$m_\ell = \lim_{N\to\infty} \frac{1}{2} x_N^\ell \mathrm{e}^{\ell\tilde{\alpha}_\ell}, \tag{18}$$

$$m_\infty = \lim_{N\to\infty} \frac{1}{2N}\sum_{\ell=K+1}^{N} x_N^\ell. \tag{19}$$

3.3. Phase phenomenology of the ensemble

We will now demonstrate that the solutions to the coupled equations (17,18) give rise to two phases of our graph ensemble. A *disconnected* phase is characterized by the fact that all nodes are typically in cycles of length K or less, so $m_\infty = 0$. A second phase, the *connected* phase, is characterized by finding a finite fraction of the nodes in longer cycles, so here $m_\infty > 0$. The transition separating the phases is marked by bifurcation of $m_\infty > 0$ solutions.

If $\lim_{N\to\infty} x_N = x < 1$, the second term of (17) vanishes for $N \to \infty$, and we immediately obtain $m_\infty = 0$. Hence we are in the disconnected phase, and here the asymptotic observables m_ℓ are simply found by solving

$$m_\infty = 0: \quad 1 = \frac{1}{2}\sum_{\ell=3}^{K} x^\ell \mathrm{e}^{\ell\tilde{\alpha}_\ell}, \quad m_\ell = \frac{1}{2} x^\ell \mathrm{e}^{\ell\tilde{\alpha}_\ell}. \tag{20}$$

The condition $x < 1$ for this solution to exist will be met for large values of $\{\tilde{\alpha}_\ell\}$. Upon reducing the control parameters $\{\tilde{\alpha}_\ell\}$, the value of x will increase, and a transition to the connected phase occurs exactly when $x = 1$. This happens at the critical manifold in the $K-2$ dimensional parameter space, defined by validity of

$$\sum_{\ell=3}^{K} \mathrm{e}^{\ell\tilde{\alpha}_\ell} = 2. \tag{21}$$

To confirm the equations of the connected phase, we need to investigate how the solution x_N of (17) scales with N as we approach $x = 1$. Substituting $x_N = 1 - \xi/N$, expanding (17) in N, and taking the limit $N \to \infty$ gives

$$m_\ell = \frac{1}{2}\mathrm{e}^{\ell\tilde{\alpha}_\ell}, \qquad m_\infty = 1 - \frac{1}{2}\sum_{\ell=3}^{K} \mathrm{e}^{\ell\tilde{\alpha}_\ell}, \tag{22}$$

and the link between ξ and m_∞ is $m_\infty = (1-\mathrm{e}^{-\xi})/2\xi$.

It turns out that all cycles of finite length $L > K$ will always have vanishing density for $N \to \infty$. This can be seen simply by replacing $K \to L$ in the previous analysis, but with $\alpha_\ell = 0$ for all $K < \ell \leq L$. The newly added control parameters with $\ell > K$ will give $\tilde{\alpha}_\ell = -\ell^{-1}\log N$, and hence $m_\ell = \lim_{N\to\infty} \frac{1}{2} x^\ell \mathrm{e}^{\ell\tilde{\alpha}_\ell} = \lim_{N\to\infty} \frac{1}{2} x^\ell/N = 0$, in both phases. We knew that in the disconnected phase all nodes will typically be in the controlled short cycles of length K or less. We may now conclude that, in the connected phase, those nodes that are not in the controlled short cycles (the fraction $m_\infty > 0$) will typically be found in cycles of *diverging* length.

The ensemble's Shannon entropy [28] is given by

$$\begin{aligned} S_N &= -\sum_{\mathbf{A}\in\mathcal{G}_N} p(\mathbf{A})\log p(\mathbf{A}) \\ &= \log N! + N\Big[\phi_N(\boldsymbol{\alpha}) - \sum_{\ell=3}^{K}\alpha_\ell\frac{\partial\phi_N(\boldsymbol{\alpha})}{\partial\alpha_\ell}\Big] \\ &= N\log(N)\Big(1-\sum_{\ell=3}^{K}\frac{m_\ell}{\ell}\Big)+\mathcal{O}(N). \end{aligned} \tag{23}$$

Since $\sum_{\ell=3}^{K}(m_\ell/\ell)\leq\frac{1}{3}\sum_{\ell=3}^{K}m_\ell\leq\frac{1}{3}$, the leading order will for large N always scale as $N\log(N)$, and be bounded according to $S_N\geq\frac{2}{3}N\log(N)+\mathcal{O}(N)$, but with a reduced prefactor if we increase the fraction of nodes in short cycles. The lower bound is achieved in the disconnected phase, when $m_3=1$ and $m_{\ell>3}=0$.

