Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
26 August 2020

REVISED
23 March 2021

ACCEPTED FOR PUBLICATION
29 March 2021

PUBLISHED
21 May 2021

PAPER

Transitions in random graphs of fixed degrees with many short cycles

Fabián Aguirre López[1,*] and Anthony C C Coolen[2,3]

[1] Université Paris-Saclay, CNRS, LPTMS, 91405, Orsay, France
[2] Department of Biophysics, Donders Institute, Radboud University, 6525AJ Nijmegen, The Netherlands
[3] London Inst for Mathematical Sciences, Royal Institution, 21 Albemarle Street, London W1S 4BS, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: fabian.aguirre-lopez@universite-paris-saclay.fr and a.coolen@science.ru.nl

Keywords: network theory, statistical physics, graph theory, statistical models, combinatorics

Abstract

We analyze maximum entropy random graph ensembles with constrained degrees, drawn from arbitrary degree distributions, and a tuneable number of three-cycles (triangles). We find that such ensembles generally exhibit two transitions, a clustering and a shattering transition, separating three distinct regimes. At the clustering transition, the graphs change from typically having only isolated cycles to forming cycle clusters. At the shattering transition the graphs break up into many small cliques to achieve the desired three-cycle density. The locations of both transitions depend nontrivially on the system size. We derive a general formula for the three-cycle density in the regime of isolated cycles, for graphs with degree distributions that have finite first and second moments. For bounded degree distributions we present further analytical results on cycle densities and phase transition locations, which, while non-rigorous, are all validated via MCMC sampling simulations. We show that the shattering transition is of an entropic nature, occurring for all three-cycle density values, provided the system is large enough.

1. Introduction

Graph theory was introduced by Euler to solve the problem of the seven bridges of Königsberg [1]. He noted that upon stripping all unnecessary details, to solve this problem, one was left only with a set of 4 nodes and 7 links between them. Since then, networks and graphs have proven to be fundamental in the modelling of many real world phenomena. While with the advent of powerful computers accessible to almost all researchers it is now typical for network scientists to work on a daily basis with networks of nodes ranging from thousands to millions, still the modelling strategy is the same: remove unnecessary details and reduce the problem to nodes and links.

For scientists, and especially those with a statistical training—used to thinking in terms of null models in hypothesis testing [2]—it is natural to ask a very simple question regarding observed networks: which are typical and which are atypical topological features? To answer this question one commonly works with random graph ensembles, designed to mimic real-world networks; see e.g. [3–6]. In addition to studying properties of graphs, one usually also seeks to understand processes for which these graphs define the interaction infrastructure, and the relation between graph topology and process efficacy. Here one would benefit from exact analytical solutions for processes defined on nontrivial graph ensembles. Unfortunately, this is hard. While there has been an explosion of exact solutions for processes on random graphs, the vast majority of these are locally tree-like graphs. The property of being locally tree-like allows one to write recursive equations, that become exact for large graphs and show very good agreement with simulations on finite ones. Ironically, this property that makes the models solvable is the same property that makes them unrealistic.

In addition to the lack of solvable models on graphs with many short cycles, there is also the fact that there is no easily controllable random graph ensemble that generates graphs with given numbers of links and short cycles in the sparse regime. The natural extension of the Erdös–Rényi ensemble [7] was presented by Strauss in

[8], which consisted in simply including an additional bias for the number of triangles in the model. Surprisingly, a very sharp transition was observed in the ensemble. When tuning the control parameters, the ensemble very quickly switched from typically sampling sparse graphs to assigning high probability to very dense graphs as well, losing any resemblance to real networks. There is a long history of attempts at understanding this transition [9–14], and many alternative random graph ensembles and algorithms have since then been presented [15–22] that possess a high number of short cycles, yet none generates easily controllable graphs. It appears very hard to access a regime where there is high number of triangles while keeping a 'nice' topology. More recent models conserve the degree sequence to avoid the condensation observed by Strauss, but still show a transition into a clustered regime [21–27]. The logical way of stopping the appearance of a clustered regime is to restrict the number of triangles in each node via a hard constraint, as proposed in [17, 18]. While there have been numerical and theoretical advances with this model [28–33], it remains difficult to keep the target degree distribution and the target total number of triangles under control [34].

In this paper we study a random graph ensemble with a tuneable number of short cycles, achieved with a soft global constraint on the number of triangles in combination with a hard constraint on the degrees, each drawn from a fixed degree distribution. This guarantees that our graphs will both be sparse and with a large number of short cycles, which are desired properties to mimic real networks. This model was previously studied in [19]. However, in that previous study the chosen MCMC move acceptance probabilities did not ensure convergence to the target distribution. This was pointed out in [5, 35], where it was shown that in edge swap graph dynamics nontrivial acceptance probabilities are needed (see also appendix A). We show in this paper that with the correct MCMC sampling the model again displays a transition into a clustered phase, and that the overall phenomenology presented in [19] coincides with our results. We then proceed to develop an extended theoretical and quantitative understanding of the behaviour of the model, including an analytic characterization of the low triangle density phase, expressions for the locations of the two (clustering and shattering) transitions, and scalar measures to probe the interactions between cycles and their relevance for the phases of the ensemble. Our results and predictions are supported via nontrivial graph sampling simulations involving different degree distributions, using the exact move acceptance probabilities of [5, 35].

2. The model

We study a random graph ensemble defined on the set of N-node graphs with a given degree distribution. A graph is an ordered pair (V, E) of nodes and edges, respectively. We model graphs through their adjacency matrices $\mathbf{A}$, defined by the entries $A_{ij} = 1$ if $(i, j) \in E$, and 0 otherwise. We will only be concerned with simple undirected graphs, which in terms of the adjacency matrix implies that $A_{ij} = A_{ji}$ and $A_{ii} = 0$ for all (i, j). The degree of a node is the number of edges connected to it, $k_i(\mathbf{A}) = \sum_j A_{ij}$. Throughout this paper we will work with graphs that have a prescribed degree sequence $\{k_i\}_{i=1,\dots,N}$. Each element of this sequence is drawn randomly and independently from a given distribution $p(k)$.

We expect that in the large N limit the empirical distribution of degrees will converge to the target distribution, $p(k) = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \delta_{k,k_i(\mathbf{A})}$. In general this property might not be true if the tails of the distribution are fat enough, but since we will focus our discussion on degree distributions with finite first and second moments this will not be a problem, as discussed in [36].

Our main interest will be to tune the number of triangles in the graphs. It will be more convenient to work with three-cycles than with triangles, for reasons that will become evident in further sections. The number of three-cycles in a graph corresponds to 6 times the number of triangles, as the former also account for starting point and directionality. There is a three-cycle around node i if there exist j and k such that $(i, j) \in E$, $(j, k) \in E$, and $(k, i) \in E$. Since our graph is simple, the indices i, j, k are all different. In the language of the adjacency matrix, the indicator function for a given three-cycle takes the simple form

$$\mathbb{I}\left[(i \to j \to k \to i) \in \mathbf{A}\right] = A_{ij}A_{jk}A_{ki}. \tag{1}$$

The total number of three-cycles is then simply the trace of the third power of the adjacency matrix,

$$\mathcal{M}(\mathbf{A}) = \sum_{ijk} \mathbb{I}\left[(i \to j \to k \to i) \in \mathbf{A}\right] = \mathrm{Tr}\left(\mathbf{A}^3\right). \tag{2}$$

We now define an ensemble of random graphs such that the average number of triangles can be controlled, using a parametrized distribution over graphs denoted by $p(\mathbf{A})$. Our choice is a maximum entropy (ME) ensemble. That is, we take $p(\mathbf{A})$ to be such that the average number of three-cycles is fixed,

$$\mathcal{M}^* = \sum_{\mathbf{A}} p(\mathbf{A})\, \mathrm{Tr}(\mathbf{A}^3), \tag{3}$$

and that the degree sequence $\mathbf{k} = \{k_i\}_{i=1,\dots,N}$ is achieved exactly. Among those distributions $p(\mathbf{A})$ that share these two properties, we choose the one that maximizes the Shannon entropy $S[p] = -\sum_{\mathbf{A}} p(\mathbf{A})\log p(\mathbf{A})$. This will guarantee that the distribution is statistically unbiased [37, 38]. The ME distribution is of an exponential form, with one tuneable parameter α,

$$p(\mathbf{A}) = \frac{1}{Z(\alpha)}\mathrm{e}^{\alpha\ \mathrm{Tr}(\mathbf{A}^3)}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}}. \tag{4}$$

The product over Kronecker deltas enforces the degree sequence of the graph. For $\alpha = 0$ the ensemble reduces to the configuration model (CM) [39], a uniform distribution over all graphs with degree sequence $\mathbf{k}$.

Our main observable of the ensemble will be the number of three-cycles *per node*. We will refer to it as the *three-cycle density*,

$$m(\alpha) = N^{-1}\left\langle \mathcal{M}(\mathbf{A})\right\rangle = N^{-1}\left\langle \mathrm{Tr}(\mathbf{A}^3)\right\rangle. \tag{5}$$

Where $\langle f(\mathbf{A})\rangle = \sum_{\mathbf{A}} p(\mathbf{A}) f(\mathbf{A})$. This quantity reflects the typical number of cycles in the neighborhood of a node. Each node can have a different maximum number of triangles, depending on its degree. Once a random graph ensemble like (4) is defined it is desirable to have both an algorithm to generate graph samples numerically and an analytic theory of its statistical properties. In order to generate samples form an ensemble such as (4) we use a Markov chain Monte Carlo (MCMC) approach. The algorithm starts with a seed graph satisfying the degree sequence, and evolves it by performing degree preserving edge swaps as shown in figure A1. Edge swaps are either accepted or rejected, with a nontrivial acceptance probability that not only takes into account the specific ensemble (4) but also the availability of possible edge swaps as the graph evolves. The theory of this MCMC algorithm was developed in [35] and presented more extensively in [5]. It is also summarized briefly in appendix A.

To find an analytic expression for the three-cycle density we need to calculate the generating function $\phi(\alpha)$:

$$\phi(\alpha) = \frac{1}{N}\log Z(\alpha) = \frac{1}{N}\log\sum_{\mathbf{A}}\mathrm{e}^{\alpha\,\mathrm{Tr}(\mathbf{A}^3)}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}} \tag{6}$$

$$m(\alpha) = \frac{\partial\phi(\alpha)}{\partial\alpha} = \left\langle \frac{1}{N}\mathrm{Tr}(\mathbf{A}^3)\right\rangle. \tag{7}$$

Ideally, knowledge of the functions $\phi(\alpha)$ and $m(\alpha)$ would allow us to generate random graphs with any desired three-cycle density. Although it is not possible to calculate $\phi(\alpha)$ analytically, in section 3 we will show that a small α approximation will give very good results for a wide range of values. Additionally we will give a description of the general behaviour of this ensemble for the whole range of α values. Another important observable of the ensemble reports on the amount of *interaction* between the triangles in the graph, i.e. the number of edges and nodes that different triangles share. This varies in a nontrivial way with different values of α and different system sizes N. To measure the degree of interaction between three-cycles, we define

$$r(\mathbf{A}) = \frac{\#\text{nodes in three}-\text{cycles}}{\#\text{of three}-\text{cycles}} = \frac{\sum_{i=1}^{N}\Theta[(\mathbf{A}^3)_{ii}]}{\mathrm{Tr}(\mathbf{A}^3)} \in [0, 1/2], \tag{8}$$

where $\Theta(x) = 1$ if $x > 0$ and zero otherwise. This ratio of nodes in three-cycles to three-cycles is independent of the total number of three-cycles in the graph. If $r(\mathbf{A}) = 1/2$, the three-cycles are all non-interacting in the sense that they do not share any nodes. If $r(\mathbf{A}) < 1/2$, then three-cycles are sharing nodes. Motivated by this intuition, we define $r(\mathbf{A}) = 1/2$ in the case of $\mathrm{Tr}(\mathbf{A}^3) = 0$, since no three-cycles implies there are no three-cycles interacting. Some simple examples are shown on the top row of figure 5. In the particular case where graphs form cliques of $q + 1$ nodes, we would have $r(\mathbf{A}) = 1/(q^2 - q)$; this is a natural lower bound for graphs of maximum degree q.

