Limit theorems for the Erdős–Rényi random graph conditioned on being a cluster graph*

Martijn Gösgens[†1], Lukas Lüchtrath[‡2], Elena Magnanini[§2], Marc Noy[¶3], and Élie de Panafieu [‖4]

[1]Eindhoven University of Technology
[2]Weierstrass Institute, Berlin
[3]Universitat Politènica de Catalunuya, Barcelona
[4]Nokia Bell Labs France, Massy, France

Abstract

We investigate the structure of the random graph $G(n, p)$ on n vertices with constant (not depending on n) connection probability p, conditioned on the rare event that every component is a clique. We show that a phase transition occurs at $p = 1/2$, contrary to the dense $G(n, p)$ model. Our proofs are based on probabilistic methods, generating functions and analytic combinatorics.

1 Introduction

A cluster graph is a graph that is the disjoint union of complete graphs. In this paper, we consider the Erdős–Rényi (ER) random graph $G(n, p)$ on n vertices with connection probability p, conditioned on the rare event of being a cluster graph; in our situation $p \in (0, 1)$ does not depend on n. We refer to such a graph as a random cluster graph (RCG). The initial motivation for our study was the observation that a random cluster graph is a good candidate for a Bayesian prior distribution in the context of community detection [3], which is the task of partitioning the nodes of a network into communities.

Secondly, it is an interesting probabilistic object due to its rare event character. Forming a cluster graph is no standard behaviour of the ER random graph and it is fascinating how drastically its behaviour is effected by this conditioning; an evidence of this fact is that the random graph obtained after this conditioning overcomes a phase transition in p (that is not present in the dense ER model).

Finally, when ignoring the edges and only considering each cluster as a set, a cluster graph represents a partition of the whole vertex set. The case $p = 1/2$ then coincides with the uniform distribution over set partitions. Uniform set partitions are standard objects in enumerative and probabilistic combinatorics [4]. Varying the value of p is a natural way of weighting partitions and thus the RCG gives rise to more general, non-uniform underlying distributions.

After stating our main results, we briefly explain the proof techniques, based on probabilistic methods and analytic combinatorics [2]. We conclude with a sketch of further results and concluding remarks.

*This project has been initiated during the RandNET Summer School and Workshop on Random Graphs in Eindhoven in August 2022. It was supported by the RandNET project, MSCA-RISE - Marie Sklodowska-Curie Research and Innovation Staff Exchange Programme (RISE), Grant agreement 101007705

†Email: research@martijngosgens.nl
‡Email: luechtrath@wias-berlin.de
§Email: magnanini@wias-berlin.de
¶Email: marc.noy@upc.edu
‖Email: depanafieuelie@gmail.com

2 Main results

We let $\mathbf{CG}_{n,p}$ denote a random cluster graph with parameters n and p. Our main quantities of interest are the number of connected components (clusters) in $\mathbf{CG}_{n,p}$, denoted by $\mathbf{C}_{n,p}$, the number of edges denoted by $\mathbf{M}_{n,p}$, and the degree $\mathbf{D}_{n,p}$ chosen independent and uniformly at random from the vertex set. Our main results concerning these parameters are the following.

Theorem 1 (Number of clusters in the RCG)**.** *Consider the random cluster graph $\mathbf{CG}_{n,p}$ on $n \in \mathbb{N}$ vertices and ER edge probability $p \in (0,1)$ and the number of its clusters $\mathbf{C}_{n,p}$.*

1. *If $p > 1/2$, then*
$$\lim_{n\to\infty} \mathbb{P}(\mathbf{C}_{n,p} = 1) = 1.$$
Put differently, $\mathbf{CG}_{n,p} = K_n$ with high probability.

2. *If $p = 1/2$, then $\mathbf{C}_{n,p}$ obeys a central limit theorem. That is,*
$$\frac{\mathbf{C}_{n,p} - \mathbb{E}\mathbf{C}_{n,p}}{\sqrt{\mathrm{Var}(\mathbf{C}_{n,p})}} \longrightarrow \mathcal{N}(0,1),$$
in distribution, as $n \to \infty$. Moreover,
$$\mathbb{E}\mathbf{C}_{n,p} \sim \frac{n}{\log n} \quad \text{and} \quad \mathrm{Var}(\mathbf{C}_{n,p}) \sim \frac{n}{\log(n)^2}.$$