3.4. Spectral densities of adjacency matrices

A graph can be represented uniquely by its adjacency matrix $\{A_{ij}\}$, where $A_{ij}\in\{0,1\}$, and $A_{ij}=1$ if and only if there is a link from j to i. The set $\mathcal{G}_N$ contains only simple nondirected graphs, so our adjacency matrices are symmetric and with zero diagonal elements. The eigenvalue density of the adjacency matrix of a graph $\mathbf{A}$,

$$\varrho(\mu|\mathbf{A})=\frac{1}{N}\sum_{i=1}^{N}\delta[\mu-\mu_i(\mathbf{A})], \tag{24}$$

contains valuable information on the statistics of cycles in the graph. Here the sum runs over the set of (real) eigenvalues $\{\mu_i(\mathbf{A})\}_{i=1,\ldots,N}$ of $\mathbf{A}$, taking into account multiplicities. For instance, the number of closed paths in $\mathbf{A}$ is proportional to $\int\mathrm{d}\mu\ \varrho(\mu|\mathbf{A})\mu^\ell$. Our main quantity of interest will be the expected density, averaged over the ensemble probabilities (1), in the infinite size limit,

$$\varrho(\mu)=\lim_{N\to\infty}\sum_{\mathbf{A}\in\mathcal{G}_N}p(\mathbf{A})\varrho(\mu|\mathbf{A}). \tag{25}$$

The adjacency matrix of a graph that consists of a single cycle of length ℓ has the Toeplitz form, and is therefore diagonalized trivially, leading to the density

$$\varrho_\ell(\mu)=\frac{1}{\ell}\sum_{r=0}^{\ell-1}\delta\Big(\mu-2\cos(2\pi r/\ell)\Big). \tag{26}$$

If the cycle length ℓ diverges, this density becomes continuous (in a distributional sense), see e.g. [29],

$$\varrho_\infty(\mu)=\lim_{\ell\to\infty}\varrho_\ell(\mu)=\frac{1}{\pi}\frac{\theta(2-|\mu|)}{\sqrt{4-\mu^2}}. \tag{27}$$

The set of eigenvalues for each $\mathbf{A}\in\mathcal{G}_N$ will just be the union of all the sets of eigenvalues of the disjoint cycles of which it is composed, taking multiplicities into account:

$$\begin{aligned} \varrho(\mu|\mathbf{A}) &= \frac{1}{N}\sum_{\ell=3}^{N}n_\ell(\mathbf{A})\sum_{r=0}^{\ell-1}\delta\Big(\mu-2\cos\Big(\frac{2\pi r}{\ell}\Big)\Big) \\ &= \sum_{\ell=3}^{N}\frac{\ell n_\ell(\mathbf{A})}{N}\varrho_\ell(\mu). \end{aligned} \tag{28}$$

Upon averaging over the ensemble, using (3) and our earlier observation that for $N \to \infty$ the fraction of nodes in cycles of finite length $L > K$ vanishes, we immediately obtain the asymptotic ensemble-averaged spectrum corresponding to (1), expressed in terms of (26) and (27):

$$\varrho(\mu) = \sum_{\ell=3}^{K} m_\ell \varrho_\ell(\mu) + m_\infty \varrho_\infty(\mu). \tag{29}$$

Since we are working with regular graphs, we can immediately recover the spectrum of the Laplacian operator ($\mathbf{L} = 2\mathbf{I} - \mathbf{A}$) by the change of variable $\mu \to 2 - \lambda$.

4. Grand Canonical approach

Within the canonical approach one finds that, if N is sufficiently large, graphs generated randomly from (1) will all display the same values of the main intensive quantities, such as the fraction of ℓ-cycles (modulo finite size fluctuations). We expect a similar claim to hold if we sample randomly both the graphs and the number N of nodes, i.e. if we work with grand canonical graph ensembles. The grand partition function of our ensemble with weights $w_N = \mathrm{e}^{-\mu N}/N!$ (where $\mu > 0$) is given by

$$Q(\boldsymbol{\alpha}) = \sum_{N=1}^{\infty} w_N Z_N(\boldsymbol{\alpha}), \tag{30}$$

with $Z_N(\boldsymbol{\alpha})$ defined in (2). The divisor $N!$ in w_N will simplify our calculation, without losing the benefits of the thermodynamic limit (since we will find that for $\mu \to 0$ the expected system size still diverges). Direct calculation of $Q(\boldsymbol{\alpha})$ now circumvents the integration over ω:

$$\begin{aligned}
Q(\boldsymbol{\alpha}) &= \sum_{\boldsymbol{n}} \Big(\prod_{\ell=3}^{\infty} \frac{\mathrm{e}^{-\mu \ell n_\ell}}{(2\ell)^{n_\ell} n_\ell!} \Big) \left(\prod_{\ell=3}^{K} \mathrm{e}^{\ell \alpha_\ell n_\ell} \right) \\
&= \left[\prod_{\ell=3}^{K} \left(\sum_{n \geq 0} \frac{\mathrm{e}^{(\alpha_\ell - \mu)\ell n}}{(2\ell)^n n!} \right) \right] \left[\prod_{\ell > K} \left(\sum_{n \geq 0} \frac{\mathrm{e}^{-\mu \ell n}}{(2\ell)^n n!} \right) \right] \\
&= \exp\left(\sum_{\ell=3}^{K} \frac{\mathrm{e}^{(\alpha_\ell - \mu)\ell}}{2\ell} + \sum_{\ell > K} \frac{\mathrm{e}^{-\mu \ell}}{2\ell} \right) \\
&= \exp\left(\sum_{\ell=3}^{K} \frac{\mathrm{e}^{(\alpha_\ell - \mu)\ell}}{2\ell} - \frac{1}{2}\log(1 - \mathrm{e}^{-\mu}) - \sum_{\ell=1}^{K} \frac{\mathrm{e}^{-\mu \ell}}{2\ell} \right),
\end{aligned} \tag{31}$$

where we used $\sum_{\ell > 0} x^\ell/\ell = -\log(1-x)$. From $Q(\boldsymbol{\alpha})$ we obtain, in turn, the grand potential $\Omega(\boldsymbol{\alpha}) = -\log Q(\boldsymbol{\alpha})$:

$$\Omega(\boldsymbol{\alpha}) = \sum_{\ell=1}^{K} \frac{\mathrm{e}^{-\mu \ell}}{2\ell} - \sum_{\ell=3}^{K} \frac{\mathrm{e}^{(\alpha_\ell - \mu)\ell}}{2\ell} + \frac{1}{2}\log(1 - \mathrm{e}^{-\mu}). \tag{32}$$

Its partial derivatives with respect to μ and $\boldsymbol{\alpha}$ yield the average system size, via $\langle N \rangle = \partial \Omega(\boldsymbol{\alpha})/\partial \mu$, and the average number of length-ℓ cycles (for $\ell = 3, \ldots, K$), via

$\langle n_\ell(\mathbf{A})\rangle = -\ell^{-1}\partial\Omega(\boldsymbol{\alpha})/\partial\alpha_\ell$. We thereby find that

$$\begin{aligned}\langle N\rangle &= \frac{1}{2}\frac{\mathrm{e}^{-\mu}}{1-\mathrm{e}^{-\mu}} + \frac{1}{2}\sum_{\ell=3}^{K}\mathrm{e}^{(\alpha_\ell-\mu)\ell} - \frac{1}{2}\sum_{\ell=1}^{K}\mathrm{e}^{-\mu\ell} \\ &= \frac{1}{2}\frac{\mathrm{e}^{-\mu(K+1)}}{1-\mathrm{e}^{-\mu}} + \frac{1}{2}\sum_{\ell=3}^{K}\mathrm{e}^{(\alpha_\ell-\mu)\ell} \end{aligned} \tag{33}$$

and

$$\langle n_\ell(\mathbf{A})\rangle = \frac{1}{2\ell}\mathrm{e}^{(\alpha_\ell-\mu)\ell}. \tag{34}$$

Clearly, $\langle N\rangle$ diverges for $\mu\to 0$, which gives our thermodynamic limit. In this limit we can then work out for $\ell\in\{3,\ldots,K\}$ the ratios

$$\lim_{\mu\downarrow 0}\frac{\ell\langle n_\ell\rangle}{\langle N\rangle} = \lim_{\mu\downarrow 0}\frac{\mathrm{e}^{(\alpha_\ell-\mu)\ell}}{\frac{\mathrm{e}^{-\mu(K+1)}}{1-\mathrm{e}^{-\mu}} + \sum_{\ell'=3}^{K}\mathrm{e}^{(\alpha_{\ell'}-\mu)\ell'}} = 0. \tag{35}$$

Similar to the canonical case, any μ-independent $\boldsymbol{\alpha}$ will asymptotically always yield a vanishing fraction of nodes in cycles of length $\ell\leq K$. It is clear from (34) that without a re-parametrization, the expected value of ℓ-cycles only increases exponentially with α_ℓ in the thermodynamic limit. We need to re-parametrize in such a way that the expected number of ℓ-cycles increases as the expected system size increases. The re-parametrization required to obtain a non-trivial thermodynamic limit is $\alpha_\ell = \tilde{\alpha}_\ell + \ell^{-1}\log\langle N\rangle$. Upon following this prescription, we then reproduce the canonical result

$$\frac{\ell\langle n_\ell\rangle}{\langle N\rangle} = \frac{1}{2}\mathrm{e}^{(\tilde{\alpha}_\ell-\mu)\ell}, \tag{36}$$

and our expression (33) for $\langle N\rangle$ now becomes

$$\langle N\rangle = \frac{\mathrm{e}^{-\mu(K+1)}}{(1-\mathrm{e}^{-\mu})\big(2-\sum_{\ell=3}^{K}\mathrm{e}^{(\tilde{\alpha}_\ell-\mu)\ell}\big)}. \tag{37}$$