2.1. Main results

We will now outline the main results of our analysis of the ensemble (4). We found the same initial behaviour for all degree distributions as the triangle-inducing control parameter α is increased form $\alpha = 0$. This behaviour depends only on the first two moments of the degree distribution, $c = \overline{k}$ and $\overline{k^2}$ (where $\overline{f(k)} = N^{-1}\sum_{i=1}^{N} f(k_i)$), and on the maximum degree $q = \max_{i=1,\dots,N}\{k_i\}$ (for bounded degree distribution). We will only consider the case where N is sufficiently large, $Np(q) \gg q + 1$, so that all degrees are typically represented in the graph with an extensive number of nodes.

In figure 1 we show the results of numerical sampling of graphs from (4), using an appropriate MCMC process (simulation details are given in sections 3 and 4). The triangle density $m(\alpha)$ increases with α, as expected.

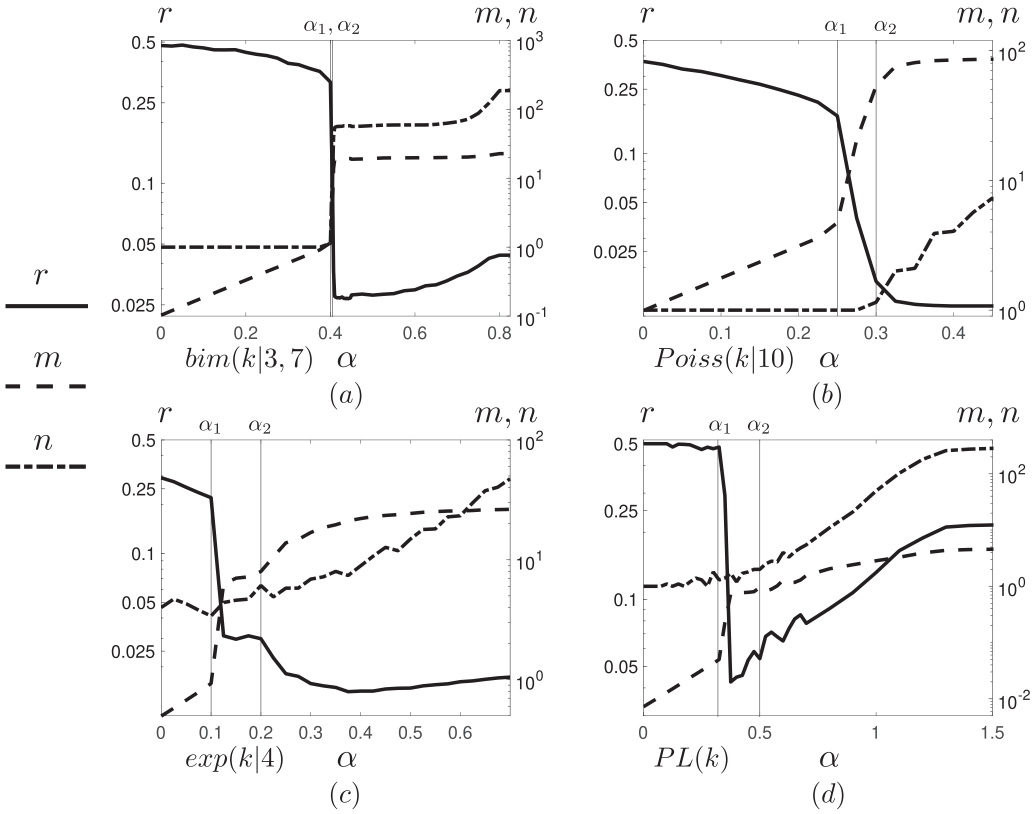

Figure 1. Values of the three main topological observables as measured in numerical sampling simulations of the random graph ensemble (4), all with $N = 1000$ and averaged over 20 samples. The cycle interaction observable $r(\alpha)$, defined in (8), is shown as solid lines with values on the left vertical axis. The triangle density $m(\alpha)$ is shown as dashed lines with values on the right vertical axis. The number of connected graph components $n(\alpha)$ is shown as dashed-dotted line with values on the right vertical axis. Different panels refer to different degree distributions: (a) $p(k) = \mathrm{bim}(k|3, 7)$; (b) $p(k) = \mathrm{Poiss}(k|10)$; (c) $p(k) = \exp(k|4)$; (d) $p(k) = \mathrm{PL}(k)$. See table 1 for the corresponding definitions. Error bars are omitted to avoid clutter; see figures 5 and 6 for examples of typical error bar values. Approximate positions for $\alpha_1(N)$ and $\alpha_2(N)$ are shown, see text for definitions.

We observe distinct regimes of α-values, as had already been observed for regular graphs in [27]. Interestingly, to understand properly the nature of the different regimes of the ensemble it is necessary to also look at two other graph observables: the level of interaction between cycles, measured with $r(\mathbf{A})$ as defined in (8), and the number $n(\mathbf{A})$ of connected components of the graph. We define their respective ensemble averages as $r(\alpha) = \langle r(\mathbf{A})\rangle$ and $n(\alpha) = \langle n(\mathbf{A})\rangle$. The observed regimes (see figure 1) are the following:

- $\alpha \in [0, \alpha_1(N)]$: *connected regime*

  The three-cycle density $m(\alpha)$ grows exponentially with α, following

  $$m(\alpha) \approx \frac{1}{N}\left(\overline{k^2}/c - 1\right)^3 \mathrm{e}^{6\alpha}. \tag{9}$$

  This formula is one of the main results of this letter, it is derived in 3. It corresponds to the straight lines observed before α_1 in figure 1. Only the proportionality constant and the transition point $\alpha_1(N)$ depend on the degree distribution. This formula allows for an explicit calculation of α, given a desired three-cycle density, simply by inversion. For $\alpha = 0$ it reproduces the rigorous result for the three-cycle density for large graphs in [39]. The degree of interaction between cycles is as low as $r(\mathbf{A}) \approx 1/2$ for large graphs. The number $n(\alpha)$ of components of the graph is the same as in the $\alpha = 0$ case. It is relatively easy to obtain samples in this regime with the MCMC edge swap dynamics.
- $\alpha \in [\alpha_1(N), \alpha_2(N)]$: *clustered regime*

  Here the triangle density $m(\alpha)$ grows faster than (9). Depending on the chosen degree distribution, this growth may exhibit sudden jumps or may be more smooth. The main difference with the previous regime is that cycles start sharing edges. This follows from the observed drop of $r(\alpha)$, clearly observed in all cases in figure 1. Nodes start to form clusters of similar degree. We call this the clustered regime, and $\alpha_1(N)$ the clustering transition point.
- $\alpha \in [\alpha_2(N), \infty)$: *disconnected regime*

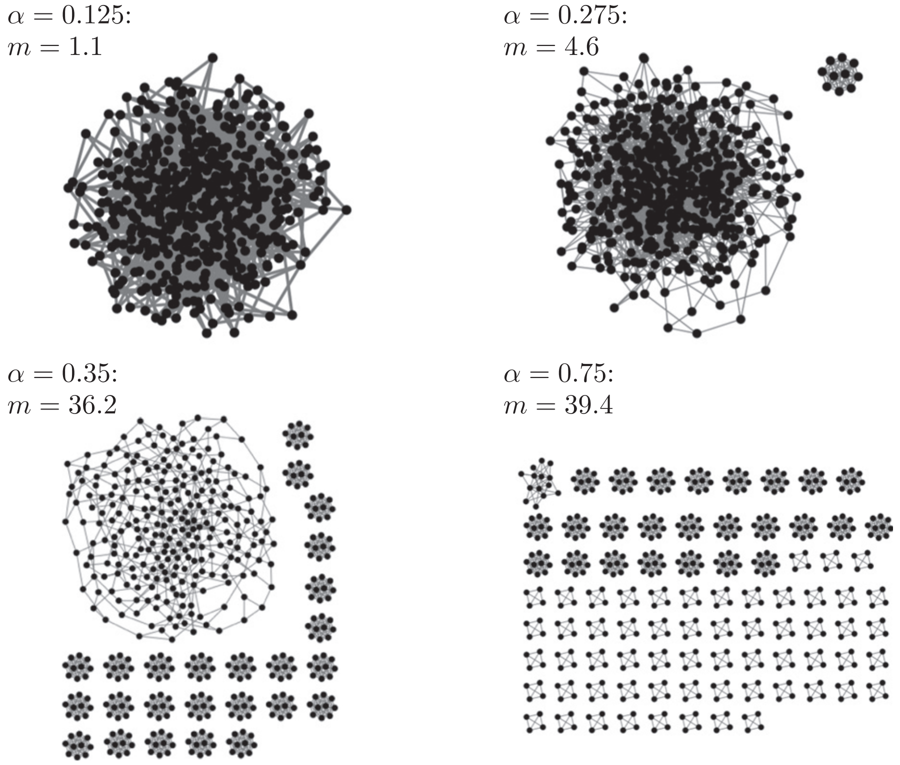

Figure 2. Results of numerical sampling simulations of the random graph ensemble (4), with $N = 500$ and $p(k) = \frac{1}{2}\delta_{k,3} + \frac{1}{2}\delta_{k,9}$. The four different images correspond to four different values of α, with different cycle densities $m(\alpha)$, as indicated.

There is a drastic topological change associated with a second transition at $\alpha_2(N)$: the graph breaks down into small disconnected cliques. Cliques of $k+1$ nodes maximize the number of cycles around a node of degree k, see figures 2 and 3. Cliques associated with the maximum degree, with $q+1$ nodes, will appear first, followed by those of the second largest degree, and so on. If, due to finite size effects, there are insufficient nodes to generate cliques, the graphs break down into small incomplete cliques. We call the transition at $\alpha_2(N)$ the *shattering* transition, and this phase $\alpha > \alpha_2(N)$ the disconnected or shattered phase. The rest of the nodes, those unable due to degree constraints to form cliques, will continue to be connected and follow qualitatively similar regimes, but now for a new degree distribution that excludes the separated nodes.

In the previous list we have presented a very general picture regarding the phenomenology of ensemble (4). In particular panel (b) in figure 1 follows closely this qualitative picture. We expect the behaviour in the connected phase, $\alpha < \alpha_1(N)$, to be general. For the behaviour between $\alpha_1(N)$ and $\alpha_2(N)$ we observe a dependence on the the degree distribution, but still a strong similarity in the sharp drop in the value of $r(\alpha)$. The details for the behaviour for $\alpha > \alpha_2(N)$ have a much stronger dependence on the degree sequence and on the system size, an extensive exploration of this regime is beyond the scope of this paper since the MCMC simulations would require an enormous amount of computation time and power. Before developing the theory in sections 3 and 4, we would like to present the following list of subtleties necessary to understand the further theoretical arguments.

The transitions at $\alpha = \alpha_{1,2}(N)$ are not phase transitions in the conventional sense—they depend on N, which is taken to be large but still finite—so they are not marked by non-analyticities in the thermodynamic limit. The definitions given for $\alpha_1(N)$ and $\alpha_2(N)$ are instead of a descriptive nature, marking the α-values where $\langle r(\mathbf{A})\rangle$ drops for the first time and where $\langle n(\mathbf{A})\rangle$ increases for the first time, respectively. As will become apparent in the next section, the system size N affects severely the ensemble. Additionally, hysteresis has been observed in [19] for this ensemble. This implies that certainly the observed values of $\alpha_{1,2}(N)$ would be shifted to the left from what is observed in figure 1 if the MCMC simulations were to be performed starting with seed graphs with high three-cycle density. In this letter we will only focus on writing an approximate theory of what happens when going left to right, that is starting at $\alpha = 0$ and increasing its value gradually. The analytic predictions obtained will be validated by the MCMC sampling performed with this prescription.

We make a distinction between bounded and unbounded degree distributions $p(k)$, since boundedness affects the way in which the ensemble behaves with increasing N, e.g. in the asymptotics of $\alpha_1(N)$ and $\alpha_2(N)$.

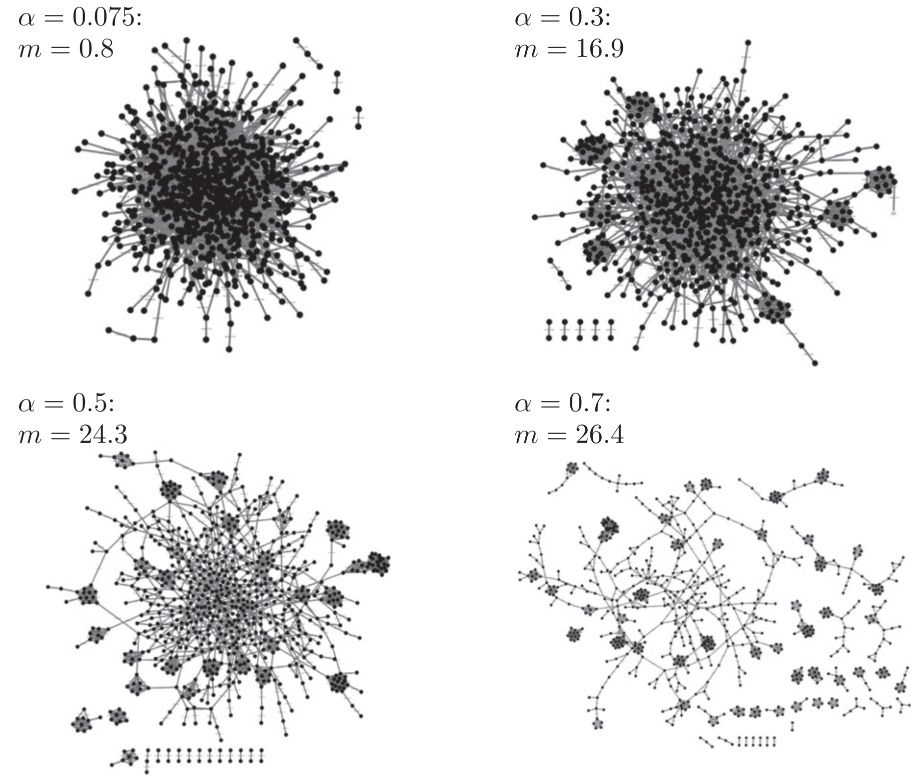

Figure 3. Results of numerical sampling simulations of the random graph ensemble (4), with $N = 1000$ and $p(k) = \frac{1}{4}\left(\frac{4}{5}\right)^k$. The four different images correspond to four different values of α, with different cycle densities $m(\alpha)$, as indicated. Zero degree nodes are omitted.

For large graphs with bounded degree distributions both transitions are close, $\alpha_1(N) \approx \alpha_2(N)$. There is a sudden appearance of disconnected cliques of $q+1$, nodes giving rise to a sharp jump in $m(\alpha)$. Strong numerical evidence and mathematical arguments support the proposition that $\alpha_1(N)$ and $\alpha_2(N)$ both scale as $\mathcal{O}\left(\log N\right)$. For graphs with unbounded $p(k)$, the maximum degree present in the graph will diverge slowly with N. Hence there are not many nodes of large degree to create cliques, and the structures created when the graphs shatter are less clear. The asymptotics of $\alpha_1(N)$ and $\alpha_2(N)$ should depend heavily on the tail of $p(k)$, as this tail governs the growth of the maximum degree with N. Nevertheless, in both cases the ensemble will end in a set of disconnected cliques, as this is the graph that maximizes the number of cycles around each node. The difference between bounded and unbounded $p(k)$ can be seen clearly when comparing figures 3 and 2. For the graph with a bimodal degree distribution in figure 3 the cliques appear immediately as the graph clusters, while for the one with an exponential distribution in figure 2 one can see clusters appearing before the breaking down of the graph.

Expressions like (6) are hard to evaluate analytically, especially for a finite N. The typical approach of statistical mechanics would be to derive exact results in the limit $N \to \infty$, and then to show they are a good approximation for finite N. In contrast, here it is important *not* to take the limit $N \to \infty$, but rather to work with asymptotically vanishing expressions for the three-cycle density, $m(\alpha) = \mathcal{O}\left(N^{-\delta}\right)$. A clear example is that of the *connected non interacting regime*; here equation (9) shows correctly that $\lim_{N\to\infty} m(\alpha) = 0$, but it is the way in which $m(\alpha)$ approaches 0 that gives us formula (9), which is seen to be very accurate. One would normally rescale α with N to avoid this effect, but it will become clear that in that case $m(\alpha_1(N)) \to 0$ for any proper scaling of α with N, meaning that the description of the first regime would vanish, which is not something we want.

Regarding the sampling, we note that convergence from a given seed graph towards equilibration requires increasing numbers of edge swaps as α is increased. Only for values in the connected regime $\alpha \in [0, \alpha_1(N))$ will equilibration be fast enough to sample graphs in a reasonable amount of time on a personal computer. Close to the transitions there is a significant divergence of relaxation times. We conjecture that the main reason for this change is precisely the clustering of triangles: in order to break a clique one has to destroy many triangles, an event that becomes extremely unlikely during the dynamics for large graphs. Therefore we expect there to be an effective breaking of ergodicity when sampling with MCMC for $\alpha \geqslant \alpha_2(N)$ and large N. Therefore, even though we performed extensive numeric simulations, we have no guarantee that the MCMC was properly equilibrated close to the transitions. Due to hysteresis, we would require repetition of the MCMC sampling starting from different seeds with very different initial three-cycle densities. Rather than pursuing this path, we will focus only on painting a general picture of what happens when increasing α starting from $\alpha = 0$.

For the above reasons, from the point of view of applied network science, working with ensemble (4) has to be done carefully. Given a seed network, it is possible to randomize via edge swaps while retaining the value of the three-cycle density, but there will be two problems. First, it could be that it takes a long time to sample correctly. Second, it could be that samples generated with the same three-cycle density have completely different topologies, according to their values of $r(\mathbf{A})$. The first problem is a matter of computing power and speed. The second problem is more tricky, and essentially unsolvable without modifying (4). If the graph one wants to randomize has a value of $r(\mathbf{A})$ that deviates significantly from $\langle r(\mathbf{A})\rangle$, then all samples will be typically very different in structure, even though they share the same three-cycle density.

3. The connected regime

We will now present an effective approximation for the generating function (6). It is analogous to the one presented in [27], but generalized for an arbitrary degree distribution $p(k)$ with finite first and second moments. We use a small α (or large N) approximation to derive (9), using a known result about the distribution of triangles in the CM [39]. It is found to give very good results, suggesting it could be exact asymptotically, at least for bounded degree distributions. If we denote by $T(\mathbf{A})$ the number of triangles in $\mathbf{A}$, we have (due to overcounting):

$$\mathrm{Tr}(\mathbf{A}^3) = 6T(\mathbf{A}). \tag{10}$$

We can therefore calculate the generating function (6) as follows:

$$\begin{aligned}\phi(\alpha) &= \frac{1}{N}\log\sum_{\mathbf{A}} \mathrm{e}^{6\alpha T(\mathbf{A})}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}}\\ &= \frac{1}{N}\log\sum_{T}\mathrm{e}^{6\alpha T}P_N(T) + \frac{1}{N}\log\mathcal{N}_{\mathbf{k}}.\end{aligned} \tag{11}$$

Where we have introduced,

$$P_N(T) = \frac{1}{\mathcal{N}_{\mathbf{k}}}\sum_{\mathbf{A}}\delta_{T,T(\mathbf{A})}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}} \tag{12}$$

$$\mathcal{N}_{\mathbf{k}} = \sum_{\mathbf{A}}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}}. \tag{13}$$

Note that we have introduced $P_N(T)$ by multiplying and dividing by the total number of graphs with a given degree sequence $\mathbf{k}$, denoted by $\mathcal{N}_{\mathbf{k}}$. Our approximation now consists in replacing $P_N(T)$ by the known asymptotic distribution of isolated triangles, that is triangles that do not share edges or nodes. The latter was computed rigorously in [39]:

$$P_N(T) \approx \mathrm{Poiss}(T|\lambda_t) = \mathrm{e}^{-\lambda_t}\frac{(\lambda_t)^{\mathrm{T}}}{T!} \tag{14}$$

$$\lambda_t = \frac{1}{6}\left(\frac{\sum_{i=1}^{N}k_i(k_i-1)}{\sum_{i=1}^{N}k_i}\right)^3 = \frac{1}{6}\left(\frac{\overline{k^2}}{c}-1\right)^3. \tag{15}$$

This then leads us to the the following approximation for (11) and $m(\alpha)$:

$$\phi(\alpha) \approx \frac{1}{N}\lambda_t\left(\mathrm{e}^{6\alpha}-1\right) + \frac{1}{N}\log\mathcal{N}_{\mathbf{k}} \tag{16}$$

$$m(\alpha) \approx \frac{1}{N}6\lambda_t\mathrm{e}^{6\alpha} = \frac{1}{N}\left(\overline{k^2}/c-1\right)^3\mathrm{e}^{6\alpha}. \tag{17}$$

This formula has a simple interpretation. At $\alpha = 0$ it correctly predicts the expected number of triangles in a CM, where one pictures these triangles to be very far away from each other. When $\alpha > 0$ this number of triangles is multiplied by $\mathrm{e}^{6\alpha}$, giving another finite but larger amount of triangles when $N \to \infty$. In this scenario we would view these triangles to be simply further and further apart as the system size grows. This picture will be revisited in the next section.