The re-parametrization of $\boldsymbol{\alpha}$ now depends on $\boldsymbol{\alpha}$ itself, via $\langle N\rangle$, and has to be consistent with a nonnegative value for (37), i.e. with $\frac{1}{2}\sum_{\ell=3}^{K}\mathrm{e}^{(\tilde{\alpha}_\ell-\mu)\ell}\leq 1$. Expression (36) gives us the physical interpretation $\sum_{\ell=3}^{K}\ell\langle n_\ell\rangle/\langle N\rangle\leq 1$. In the limit $\mu\downarrow 0$ the condition becomes $\sum_{\ell=3}^{K}\mathrm{e}^{\ell\tilde{\alpha}_\ell}\leq 2$. In the case of inequality we have $\sum_{\ell=3}^{K}\ell\langle n_\ell\rangle/\langle N\rangle<1$, so we are in the connected phase. The case of equality reproduces our earlier phase transition condition (21) and we enter the disconnected phase; here the thermodynamic limit is reached already for nonzero μ, and we can again recover our canonical equations, with $\exp(-\mu)$ now playing the role of the canonical order parameter x.

5. Numerical simulations

Calculating $Z_N(\boldsymbol{\alpha})$ by numerical enumeration for nontrivial values of N is not a realistic option, since the size of the set $\mathcal{G}_N$ grows super-exponentially with N. Instead, to test our theoretical predictions we have sampled graphs from the ensemble (1) using the Markov Chain Monte Carlo (MCMC) method described in e.g. [30] or [6]. Starting from an arbitrary 2-regular N-node graph, this stochastic process is based on executing repeated (degree-preserving) edge swap moves with appropriate nontrivial move acceptance probabilities, constructed such that the Markov chain's equilibrium

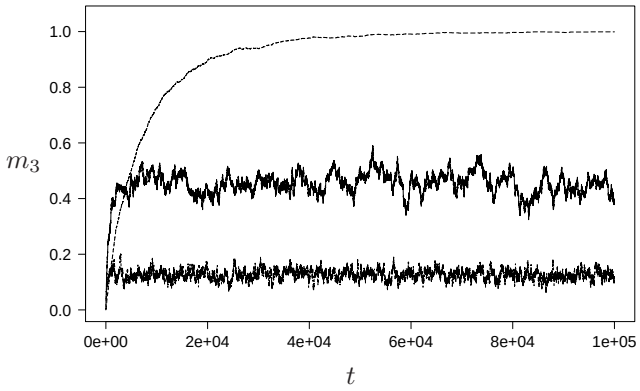

Figure 1. Examples of the evolution of the fraction m_3 of nodes in triangles, measured during MCMC simulations, for $K = 3$. Time is defined as the number of accepted edge swap moves per link. The bottom two curves correspond to the connected phase of the ensemble, equilibrating to the values $m_3 = 0.125$ for $\tilde{\alpha}_3 = \frac{1}{3}\log(0.25)$, and to $m_3 = 0.45$ for $\tilde{\alpha}_3 = \frac{1}{3}\log(0.9)$. The top curve corresponds to the disconnected phase, here the MCMC process is equilibrating to the value $m_3 = 1$.

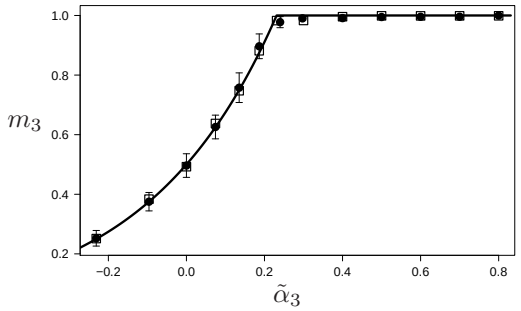

Figure 2. Values of m_3 shown versus $\tilde{\alpha}_3$ for ensembles with $K = 3$. Numerical results, measured upon equilibration of the MCMC processes, are shown as black dots with error bars for $N = 1000$, and as squares for $N = 5000$ (error bars for $N = 5000$ are not shown; their sizes are similar to or smaller than the squares). The solid line is the prediction of (38).

distribution is the target measure (1). In each simulation experiment, the MCMC process was first run for 10^5 to 10^6 accepted moves per link, and equilibration was confirmed by measuring the Hamming distance between the instantaneous and the initial state. After this randomization stage, the instantaneous state $\mathbf{A}$ arrived at by the chain was defined to be our graph sample. We have limited our simulations to ensembles with $K = 3$ and $K = 4$. The degree of equilibration achieved by the MCMC during a run of 10^5 accepted moves per link is illustrated in Figure 1, where we show typical evolution curves of the order parameter m_3 during the stochastic process.

For $K = 3$ we have just one control parameter $\tilde{\alpha}_3$, and the order parameter is the fraction m_3 of nodes in triangles. The theory claims that, for large N, the graphs from our ensemble will be collections of triangles and large rings. The key equations

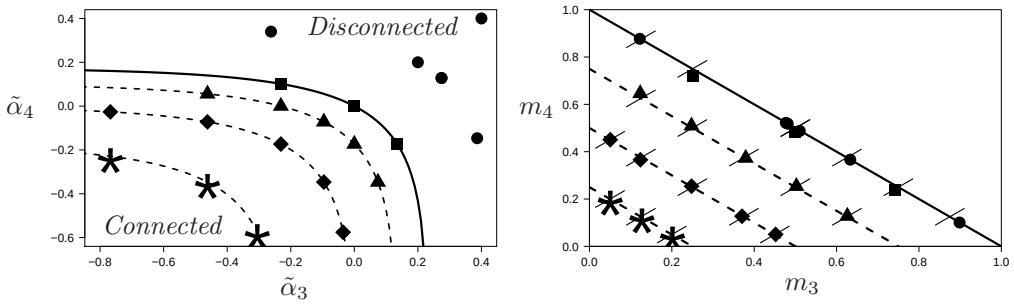

Figure 3. Left panel: the plane of control parameters for $K = 4$. The solid black line is the critical line $e^{3\tilde{\alpha}_3} + e^{4\tilde{\alpha}_4} = 2$ (here $m_\infty = 0$). The dashed lines correspond to parameter combinations with constant m_∞, taking the values $m_\infty \in \{0.75, 0.5, 0.25\}$, from bottom to top. The markers represent parameter combinations chosen for MCMC simulations. Right panel: the fractions (m_3, m_4) associated with the control parameter combinations in the left panel. Here the markers represent the simulation results, measured after execution of 10^4 accepted moves per node in the MCMC to secure equilibration. The results are indeed found on the respective lines predicted by the theory. Note that the theory predicts that all parameter combinations in the disconnected phase $e^{3\tilde{\alpha}_3} + e^{4\tilde{\alpha}_4} \geq 2$, should be mapped to the line $m_1 + m_2 = 1$ in the right panel. Error bars were omitted, as they are as big as or smaller than the markers.

(20,21,22) reduce to the following predictions, with $\tilde{\alpha}_c = \frac{1}{3}\log(2) \approx 0.23105\ldots$:

$$\begin{aligned} &\tilde{\alpha}_3 < \tilde{\alpha}_c: \quad m_3 = \tfrac{1}{2}e^{3\tilde{\alpha}_3}, \quad \textit{connected phase}, \\ &\tilde{\alpha}_3 > \tilde{\alpha}_c: \quad m_3 = 1, \quad \textit{disconnected phase}. \end{aligned} \tag{38}$$

Numerical simulations with sizes $N = 1000$ and $N = 5000$ show excellent agreement with these predictions, as shown in Figure 2, both in terms of the values of m_3 and in terms of the location of the transition.

For $K = 4$ we have two control parameters, $\tilde{\alpha}_3$ and $\tilde{\alpha}_4$, and the theory claims that for large N the graphs from our ensemble will now be collections of triangles, squares and large rings. Here the key equations (20,21,22) predict that the transition line in parameter space is given by $e^{3\tilde{\alpha}_3} + e^{4\tilde{\alpha}_4} = 2$, and that the fractions m_3 and m_4 of nodes found in triangles and squares, respectively, are solved (together with the auxiliary order parameter x, in the disconnected phase) from:

$$\begin{aligned} e^{3\tilde{\alpha}_3} + e^{4\tilde{\alpha}_4} < 2: \quad & m_3 + m_4 < 1, \quad \textit{connected phase}, \\ & m_3 = \tfrac{1}{2}e^{3\tilde{\alpha}_3}, \quad m_4 = \tfrac{1}{2}e^{4\tilde{\alpha}_4}, \\ e^{3\tilde{\alpha}_3} + e^{4\tilde{\alpha}_4} > 2: \quad & m_3 + m_4 = 1, \quad \textit{disconnected phase}, \\ & m_3 = \tfrac{1}{2}x^3 e^{3\tilde{\alpha}_3}, \quad m_4 = \tfrac{1}{2}x^4 e^{4\tilde{\alpha}_4}. \end{aligned} \tag{39}$$

Figure 3 (left panel) shows the resulting predicted phase diagram in the $(\tilde{\alpha}_3, \tilde{\alpha}_4)$ plane. The mapping $(\tilde{\alpha}_3, \tilde{\alpha}_4) \mapsto (m_3, m_4)$ will map the lower region of the phase diagram (the connected phase) to the interior of the triangle $m_3 + m_4 < 1$ in the right panel of Figure 3. The upper region of the phase diagram on the left (the disconnected phase), including the critical line, will be mapped to the line $m_3 + m_4 = 1$ in the right panel. To test also these predictions against numerical simulations, we have chosen multiple points $(\tilde{\alpha}_3, \tilde{\alpha}_4)$ in both regions of the phase diagram, grouped such that the