Table 1. Different degree distributions used for numerical experiments.

Type	Name	Formula $p(k)$	Parameter values
Unbounded	Exponential	$\exp(k\vert c)=\left(\frac{c}{c+1}\right)^k\frac{1}{c+1}$	$c=3,4,5,10$
Unbounded	Poissonian	$\mathrm{Poiss}(k\vert c)=\mathrm{e}^{-c}\frac{c^k}{k!}$	$c=3,4,5,10$
Unbounded	Power law	$\mathrm{PL}(k)=Ak^{-\gamma}k\geqslant 2$	$\gamma=4\ (\overline{k}\approx 2.5)$
Bounded	Bimodal	$\mathrm{bim}(k\vert 3,q)=\frac{1}{2}(\delta_{k,3}+\delta_{k,q})$	$q=5,7,9$
Bounded	Uniform	$u(k)=\frac{1}{5}\sum_{j=1}^{5}\delta_{k,j}$	—
Bounded	Non uniform	$v(k)=\sum_{j=1}^{5}w_j\delta_{k,j}$	$\boldsymbol{w}=(\frac{1}{10},\frac{2}{10},\frac{3}{10},\frac{3}{10},\frac{1}{10})$

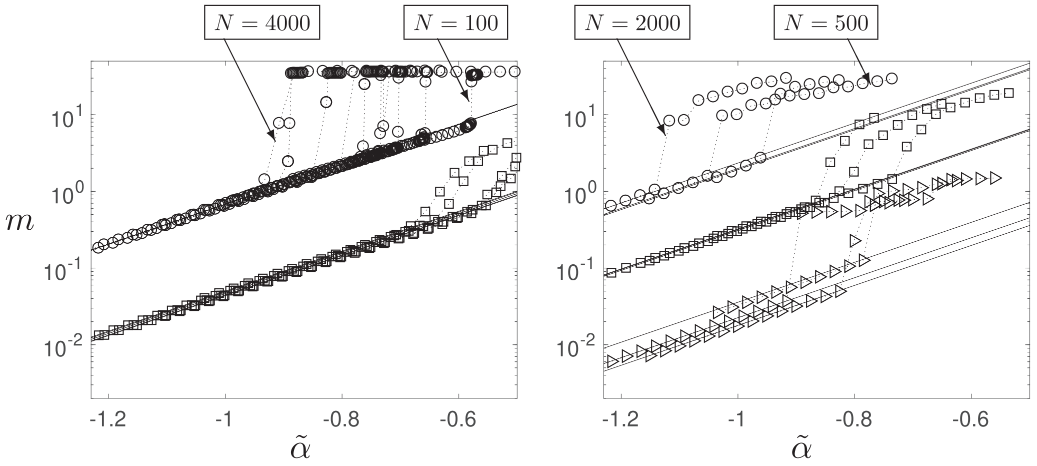

Figure 4. The three-cycle density m as measured in numerical MCMC simulations (see text for details) of the ensemble (4), plotted against the rescaled control parameter $\tilde{\alpha}=\alpha-\frac{1}{6}\log N$. Left panel: $p(k)=\frac{1}{2}\delta_{k3}+\frac{1}{2}\delta_{k9}$ with circles, for system sizes $N=100,200,300,400,500,750,1000,2000,4000$, from right to left. And, $p(k)=\frac{1}{5}\sum_{j=1}^{5}\delta_{kj}$ with squares, for system sizes $N=500,750,1000,2000$, from right to left. Right panel: $p(k)=\exp(k|5)$ (circles), $p(k)=\mathrm{Poiss}(k,5)$ (squares), and $p(k)=PL(k)$ (triangles), all for system sizes $N=500,1000,2000$. See table 1 for the relevant definitions. Error bars were omitted for clarity. The solid lines correspond to the corresponding theoretical prediction (18). Markers are joined with dotted lines and the system size of certain curves is indicated for better readability.

We have tested the above approximation extensively with numerical simulations. We generated samples from (4) for many different degree distributions, shown in table 1. The results are shown in figure 4, where we have plotted the results for systems of multiple sizes $N\sim 100$–4000. In order to have a better visualization, we plotted the cycle densities against a rescaled parameter $\tilde{\alpha}$, defined via $\alpha=\tilde{\alpha}+\frac{1}{6}\log N$,

$$m\left(\tilde{\alpha}+\frac{1}{6}\log N\right)\approx\left(\frac{\overline{k^2}}{c}-1\right)^3\mathrm{e}^{6\tilde{\alpha}}\quad\text{for}\quad\tilde{\alpha}\leqslant\tilde{\alpha}_1(N).\tag{18}$$

For the MCMC we used a different initial graph for each size and each degree distribution, $p(k)$. To generate each initial graph we used a sample from the CM for a given degree sequence sampled from the corresponding $p(k)$. In this regime, $\alpha\in[0,\alpha_1(N)]$, we used waiting times of 2×10^4 attempted edge swaps per link (AESPL), and subsequently recorded 20 samples spaced by 2×10^3 AESPL. To show the accuracy of the theory with a modest number of samples, we plot the average of the three-cycle density over the full time series of cycle densities between samples. We do this to reduce noise, and because our theory refers to the average (7), not to graph instances, since there is no self averaging at finite sizes. For graphs larger than 500 nodes, error bars are of the order of magnitude of the markers. For smaller graphs the error bars can be appreciated on the right panel of figure 6. In the remaining three-cycle density plots the error bars were omitted, in order to avoid cluttering of figures. Note that the scaling in (18) collapses all curves of the same degree distribution, up to a certain value $\tilde{\alpha}_1(N)$. As we will show in the next section, the three-cycle density at the transition vanishes as $N\to\infty$, $m(\tilde{\alpha}_1(N))\to 0$. This can be clearly seen in figure 4.

The accuracy of (9) suggests that it could be the exact asymptotic result when $N\to\infty$. This would imply that a bias of the form (4) with $\alpha=\mathcal{O}(1)$ only modifies the number of expected triangles in large graphs by an $\mathcal{O}(1)$ amount, implying that the three-cycle density will still vanish asymptotically. To achieve a nonvanishing three-cycle density in the asymptotic limit, a different scaling of α should be introduced, as was done in [22] for two-regular graphs, i.e. for $p(k) = \delta_{k,2}$. However, as will be discussed in the next section, for general degree distributions the effect of scaling α with N is much more complicated than in the two-regular case.

4. The clustered and disconnected regimes

4.1. General results

We next investigate the behaviour of the ensemble beyond the clustering transitions, i.e. for $\alpha > \alpha_1(N)$, where (9) no longer reproduces the correct three-cycle density. For the two-regular case, the only three-cycle that can exist inside a graph is an isolated cycle, therefore it is possible for (18) to be exact asymptotically. For other degree distributions, many other cycle structures can appear in a graph. As we will show, it seems that structures with strongly interacting triangles dominate entropically. Therefore the statistics of different local structures needs to be taken into account, making (9) insufficient to describe the ensemble for all values of α.

In the regime $\alpha < \alpha_1(N)$, the desired three-cycle density is achieved by creating further triangles that are independent and far from each other, without sharing nodes. For $\alpha > \alpha_1(N)$, in contrast, the desired three-cycle density is achieved by creating triangles that share as many edges as possible. This qualitative change appears to be purely entropic, since the latter regime appears for all cycle densities as long as the system is large enough, that is even for very small values of m. Put differently, the transition at $\alpha_1(N)$ does *not* happen because there are too many triangles which need to share nodes due to of lack of space in the graph, as one might guess initially. The transition happens because for a given three-cycle density the number of graphs one can create by 'putting triangles aside' in small clusters is larger than the number of graphs one can create by embedding them in the graph in a non-interacting way. While we cannot prove this assertion rigorously, extensive numerical experiments support this claim.

We measured the interaction between cycles in samples of (4) using the observable $r(\mathbf{A})$ defined in (8). The empirical value $r(\alpha) = \langle r(\mathbf{A})\rangle$ was measured in all the numerical experiments listed in table 1. For values $\alpha > \alpha_1(N)$ we increased the number of AESPL by a factor ten, giving waiting times of 2×10^5 AESPL and inter-sample intervals of 2×10^4 AESPL. For each degree distribution and each size 20 samples were taken. In all experiments we observed the same behaviour as shown for the two cases in figure 5. An initial phase of $\langle r(\mathbf{A})\rangle \approx 1/2$, indicating non-interacting cycles, is followed by a sudden drop to $\langle r(\mathbf{A})\rangle = r_{\min}(N) < 1$, indicating interacting cycles. At the value of α marking this sudden drop, which we defined to be $\alpha_1(N)$, the graph has become clustered in order to achieve the desired three-cycle density. This α value coincides precisely with the point where formula (9) stops working, as can be seen in figure 1.

When increasing the system size N, it is clear that the initial parts of the curves tend to flatten to plateaux at the level $r = 1/2$. This is consistent with the fact that equation (9), which accurately describes the three-cycle density in this regime, was derived assuming an underlying Poissonian distribution of triangles; the latter assumes, in turn, that the triangles are non-interacting [39].

The remaining question is how the two values $\alpha_1(N)$ and $r_{\min}(N)$ depend on N. For $r_{\min}(N)$ the following possibilities must be considered:

(a) $\lim_{N\to\infty} r_{\min}(N) = r^* > 0$.

(b) $\lim_{N\to\infty} r_{\min}(N) = 0$.

Given that for a finite graph $r(\mathbf{A})$ is always bounded from below by $r = 1/(q^2 - q)$, the second option is only a possibility for graphs with unbounded degree distribution. For $\alpha_1(N)$ we have the following possibilities, with their different physical implications:

(a) $\lim_{N\to\infty} \alpha_1(N) = \infty$, asymptotically the three-cycle density vanishes for all values of α.

(b) $\lim_{N\to\infty} \alpha_1(N) = \alpha^* > 0$, there is a first order phase transition at α^*.

(c) $\lim_{N\to\infty} \alpha_1(N) = 0$, all α values have a finite density three-cycle density $m(\alpha) > 0$.

We made the distinction between bounded and unbounded distributions precisely because we believe that the behaviour of the ensemble for these distribution families might not be the same. As can already be seen in the bound $r(\mathbf{A}) \geqslant 1/(q^2 - q)$, if q is growing with N, then r can approach the value 0 arbitrarily closely, contrary to the bounded case. This can also be appreciated in figure 5, for the exponentially distributed degree distribution $\langle r(\mathbf{A})\rangle$ appears to reach a lower value for larger N. Therefore we expect $r_{\min}(N) \to r^*$ as $N \to \infty$ for bounded degree distributions and $r_{\min}(N) \to 0$ for unbounded ones. For the behaviour of $\alpha_1(N)$, all numerical simulations suggest $\alpha_1(N) \to \infty$. In the next section we will also give theoretical arguments in favor of this idea for bounded degree distributions. For the case of unbounded degree distributions more extensive numerical simulations on much larger graphs should explored in order to observe a deviation from logarithmic growing.