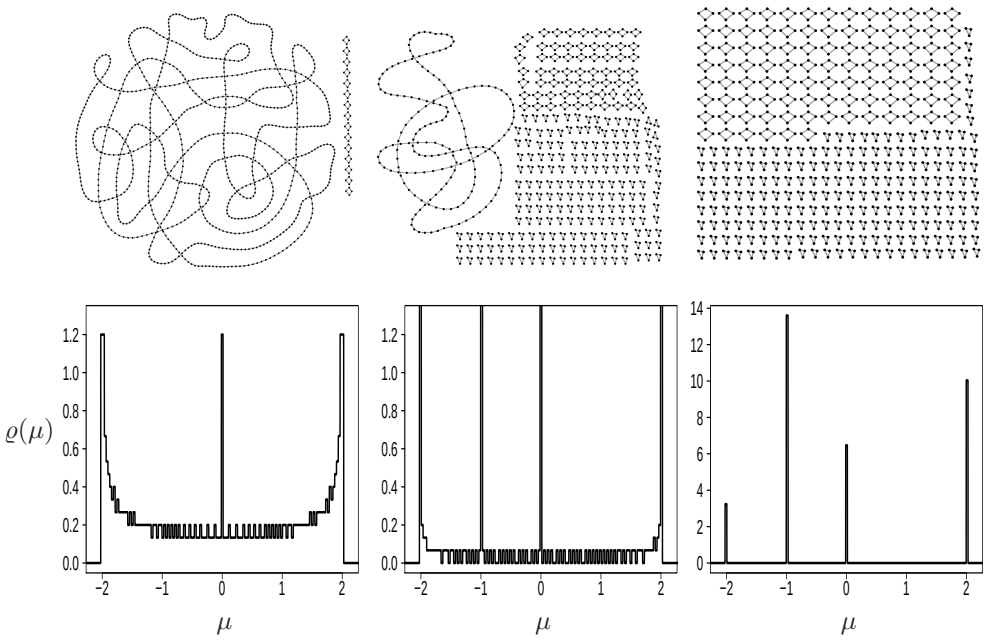

Figure 4. Top row: typical graphs sampled numerically via MCMC from the canonical ensemble (1) of 2-regular nondirected simple graphs, for $N = 1000$. Left: $(m_3, m_4) = (0.0, 0.06)$ and $m_\infty = 0.94$. Middle: $(m_3, m_4) = (0.25, 0.56)$ and $m_\infty = 0.19$. Right: $(m_3, m_4) = (0.39, 0.61)$ and $m_\infty = 0$. The bottom row shows the eigenvalue spectra of the corresponding three adjacency matrices, computed by direct numerical diagonalization. The locations of the peaks are seen to agree with the theoretical predictions of (29). Note the different scale in the third spectrum graph, to emphasize the weights of the δ-peaks.

predicted values of $m_\infty = 1 - m_3 - m_4$ were always in the set $\{0.25, 0.5, 0.75\}$. The prediction would therefore be that in the (m_3, m_4) plane these groups of points should be found on the lines $m_3 + m_4 = 1 - m_\infty$. Upon measuring the fractions m_3 and m_4 via MCMC in the corresponding graph ensembles, these predictions are once more validated convincingly. See Figure 3.

Graphs sampled from our ensemble with $K = 4$ do indeed typically consist of controlled numbers of triangles and squares, and a long ring. Figure 4 shows examples of such graphs, obtained via MCMC, together with the eigenvalue spectra of their adjacency matrices (obtained by numerical diagonalization). Also the observed spectra agree with the corresponding theoretical predictions (29).

6. Discussion

In this paper we presented an analytical solution for an exponential random graph ensemble with a controllable density of short cycles. Whereas one would normally not expect such non-treelike graph ensembles to be solvable, here this is possible as a consequence of imposing a local degree constraint of strict 2-regularity. We found a second order phase transition, which separates a connected phase with large and small cycles from a disconnected phase where the graphs are typically formed only of