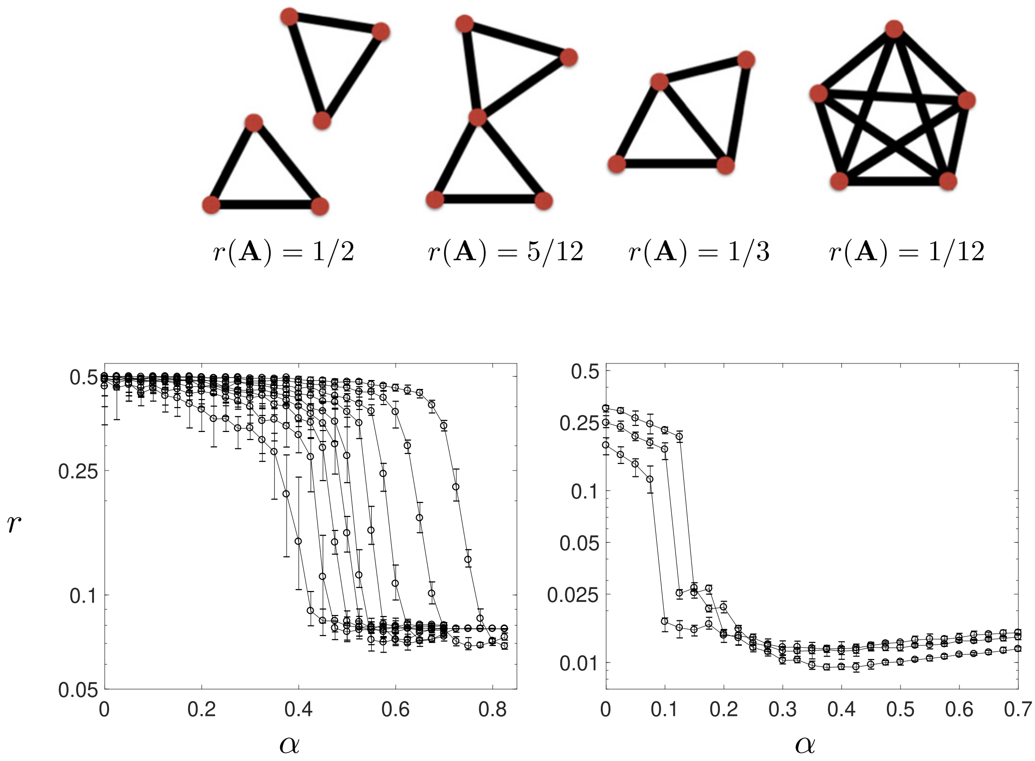

Figure 5. Top row: examples of small graphs and their corresponding values of $r(\mathbf{A})$. Bottom row: plots of $r(\alpha)$ as measured in simulations, shown versus α, values correspond to averages over 20 samples with standard deviations shown as error bars. See text for sampling details. Left: $p(k) = \frac{1}{2}\delta_{k3} + \frac{1}{2}\delta_{5k}$, with graph sizes $N = 100, 200, 300, 400, 500, 750, 1000, 2000, 4000$ (from left to right). Right: $p(k) = \mathrm{e}^{-5}5^k/k!$, with graph sizes $N = 500, 1000, 2000$ (from left to right).

For the case of bounded distributions, the maximum degree q asymptotically provides sufficiently many nodes to create cliques that will achieve the desired three-cycle density, see for example figure 3. If the desired three-cycle density is higher, then this density will be realized via cliques of the next highest degree $k < q$, in descending order. For unbounded degree distributions, this picture changes. Here one cannot guarantee the abundance of such cliques, therefore the observed topology seems to remain connected for larger values of α, in what we have called the clustered regime.

4.2. Results for bounded degree distributions

In this subsection we develop a further theoretical description of our graph ensemble for the case of bounded degree distributions. As mentioned before, numerical simulations suggest the need to include the statistics of the cliques formed by nodes of maximum degree. We denote by $K_q(\mathbf{A})$ the number of fully connected cliques of $q + 1$ nodes, and by $T^0(\mathbf{A})$ the number of triangles that are not in cliques of degree q, we use the superscript ^{0} to differentiate from the counting of *all* triangles $T(\mathbf{A})$. Since these two cases are mutually exclusive, we can then decompose the total number of three-cycles in the following way:

$$\mathrm{Tr}(\mathbf{A}^3) = 6T^0(\mathbf{A}) + (q+1)q(q-1)K_q(\mathbf{A}). \qquad (19)$$

With this decomposition we can write the partition function as

$$\begin{aligned}\phi(\alpha) &= \frac{1}{N}\log\sum_{\mathbf{A}}\mathrm{e}^{6\alpha T^0(\mathbf{A})+\alpha(q+1)q(q-1)K_q(\mathbf{A})}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}}\\ &= \frac{1}{N}\log\sum_{T^0,K}P_N(T^0,K)\mathrm{e}^{6\alpha T^0}\mathrm{e}^{q(q^2-1)\alpha K} + \frac{1}{N}\log\mathcal{N}_{\mathbf{k}},\end{aligned} \qquad (20)$$

where we introduced $\mathcal{N}_{\mathbf{k}} = \sum_{\mathbf{A}}\prod_{i\leqslant N}\delta_{k_i,\sum_j A_{ij}}$, and the joint distribution of triangles and cliques for the unbiased CM,

$$P_N(T^0,K) = \frac{1}{\mathcal{N}_{\mathbf{k}}}\sum_{\mathbf{A}}\delta_{T^0,T^0(\mathbf{A})}\delta_{K,K_q(\mathbf{A})}\prod_{i=1}^{N}\delta_{k_i,\sum_j A_{ij}}. \qquad (21)$$

Our main approximation consists in assuming that asymptotically the random variables T^0 and K become independent, each described by a Poisson distribution. This means that we again assume the main contribution of triangles for $T(\mathbf{A})$ to come from isolated triangles. Since isolated triangles and cliques are almost independent, and are rare events in the CM, one could argue that according to the *Poisson paradigm* in [40], they should both be Poissonian random variables. For a similar argument regarding cycles of different lengths see [41]. In fact since cliques are so rare, we can use the same distribution as in T for T^0, thus we put

$$P_N(T^0,K)\sim \mathrm{Poiss}(T^0|\lambda_t)\mathrm{Poiss}(K|\lambda_{K_q}(N)). \tag{22}$$

We can then immediately proceed to calculate the partition function,

$$\phi(\alpha)\approx\frac{\lambda_t}{N}\left(\mathrm{e}^{6\alpha}-1\right)+\frac{\lambda_{K_q}(N)}{N}\left(\mathrm{e}^{q(q^2-1)\alpha}-1\right)+\frac{1}{N}\log\mathcal{N}_{\mathbf{k}}, \tag{23}$$

which leads to the following expression for the three-cycle density,

$$m(\alpha)\approx\frac{6\lambda_t}{N}\mathrm{e}^{6\alpha}+\frac{q(q^2-1)}{N}\lambda_{K_q}(N)\mathrm{e}^{q(q^2-1)\alpha}. \tag{24}$$

Contrary to the regular case discussed in [27], there is for an arbitrary $p(k)$ no established rigorous result for the expected number $\lambda_K(N)$ of cliques. Nevertheless, there is a good idea of what its scaling with N should be [42]. The expected number of isomorphisms of a given strictly balanced graph H (see [42] for definition) is expected to be $\mathcal{O}\left(N^{v(H)-e(H)}\right)$, where $e(H)$ and $v(H)$ are the number of edges and nodes of H respectively. In the case of a clique of $q+1$ nodes these numbers are, $e(K_q)=\frac{1}{2}q(q+1)$ and $v(K_q)=q+1$. Therefore,

$$\lambda_K(N)=\mathcal{O}\left(\frac{1}{N^{\frac{1}{2}q(q-1)-1}}\right)\sim\frac{c_q}{N^{\frac{1}{2}q(q-1)-1}q(q^2-1)}. \tag{25}$$

We have included the factor $q(q^2-1)$ in the denominator for convenience. With this expression we obtain the following result for small values α

$$m(\alpha)\approx\frac{1}{N}\left(\overline{k^2}/c-1\right)^3\mathrm{e}^{6\alpha}+\frac{c_q}{N^{\frac{1}{2}q(q-1)}}\mathrm{e}^{\alpha q(q^2-1)}. \tag{26}$$

The first term corresponds to the contribution from isolated triangles at low density, to be denoted by $m_t(\alpha)$. The second term represents triangles in the previously described cliques, we denote it as $m_K(\alpha)$. The latter is bounded since the number of cliques of $q+1$ nodes is bounded. This then gives

$$m_K(\alpha)\approx\begin{cases}N^{-\frac{1}{2}q(q-1)}c_q\mathrm{e}^{q(q^2-1)\alpha} & \text{if}\quad \alpha\leqslant\alpha_2(N)\\ \quad p(q)q(q-1) & \text{if}\quad \alpha\geqslant\alpha_2(N)\end{cases}. \tag{27}$$

It is convenient to define the shattering transition as the point where all the cliques of degree q have appeared. This automatically gives an estimate of how $\alpha_2(N)$ behaves with N. Here we can see that $\alpha_2(N)$ diverges logarithmically with N:

$$\alpha_2(N)=\frac{1}{2(q+1)}\log N+\frac{1}{q(q^2-1)}\log\left[\frac{p(q)q(q-1)}{c_q}\right]. \tag{28}$$

This result depends on the degree distribution $p(k)$ explicitly through q and $p(q)$, but also implicitly through c_q. Since we do not generally know c_q, we cannot test the accuracy of the above prediction directly. Only for regular graphs c_q is available, leading to accurate predictions for $\alpha_2(N)$ [22]. However, alternative tests are possible. Equations (28) predicts a collapse of the various $\alpha_2(N)$ curves under the following change of variable, $\alpha=\gamma+\frac{1}{2(q+1)}\log N$,

$$m\left(\gamma+\frac{1}{2(q+1)}\log N\right)\approx\begin{cases}N^{-\frac{q-2}{q+1}}\mathrm{e}^{6\gamma} & \text{for}\quad \gamma\leqslant\gamma_1(N)\\ c_q\mathrm{e}^{q(q^2-1)\gamma} & \text{for}\quad \gamma_1(N)\leqslant\gamma\leqslant\gamma_2(N)\end{cases}. \tag{29}$$

Even though it is hard to sample graphs very precisely in the clustering regime, given that the waiting time of the MCMC algorithm is very large, overall the transition points of the curves do collapse nicely, as can be seen in figure 6. We stress that close the transition waiting times were so long that points on the steep part of the left panel on figure 6 were probably not equilibrated for system sizes $N\geqslant 1000$. For this reason we show in the right panel smaller system sizes $N=200, 300, 400$. In these cases we did not observe any drift in the empirical value of $m(\alpha)$ in the time scale of our simulations for the values close to the transition (except *at* the transition). The reader should keep in mind this still might not be the true value for (4) due to the presence

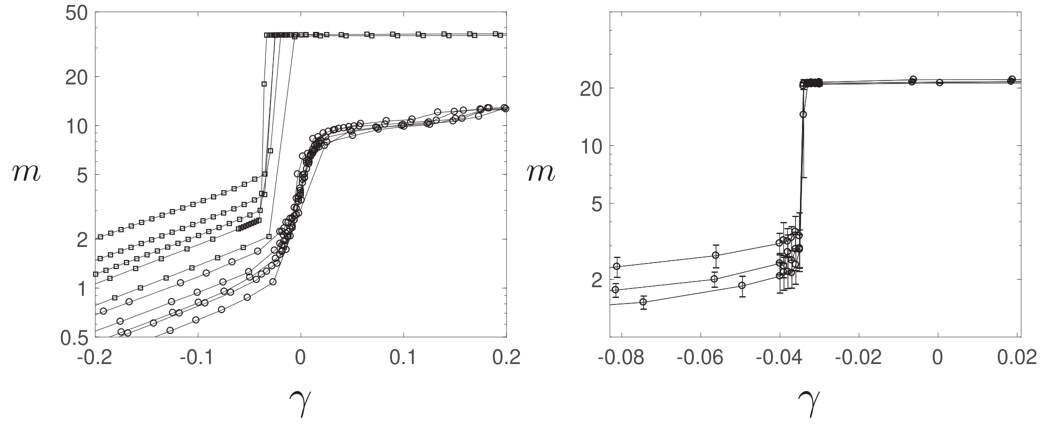

Figure 6. Plots of m against rescaled variable $\gamma = \alpha - [2(q+1)]^{-1}\log N$, showing the collapse of the second (shattering) transition point for different system sizes, predicted by (29). Left: $p(k) = \frac{1}{2}\delta_{k3} + \frac{1}{2}\delta_{k5}$ (circles), and $p(k) = \frac{1}{2}\delta_{k3} + \frac{1}{2}\delta_{k9}$ (squares). System sizes were $N = 200, 300, 400, 500, 750$, from bottom to top. Averages over 20 samples, error bars are omitted to reduce cluttering. Right: close-up in the neighbourhood of the shattering transition, for $p(k) = \frac{1}{2}\delta_{k3} + \frac{1}{2}\delta_{k7}$, for system sizes $N = 200, 300, 400$ (from bottom to top). Here the error bars correspond to average plus/minus one standard deviation.

Table 2. Comparison of the slope of $\alpha_2(N)$ plotted against $\log N$, as measuerd from data in figure 7, versus the theoretically predicted value $[2(q+1)]^{-1}$ of (28). The degree distributions $\mathrm{bim}(k|a,b)$, $u(k)$ and $v(k)$ are defined as in table 1. Standard deviations shown in parentheses.

$p(k)$	$\mathrm{bim}(k\|3,5)$	$\mathrm{bim}(k\|3,7)$	$\mathrm{bim}(k\|3,9)$	$u(k)$	$v(k)$
Theory	$0.08\overline{3}$	0.0625	0.05	$0.08\overline{3}$	$0.08\overline{3}$
Simulation	0.079(0.005)	0.066(0.002)	0.057(0.003)	0.10(0.01)	0.10(0.01)

of hysteresis, but as discussed in section 2, we will only work with the value obtained when increasing α from $\alpha = 0$. We do see an almost perfect collapse of the transitions points of the curves. As described before, we used waiting times of 2×10^5 AESPL and inter-sample intervals of 2×10^4 AESPL. The values reported are the averages over the full time series as described in the previous section for figure 4.

The prefactor slope of $\frac{1}{2(q+1)}$ for the term proportional to $\log N$ in $\alpha_2(N)$ in (28) was also tested. To do this we estimated $\alpha_2(N)$ as the first jump of $m(\alpha)$. This was easily detected with a change of sign in the difference with prediction (9), as $m(\alpha)$ was always overestimated before the transition. This property can be seen clearly in figure 8. Results are presented in table 2 and in panel (c) of figure 7. We find a very good agreement for the bimodal distributions. For distributions $u(k)$ and $v(k)$ the prediction is close enough to the predicted value $0.8\overline{3}$, but the observed value of 0.10(0.01) in both cases is actually closer to what we would observe with $q = 4$. This is consistent with the fact that, for these particular distributions, both degrees have a similar density and $k = 4$ is more abundant in the case of $v(k)$.

With our estimate for $\alpha_2(N)$ we can also derive an upper bound on the three-cycle density achieved in the connected regime,

$$m_u = \frac{1}{N}\left(\overline{k^2}/c - 1\right)^3 \mathrm{e}^{6\alpha_2(N)} = \frac{1}{N^{\frac{q-2}{q+1}}}(\overline{k^2}/c - 1)^3 (p(q)q(q-1)/c_q)^{\frac{6}{q(q^2-1)}}. \qquad (30)$$

This value corresponds to the three-cycle density that would be reached if the contribution of cliques were not present. Given that cliques appear before, it becomes impossible to reach this density in the connected phase. Even though c_q is unknown, we can conclude that m_u vanishes when $N \to \infty$, which is indeed consistent with numerical experiments, as can be seen in panels (a) and (b) of figure 7. The results are very good when looking at the chosen bimodal degree distributions, $p(k) = \frac{1}{2}\delta_{k3} + \frac{1}{2}\delta_{kq}$. Figure 7 confirms two theoretical predictions. First, we see that the last value of the three-cycle density before the steep jump into the clustered phase scales with N in the manner predicted by (30). Second, the final value of the jump at $\alpha_2(N)$ coincides with the prediction $p(q)q(q-1)$, as indicated by the dotted-dashed line in panel (a) of figure 7.

Note that even though we predict a vanishing density $m(\alpha)$ for all values of α in the asymptotic limit $N \to \infty$, the transition at $\alpha_2(N)$ does behave similar to a phase transition. A detailed exploration is presented in [19] where hysteresis is shown to be present in this model, (4). It is interesting to note that even though we did not explore the effects of hysteresis we still got good results with our theory, and even a very good match for regular graphs as shown in [27]. The reason behind this could be that the Poissonian assumption

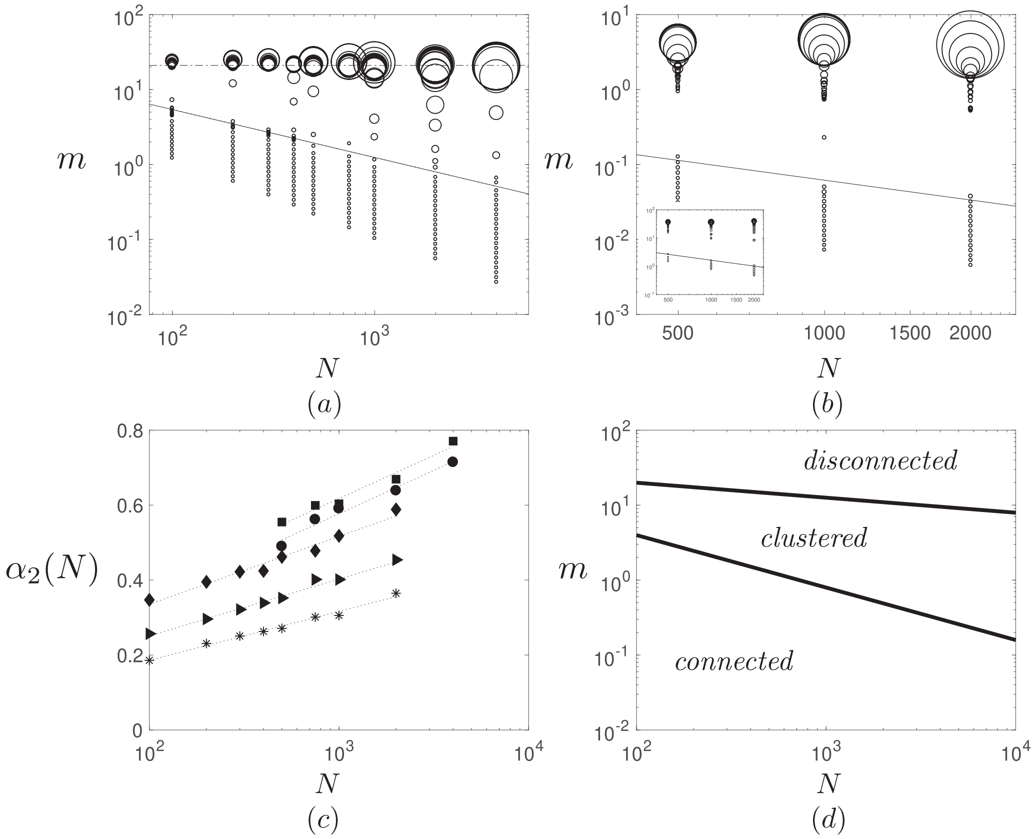

Figure 7. Top row: scatter plots of triangle density m shown versus system size N. The width of the markers is proportional to the number of connected components. (a): $p(k) = \mathrm{bim}(k|3,7)$, with dashed-dotted line corresponding to the prediction $p(q)q(q-1)$ of (26). The observed slope of $-0.64(5)$ is consistent with the predicted -0.625 of (30). (b): $p(k) = \mathrm{PL}(k)$, and inset $p(k) = \exp(k|5)$. Bottom row, here solid lines are only guides to the eye. (c): linear/log plot of $\alpha_2(N)$ versus N from simulation data. Degree distributions: $u(k)$ (circles), $v(k)$ (squares), $\mathrm{bim}(k|3,5)$ (diamonds), $\mathrm{bim}(k|3,7)$ (triangles), and $\mathrm{bim}(k|3,9)$ (asterisks). Dotted line shows linear fit in good agreement with theoretical prediction (28), see table 2. (d): conjectured phase diagram of the ensemble (4), in the (m, N) plane.

(22) implies that we only sum over states with low three-cycle density when calculating (23). As mentioned before, a different scaling of α with N could be chosen in order to make $\alpha_2(N)$ independent of N, for example the previously shown $\alpha = \gamma + (2(q+1))^{-1}\log N$. Nevertheless, as it is shown in equation (29), this different scaling would imply a vanishing of the connected phase. Since we want to write down a theory that incorporates all the regimes we chose to keep $\alpha = \mathcal{O}(1)$ and N large but finite.

As a final comment, we point out that the Poissonian assumption of (22) implies that the shattering transition is of an entropic nature. To see this, we can study the behaviour of the ratio

$$\frac{\mathcal{N}(\mathbf{k}|T)}{\mathcal{N}(\mathbf{k}|K)} = \frac{\#\text{of graphs with degree sequence } \mathbf{k} \text{ and } T \text{ isolated triangles}}{\#\text{of graphs with degree sequence } \mathbf{k} \text{ and } K \text{ q} - \text{regular cliques}}. \tag{31}$$

If we fix the three-cycle density to any arbitrary value $m^* < p(q)q(q-1)$, this value can be achieved by the following numbers of triangles or cliques.

$$T = \frac{m^*}{6}N, \qquad K = \frac{m^*}{q(q-1)}N. \tag{32}$$

Using the Poissonian assumption, we can then prove (see appendix B) that

$$\lim_{N\to\infty} \frac{\mathcal{N}(\mathbf{k}|m^*N/6)}{\mathcal{N}(\mathbf{k}|m^*N/(q^2-q))} = \lim_{N\to\infty} \mathrm{e}^{-\frac{m^*}{6}\frac{q-2}{q+1}N\,\log\,N} = 0. \tag{33}$$

Hence, no matter how small m^* is, for a large enough system there will always be infinitely many more graphs that achieve it via cliques than via isolated triangles.