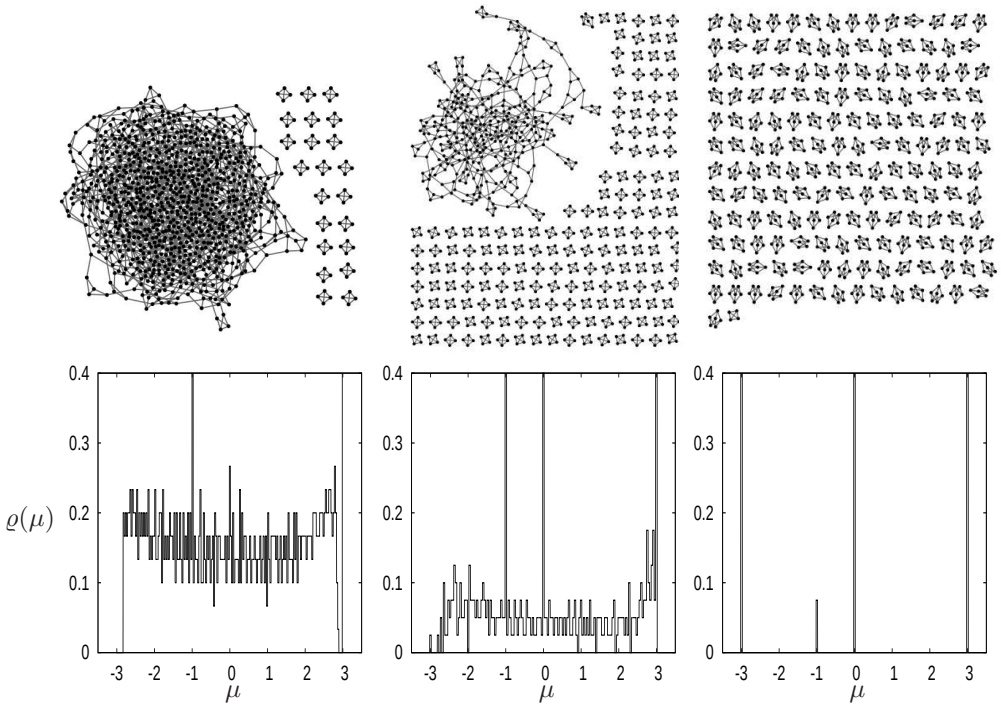

Figure 5. Top row: typical graphs sampled numerically via MCMC from the canonical ensemble (1) of 3-regular nondirected simple graphs, for $N = 1000$. Left: $(m_3, m_4) = (0.63, 0.65)$. Middle: $(m_3, m_4) = (3.34, 3.34)$. Right: $(m_3, m_4) = (0.02, 11.97)$. The bottom row shows the eigenvalue spectra of the corresponding three adjacency matrices, computed by direct numerical diagonalization. The locations of the peaks are seen to agree with the spectrum of the small subgraphs.

extensively many short cycles. The short cycles appear in controlled proportions, for which we found analytical expressions in terms of the ensemble's parameters. We also derived an analytical expression for the critical submanifold in the phase diagram, and for the expected eigenvalue spectrum of the graphs' adjacency matrices.

We analysed both the canonical and the grand canonical formulation of the ensemble. The canonical version was solved via steepest descent integration. In the grand canonical version one avoids steepest descent integration, but (as always) the chemical potential takes over the role of the steepest descent integration variable of the canonical version. In the thermodynamic limit, the canonical and grand canonical routes result in identical equations. These equations are found to give highly accurate predictions already for modest graph sizes, such as $N = 1000$, as we confirmed in numerical simulations.

The parameter K represents the largest cycle length that is controlled in our model. For $K = 3$ one controls only the number of triangles, and our ensemble becomes similar to that of Strauss [8, 11, 23] with average degree two. The remaining difference is that in the Strauss model the average degree is imposed implicitly via an overall 'soft' constraint, while in the present model all degree values are imposed as local 'hard' constraints. Due to this difference, the degeneration of the Strauss model to a phase where the complete clique has probability one (so the number of

triangles can no longer be tuned) is avoided in the present ensemble. The complete clique is simply no longer an allowed configuration, and hence the number of triangles becomes fully tuneable, if the model parameters scale appropriately with the system size. In addition, in [25] it is shown that the 'soft' version of our model would have a phase diagram reminiscent of ours. In both cases the sign of a linear combination of functions of the parameters determines the phase of the ensemble. However, in the 'soft' case of [25] there is a transition from an almost ER-like phase to a clique, while our model exhibits tuneability of the densities in both phases.

As we mentioned before, the generalization of the present model to other degree distributions has been studied for the Poissonian and the q-regular cases. Numerical explorations for 3-regular versions with $K = 4$ have been reported in [6], and show phenomenology similar to that found here for $q = 2$. In particular, one again observes a disconnected phase for large values of α_3 and α_4. One could have thought that the phenomenology of our model, in particular the emergence of a large number of small clusters, is specific to the simplifications induced by the 2-regularity condition, but this is not the case. To emphasize this fact, we redid simulations in the same fashion as in Figure 4, but now for 3-regular as opposed to 2-regular graphs, where analytical solution along the lines followed for $q = 2$ is no longer feasible. The results are shown in Figure 5. It is clear that as the bias towards increased numbers of triangles and/or squares is increased, the graph breaks down into small regular graphlets that maximize the cycle density per node. For Poissonian graphs, simulations revealed in [26] that, upon boosting triangles, they also break down into small graphlets, similar to the present model. Also the effect of boosting triangles on the spectrum of the adjacency matrix was similar to what we observed here. However, since the small graphlets that appear in Poissonian models are in general different from isolated triangles and from each other, the associated eigenvalues are described by a different distribution. Similarly, the spectrum of the large component changes in a more complicated way than just by scaling down. Yet, overall we find a similar phenomenology. In fact, our present analysis predicts that the parameters in [26] would need a scaling with N, in order for the transition not to be a finite size effect but to persist in the $N \to \infty$ limit.