Table 3. Comparison of three-cycle density, $m(\alpha)$; interaction, $r(\alpha)$; ASPL; and diameter. G297 ($N = 121$, $c = 2.5, \overline{k^2} = 7$) is compared with graphs generated according to (4) for $\alpha^* = 0.3825$ and with graphs generated with CM. Averages are over 100 samples and standard deviations are shown in parentheses. For ASPL and diameter only finite lengths where taken into account.

$p(k)$	m	r	ASPL	Diameter
G297	0.4463	0.3333…	10.9596	29
(4)	0.45(0.16)	0.40(0.08)	7.2(0.6)	18(3)
CM	0.06(0.05)	0.49(0.02)	6.5(0.3)	15(2)

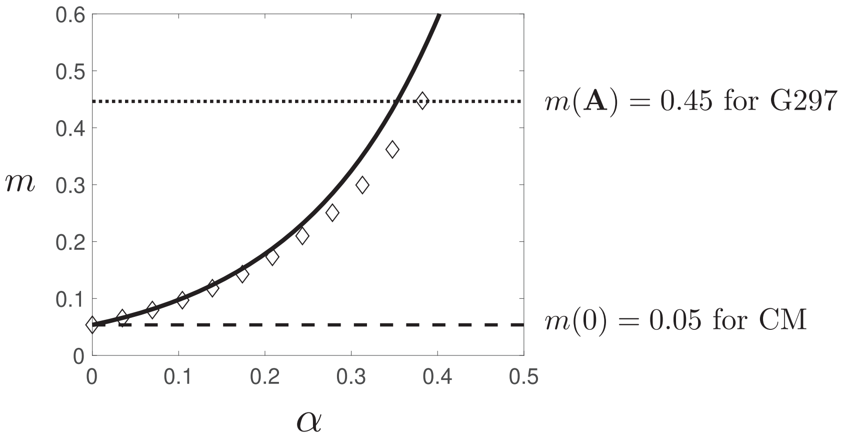

Figure 8. We show how the average three-cycle density is increased from the CM value of 0.05 by an order of magnitude up to the target value of 0.45 observed in G297. Diamonds correspond to the average over 100 samples generated with MCMC. The solid line corresponds to the theoretical prediction (9), we can see it gives a reasonable estimate for the necessary value of α.

5. Discussion

In this letter we have presented and analyzed a random graph ensemble were samples are both sparse and with a tuneable amount of three-cycles. Even though this ensemble (4) can be regarded as the simplest random graph ensemble with tuneability of short cycles, it is found to exhibit rather nontrivial behaviour. While one would hope for and expect a smooth and easy controllability of the three-cycle density via the control parameter α, we see that in fact there are very special nontrivial regimes, and there is surprisingly a very strong influence of the system size, i.e. the number of nodes in the graphs. Still, with appropriate care this ensemble could be used by practitioners of network science as a null model of networks with a nontrivial amount of three-cycles. If one has a given real network $\mathbf{A}_0$, that is to be compared with random samples having the same three-cycle density $m(\mathbf{A}_0)$, we propose the following steps should be taken:

(a) Calculate the following properties of the initial graph: $\mathbf{k}(\mathbf{A}_0)$, $m(\mathbf{A}_0)$, $r(\mathbf{A}_0)$, $n(\mathbf{A}_0)$.

(b) Sample graphs repeatedly from (4), varying α until the value α^* where observed and required cycle densities match, $m(\alpha) = m(\mathbf{A}_0)$. An initial guess for α might be $\alpha_0 = \frac{1}{6}\log\left(m(\mathbf{A}_0)N/(\overline{k^2}/c-1)^3\right)$, especially if if $\alpha_1(N) > \alpha_0$.

(c) Once cycle densities are matched, compare the other properties $r(\mathbf{A})$ and $n(\mathbf{A})$.

- If $n(\alpha) \approx n(\mathbf{A}_0)$ and $r(\alpha) \approx r(\mathbf{A}_0)$, then (4) is a suitable null model for $\mathbf{A}_0$.
- If they are different, it means that $\mathbf{A}_0$ is still extremely atypical in (4), and thus it is not a suitable null model.

As an example we compared our model to a particular graph. We chose a graph representing the structure of a protein named G297 from the public network repository [46]. Nodes represent secondary structure elements and the edges their physical proximity, as described in [47]. In table 3 we can see that its value for the interaction parameter is $r = 1/3$. This value reflects that three-cycles in this graph are interacting, i.e. sharing edges. While this means it will not look as graph from the connected regime with $r \approx 1/2$, it is still interesting to compare to graphs generated with (4), as it is still easy to generate many graphs with the desired three-cycle density in this case. In figure 8 it is shown how the target value of the three-cycle density can actually be reached by tuning α in the connected regime, in this case $\alpha^* = 0.3825$ was necessary. Although there are deviations from (9), it does give a good idea of the behaviour. Once we generate samples at the appropriate value of average three-cycle density we can compare other observables. We chose two statistics of the shortest path lengths between nodes in the graph: the average shortest path length (ASPL), and the maximum path length or diameter. For simplicity, we simply discarded all infinite values of path lengths that may occur when graphs have multiple connected components, therefore keeping our observables well defined. Increasing the number of three-cycles will decrease the shortest path lengths between certain nodes, and this indeed can be seen in table 3. It is also clear that the actual values of ASPL and the diameter of G297 are several standard deviations away from the values for the ME ensemble (4) with the same average three-cycle density $m = 0.45$. This allows us to conclude that these increased values for the ASPL and diameter are actually statistically significant and not only due to the increase in three-cycle density. One can speculate that they are due to the geometric nature of the graph, which can already be appreciated in the lower value of $r(\mathbf{A})$.

Even if all observables $m(\mathbf{A})$, $r(\mathbf{A})$ and $n(\mathbf{A})$ of initial and sampled graphs match, it still could be the case that equilibration waiting times of the MCMC are very large. For graphs of more than a thousand nodes it could take days or more to get well-mixed samples. This just shows how the applied network scientist should be cautious when applying tools like edge swapping without a proper theory.

To summarize, we present our conjectured phase diagram in figure 7 (bottom right). With an exact solution for (6) one could find an analytic expression for the phase boundaries shown. The main lessons are that the same cycle densities may have very different topologies for different system sizes, and that sampling anywhere outside the connected regime takes a very long time, potentially days or weeks for large graphs, even on fast multi-core machines. We expect that for any model, any desired three-cycle density eventually falls in the disconnected regime as N grows. For the case of bounded degree distributions with $Np(q) \gg q + 1$, the clustered region practically vanishes.

There are many directions in which to pursue further research, ranging from practical to theoretical. From a rigorous point of view it would be interesting to see how to prove or disprove any of the assertions made in this work, that is extending rigorous results of CM beyond uniform models. Additionally, longer and more extensive simulations should be carried out to try to determine the exact dependence on N of $\alpha_1(N)$ and $\alpha_2(N)$, especially to find out whether there is indeed a transition without scaling parameters for unbounded degree distributions.

The enormous waiting times seem to be due in part to the fact that in the clustered and disconnected phases many cycles have to broken in a predetermined sequence to get rid of certain structures like cliques. Given that this is unlikely, an alternative MCMC with moves that involve more edges rather than only two could be studied, in order to speed up the algorithm and let it explore more quickly the graph space.

Finally, there are many interesting questions about the spectral properties of (4) to discover. First, in [27] an analytic expression for the spectral density was found for the case of regular graphs in the connected regime. We are currently working on a generalization for an arbitrary degree distribution like in (4). The formation of clusters after the clustering transition points to a localization transition for the eigenvectors of $\mathbf{A}$. A similar observation has been made for dense graphs in [26], where its nontrivial spectral properties were found; such spectral analysis has not been done yet for the sparse case like ours.

Overall, there are many open question when it comes to presenting random counterparts of real networks. It is safe to say that they are not defined by the number of three-cycles alone. It seems like real networks occupy a very small area of the abstract graph space. Finding the correct properties that will make a ME ensemble sample from a pool of realistically looking graphs is still very much an open problem. An alternative is to impose a constraint on the full set of eigenvalues of the adjacency matrix, in this way all cycle lengths would be controlled simultaneously. This full spectral constraint has been discussed in [27, 43, 44].

Acknowledgments

FAL gratefully acknowledges financial support through a scholarship from Conacyt (Mexico). The authors thank Alexander Mozeika and Mansoor Sheikh for valuable discussions.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix A. Numerical sampling

In order for this paper to be sufficiently self-contained, we will present a brief recap of the algorithms described in [5, 35] for generating samples from nondirected random graph ensembles with hard-constrained degrees.

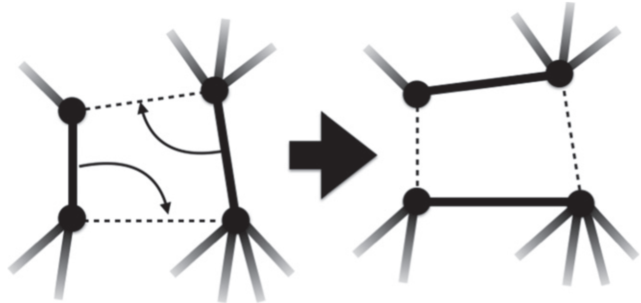

Figure A1. Edge swap for MCMC dynamics in the space of simple nondirected graphs. This is the simplest type of move in that leaves all degree invariant.

The main task is to define a Markov chain with the following characteristics:

$$p_{t+1}(\mathbf{A}) = \sum_{\mathbf{A}'\in\Omega_{\mathcal{M}}} W(\mathbf{A}|\mathbf{A}')p_t(\mathbf{A}'), \tag{A.1}$$

(a) The measure p_t converges to the invariant measure $p_\infty(\mathbf{A}) = \frac{1}{Z}\mathrm{e}^{-H(\mathbf{A})}$.

(b) The allowed transitions constitute a limited set Φ of elementary moves

$$F:\ \Omega_F \subseteq \Omega_{\mathcal{M}} \to \Omega_{\mathcal{M}}.$$

(c) For each $F\in\Phi$ there exists a unique inverse F^{-1} that acts on the same set of graphs, $\Omega_{F^{-1}} = \Omega_F$.

With these condition we will be able to define a dynamical process that will allow us to sample effectively from ensemble (4). The reason we need nontrivial moves is to be sure we respect the degree constraints; a single edge dynamics cannot achieve this. The simplest elementary move that respects the values of all degrees is called an edge swap. It involves choosing a pair of edges and interchanging them, see figure A1.