We could also combine our present model with the Erdös-Rènyi ensemble, to produce connected random graphs with a varying number of short cycles. Again, while the phenomenology of such variations could be explored via simulations, it is not clear how one would be able to obtain analytical solutions without the benefit of locally tree-like topology.

In our view, the main merit of the present model is that its analytical solution helps us understand more complicated 'loopy' graph ensembles. We are aware that the analytical route taken in this case is surely impossible for other models. Nevertheless, it provides an explicit analytical solution that reproduces the main features of non-treelike random graph ensembles with hard degree constraints. It helps us understand phenomenology that had so far only been studied numerically. It can also serve as a benchmark model against which more general solution strategies for non-treelike random graphs can be tested, such as [27], which deals with spectrally constrained maximum entropy graph ensembles. The moments of a graph's spectral density are related to its numbers of cycles, via the traces of powers of the adjacency matrix. In fact, the present model is a special case of the family of ensembles studied in [27], from which it can be obtained by choosing 2-regular degrees and an appropriate polynomial functional Lagrange parameter. The analytical and numerical results of this paper suggest that, to obtain phase transitions, the functional Lagrange parameters in

spectrally constrained maximum entropy graph ensembles [27] may need to have a specific scaling with the system size.

Acknowledgement

FAL gratefully acknowledges financial support through a scholarship from Conacyt (Mexico).

References

[1] Békéssy A, Bekessy P and Komlós J 1972 *Stud. Sci. Math. Hungar* **7** 343–353
[2] Bender E A and Canfield E R 1978 *Journal of Combinatorial Theory, Series A* **24** 296–307
[3] Molloy M and Reed B 1995 *Random structures & algorithms* **6** 161–180
[4] Newman M E, Strogatz S H and Watts D J 2001 *Physical review E* **64** 026118
[5] Catanzaro M, Boguná M and Pastor-Satorras R 2005 *Physical Review E* **71** 027103
[6] Annibale A, Roberts E *et al.* 2017 *Generating Random Networks and Graphs* (Oxford University Press)
[7] Holland P W, Laskey K B and Leinhardt S 1983 *Social networks* **5** 109–137
[8] Jonasson J 1999 *Journal of Applied Probability* **36** 852–867
[9] Holme P and Kim B J 2002 *Physical review E* **65** 026107
[10] Davidsen J, Ebel H and Bornholdt S 2002 *Physical Review Letters* **88** 128701
[11] Burda Z, Jurkiewicz J and Krzywicki A 2004 *Physical Review E* **69** 026106
[12] Krapivsky P L and Redner S 2005 *Physical Review E* **71** 036118
[13] Newman M E 2009 *Physical review letters* **103** 058701
[14] Bollobás B, Janson S and Riordan O 2011 *Random Structures & Algorithms* **38** 269–323
[15] Bianconi G, Darst R K, Iacovacci J and Fortunato S 2014 *Physical Review E* **90** 042806
[16] Granovetter M S 1973 *American journal of sociology* **78** 1360–1380
[17] Solé R V, Pastor-Satorras R, Smith E and Kepler T B 2002 *Advances in Complex Systems* **5** 43–54
[18] Vázquez A 2003 *Physical Review E* **67** 056104
[19] Marsili M, Vega-Redondo F and Slanina F 2004 *Proceedings of the National Academy of Sciences of the United States of America* **101** 1439–1442
[20] Ispolatov I, Krapivsky P and Yuryev A 2005 *Physical review E* **71** 061911
[21] Toivonen R, Onnela J P, Saramäki J, Hyvönen J and Kaski K 2006 *Physica A: Statistical Mechanics and its Applications* **371** 851–860
[22] Jackson M O and Rogers B W 2007 *The American economic review* **97** 890–915
[23] Strauss D 1986 *SIAM review* **28** 513–527
[24] Chatterjee S, Diaconis P *et al.* 2013 *The Annals of Statistics* **41** 2428–2461
[25] Yin M 2016 *Journal of Statistical Physics* **164** 241–253
[26] Avetisov V, Hovhannisyan M, Gorsky A, Nechaev S, Tamm M and Valba O 2016 *Physical Review E* **94** 062313
[27] Coolen A 2016 Replica methods for loopy sparse random graphs *Journal of Physics: Conference Series* vol 699 (IOP Publishing) p 012022
[28] Cover T M and Thomas J A 2012 *Elements of information theory* (John Wiley & Sons)
[29] McKay B D 1981 *Linear Algebra and its Applications* **40** 203–216
[30] Coolen A, De Martino A and Annibale A 2009 *Journal of Statistical Physics* **136** 1035–1067