We next need to define the transition probabilities $W(\mathbf{A}|\mathbf{A}')$ of the Markov chain. They are chosen such as to obey the detailed balance condition, with (4) as invariant measure, i.e. $W(\mathbf{A}|\mathbf{A}')p_\infty(\mathbf{A}') = W(\mathbf{A}'|\mathbf{A})p_\infty(\mathbf{A})$ for all $(\mathbf{A},\mathbf{A}')$. Together with the known ergodicity of the edge swap moves [45], detailed balance is a sufficient condition to satisfy (a). We can write the transition probabilities as

$$W(\mathbf{A}|\mathbf{A}') = \sum_{F\in\Omega'} \frac{I_F(\mathbf{A}')}{n(\mathbf{A}')}\left[\delta_{\mathbf{A},F\mathbf{A}'}A(F\mathbf{A}'|\mathbf{A}') + \delta_{\mathbf{A},\mathbf{A}'}\left[1 - A(F\mathbf{A}'|\mathbf{A}')\right]\right]. \tag{A.2}$$

with the definitions

$$\begin{aligned}
\Omega' &= \{F\in\Phi|\ \exists\mathbf{A}\in\Omega_{\mathcal{M}}\ \text{s.t.}\ F\mathbf{A}\neq\mathbf{A}\}\\
I_F(\mathbf{A}) &= \begin{cases}1 & \text{if } \mathbf{A}\to F\mathbf{A} \text{ is an allowed move}\\ 0 & \text{otherwise}\end{cases}\\
n(\mathbf{A}) &= \sum_{F\in\Omega'} I_F(\mathbf{A})\\
A(F\mathbf{A}|\mathbf{A}) &: \text{acceptance probability of move } \mathbf{A}\to F\mathbf{A}.
\end{aligned} \tag{A.3}$$

The interpretation of the above transition probabilities is as follows. At each step a candidate move is chosen uniformly at random from all possible moves, with probability $1/n(\mathbf{A})$. It is then accepted with probability $A(F\mathbf{A}|\mathbf{A})$, and otherwise rejected. The acceptance probabilities must satisfy the detailed balance condition

$$(\forall\ \mathbf{A}\in\Omega)(\forall\ F\in\Omega'):\ A(F\mathbf{A}|\mathbf{A})\mathrm{e}^{-H(\mathbf{A})}/n(\mathbf{A}) = A(\mathbf{A}|F\mathbf{A})\mathrm{e}^{-H(F\mathbf{A})}/n(F\mathbf{A}). \tag{A.4}$$

This condition is satisfied by multiple choices; here we choose

$$A(\mathbf{A}|\mathbf{A}') = \frac{1}{1+\mathrm{e}^{E(\mathbf{A})-E(\mathbf{A}')}} \tag{A.5}$$

with the effective energy $E(\mathbf{A}) = H(\mathbf{A}) + \log n(\mathbf{A})$. This expression stresses the fact that the acceptance probabilities cannot depend only on the function $H(\mathbf{A})$, but also on the current state via $n(\mathbf{A})$. In [5] it is shown that $n(\mathbf{A})$ an be written explicitly as

$$n(\mathbf{A}) = \frac{1}{4}\left(\sum_i k_i\right)^2 + \frac{1}{4}\sum_i k_i - \frac{1}{2}\sum_i k_i^2 - \frac{1}{2}\sum_{ij} k_i A_{ij} k_j + \frac{1}{4}\,\mathrm{Tr}\left(\mathbf{A}^4\right) + \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{A}^3\right). \tag{A.6}$$

Appendix B. Entropic argument

Let us assume that in the CM both T and K are Poissonian random variables,

$$P_N(T) = \mathrm{Poiss}(T|\lambda_t), \qquad Q_N(K) = \mathrm{Poiss}(K|c_q/N^{d-1}) \tag{B.1}$$

with $d = \frac{1}{2}q(q-1)$. They are simply related to the number of graphs that exist, given the prescribed degree sequence, with the stated number of triangles or cliques, so

$$\frac{P_N(T)}{Q_N(K)} = \frac{\sum_{\mathbf{A}}\delta_{T,T(\mathbf{A})}\ \prod_{i=1}^N \delta_{k_i,\sum_j\ A_{ij}}}{\sum_{\mathbf{A}}\delta_{K,K(\mathbf{A})}\ \prod_{i=1}^N \delta_{k_i,\sum_j\ A_{ij}}} = \mathrm{e}^{-\lambda_t + \frac{c_q}{N^d}}\frac{(\lambda_t)^{\mathrm{T}}}{(c_q/N^{d-1})^K}\frac{K!}{T!}. \tag{B.2}$$

We want to determine for a given three-cycle density whether asymptotically there are more graphs that realize the joint values (T, K) through triangles or through cliques. For this we need to write the number of triangles and cliques in terms of the desired three-cycle density, which gives

$$T = \frac{m}{6}N, \qquad K = \frac{m}{q(q^2-1)}N. \tag{B.3}$$

We can now inspect the asymptotic limit

$$\lim_{N\to\infty}\frac{P_N\left(\frac{m}{6}N\right)}{Q_N\left(\frac{m}{q(q^2-1)}N\right)} = \lim_{N\to\infty}\exp\left(-\lambda_t + \frac{c_q}{N^d} + \frac{m}{6}N\,\log\,\lambda_t - \frac{m}{q(q^2-1)}N\,\log(c_q)\right.$$
$$\left. + \frac{md}{q(q^2-1)}N\,\log(N) + \left(\frac{m}{q(q^2-1)}N\right)! - \left(\frac{m}{6}N\right)!\right). \tag{B.4}$$

We note, upon using Stirling's expression for the factorials, that this quantity is dominated by the $N\,\log\,N$ term, since $d = \frac{1}{2}q(q-1)$. Hence

$$\lim_{N\to\infty}\frac{P_N\left(\frac{m}{6}N\right)}{Q_N\left(\frac{m}{q(q^2-1)}N\right)} = \lim_{N\to\infty}\exp\left(-m\left(\frac{1}{6} - \frac{1}{2(q+1)}\right)N\,\log\,N\right) = 0. \tag{B.5}$$

ORCID iDs

Fabián Aguirre López https://orcid.org/0000-0002-6418-8802
Anthony C C Coolen https://orcid.org/0000-0002-6976-5875

References

[1] Euler L 1741 *Commentarii Academiae Scientiarum Petropolitanae* 128–40
[2] Casella G and Berger R L 2002 *Statistical Inference* vol 2 (Pacific Grove, CA: Duxbury)
[3] Solomonoff R and Rapoport A 1951 *Bull. Math. Biophys.* **13** 107–17
[4] Erdös P and Rényi A 1960 *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17–60
[5] Annibale A, Roberts E and Coolen A C C 2017 *Generating Random Networks and Graphs* (Oxford: Oxford University Press)
[6] Newman M E J 2003 *SIAM Rev.* **45** 167–256
[7] Erdös P and Rényi A 1959 *Publ. Math. Debr.* **6** 290–7
[8] Strauss D 1986 *SIAM Rev.* **28** 513–27
[9] Jonasson J 1999 *J. Appl. Probab.* **36** 852–67
[10] Burda Z, Jurkiewicz J and Krzywicki A 2004 *Phys. Rev.* E **69** 026106
[11] Park J and Newman M E 2005 *Phys. Rev.* E **72** 026136
[12] Chatterjee S and Diaconis P 2013 *Ann. Stat.* **41** 2428–61
[13] Horvát S, Czabarka É and Toroczkai Z 2015 *Phys. Rev. Lett.* **114** 158701
[14] Yin M 2016 *J. Stat. Phys.* **164** 241–53
[15] Holme P and Kim B J 2002 *Phys. Rev.* E **65** 026107

[16] Guo W and Kraines S B 2009 *Proc. of the 2009 Int. Conf. on Computational Aspects of Social Networks* pp 10–7

[17] Newman M E 2009 *Phys. Rev. Lett.* **103** 058701

[18] Miller J C 2009 *Phys. Rev.* E **80** 020901

[19] Foster D, Foster J, Paczuski M and Grassberger P 2010 *Phys. Rev.* E **81** 046115

[20] Bianconi G, Darst R K, Iacovacci J and Fortunato S 2014 *Phys. Rev.* E **90** 042806

[21] Tamm M, Shkarin A, Avetisov V, Valba O and Nechaev S 2014 *Phys. Rev. Lett.* **113** 095701

[22] López F A, Barucca P, Fekom M and Coolen A C C 2018 *J. Phys. A: Math. Theor.* **51** 085101

[23] Avetisov V, Hovhannisyan M, Gorsky A, Nechaev S, Tamm M and Valba O 2016 *Phys. Rev.* E **94** 062313

[24] Avetisov V, Gorsky A, Maslov S, Nechaev S and Valba O 2018 *Phys. Rev.* E **98** 032308

[25] Pospelov N, Nechaev S, Anokhin K, Valba O, Avetisov V and Gorsky A 2019 *Phys. Life Rev.* **31** 240–56

[26] Avetisov V, Gorsky A, Nechaev S and Valba O 2020 *J. Complex Netw.* **8** cnz026

[27] López F A and Coolen A C C 2020 *J. Phys. A: Math. Theor.* **53** 065002

[28] Hackett A, Melnik S and Gleeson J P 2011 *Phys. Rev.* E **83** 056107

[29] Volz E M, Miller J C, Galvani A and Ancel Meyers L A 2011 *PLoS Comput. Biol.* **7** e1002042

[30] Herrero C P 2015 *Phys. Rev.* E **91** 052812

[31] Peron T K D, Ji P, Kurths J and Rodrigues F A 2018 *Europhys. Lett.* **121** 68001

[32] Herrero C P 2019 *Phys. Rev.* E **99** 012314

[33] Cantwell G T and Newman M E J 2019 *Proc. Natl Acad. Sci.* **116** 23398–403

[34] Heath L S and Parikh N 2011 *Phys.* A **390** 4577–87

[35] Coolen A C C, De Martino A and Annibale A 2009 *J. Stat. Phys.* **136** 1035–67

[36] Arratia R and Liggett T M 2005 *Ann. Appl. Prob.* **15** 652–70

[37] Cover T M and Thomas J A 2012 *Elements of Information Theory* (New York: Wiley)

[38] Jaynes E T 1957 *Phys. Rev.* **106** 620

[39] Bollobás B 1980 *Eur. J. Comb.* **1** 311–6

[40] Alon N and Spencer J H 2004 *The Probabilistic Method* (New York: Wiley)

[41] Wormald N C *et al* 1999 *London Mathematical Society Lecture Note Series* pp 239–98

[42] Bollobás B 2001 *Random Graphs* vol 73 (Cambridge: Cambridge University Press)

[43] Coolen A C C 2016 *J. Phys.: Conf. Ser.* **699** 012022

[44] Roberts E and Coolen A C C 2014 *ESAIM: Proc.* **47** 97–115

[45] Eggleton R B and Holton D A 1981 Simple and multigraphic realizations of degree sequences *Combinatorial Mathematics VIII* (Berlin: Springer) pp 155–72

[46] Rossi R A and Ahmed N K 2015 The network data repository with interactive graph analytics and visualization http://networkrepository.com

[47] Borgwardt K M, Ong C S, Schonauer S, Vishwanathan S V N, Smola A J and Kriegel H-P 2005 *Bioinform.* **21** i47–56