3. *If $p < 1/2$, then $\mathbf{C}_{n,p}$ obeys a central limit theorem. That is,*
$$\frac{\mathbf{C}_{n,p} - \mathbb{E}\mathbf{C}_{n,p}}{\sqrt{\mathrm{Var}(\mathbf{C}_{n,p})}} \longrightarrow \mathcal{N}(0,1),$$
in distribution, as $n \to \infty$. Moreover,
$$\mathbb{E}\mathbf{C}_{n,p} \sim \sqrt{\frac{\log(1-p) - \log p}{2}} \frac{n}{\sqrt{\log n}} \quad \text{and} \quad \mathrm{Var}(\mathbf{C}_{n,p}) = \Theta\left(\frac{n}{\log(n)^{3/2}}\right).$$

Theorem 2 (Number of edges in the RCG)**.** *Consider the random cluster graph $\mathbf{CG}_{n,p}$ on $n \in \mathbb{N}$ vertices and ER edge probability $p \in (0,1)$ and its number of edges $\mathbf{M}_{n,p}$.*

1. *If $p > 1/2$, then*
$$\lim_{n\to\infty} \mathbb{P}\left(\mathbf{M}_{n,p} = \binom{n}{2}\right) = 1.$$

2. *If $p = 1/2$, then $\mathbf{M}_{n,p}$ obeys a central limit theorem. That is,*
$$\frac{\mathbf{M}_{n,1/2} - \mathbb{E}\mathbf{M}_{n,1/2}}{\sqrt{\mathrm{Var}(\mathbf{M}_{n,1/2})}} \longrightarrow \mathcal{N}(0,1)$$
in distribution as $n \to \infty$. Moreover,
$$\mathbb{E}\mathbf{M}_{n,1/2} \sim n \log n \quad \text{and} \quad \mathrm{Var}(\mathbf{M}_{n,1/2}) = \Theta(n \log(n)^2).$$

3. *If $p < 1/2$, then $\mathbf{M}_{n,p}$ obeys a central limit theorem. That is,*
$$\frac{\mathbf{M}_{n,p} - \mathbb{E}\mathbf{M}_{n,p}}{\sqrt{\mathrm{Var}(\mathbf{M}_{n,p})}} \longrightarrow \mathcal{N}(0,1)$$
in distribution as $n \to \infty$. Moreover,
$$\mathbb{E}\mathbf{M}_{n,p} \sim n\sqrt{\frac{\log n}{2(\log(1-p) - \log p)}} \quad \text{and} \quad \mathrm{Var}(\mathbf{M}_{n,p}) = \Theta\left(n \log(n)^{3/2}\right).$$

Theorem 3 (Degree distribution of the RCG)**.** *Consider the random cluster graph* $\mathbf{CG}_{n,p}$ *on* $n \in \mathbb{N}$ *vertices and ER edge probability* $p \in (0,1)$ *and the degree* $\mathbf{D}_{n,p}$ *of a uniformly chosen vertex.*

1. *If* $p > 1/2$*, then*
$$\lim_{n\to\infty} \mathbb{P}(\mathbf{D}_{n,p} = n-1) = 1.$$
2. *If* $p = 1/2$*, then for a Poisson random variable* X_n *with parameter* $\log n - \log\log n + o(1)$*, we have*
   (a) *for all* $z \in \mathbb{C}$*,*
   $$\mathbb{E} z^{\mathbf{D}_{n,1/2}} \sim \mathbb{E} z^{X_n}.$$
   That is, the probability generating function of $\mathbf{D}_{n,1/2}$ *and the one of* X_n *are asymptotically the same.*
   (b) *Additionally,*
   $$\lim_{n\to\infty} \mathrm{d}_{TV}(\mathbf{D}_{n,1/2}, X_n) = 0.$$
3. *If* $p < 1/2$*, then* $\mathbb{E}\mathbf{D}_{n,p} = \Theta(\sqrt{\log n})$*. Moreover, for each* $\lambda \in [0,1)$ *there exists a subsequence* $(n_k)_{k\in\mathbb{N}}$ *such that*
$$\mathbf{D}_{n_k,p} - \left\lfloor \sqrt{\frac{2\log n_k}{\log(1-p) - \log p}} - 1 - \frac{1}{\log(1-p) - \log p} \right\rfloor \longrightarrow X_\lambda$$
in distribution as $k \to \infty$*, where* X_λ *is defined by*
$$\mathbb{P}(X_\lambda = d) = \frac{\left(\frac{p}{1-p}\right)^{(d-\lambda)^2/2}}{\sum_{d'\in\mathbb{Z}} \left(\frac{p}{1-p}\right)^{(d'-\lambda)^2/2}}$$
for all $d \in \mathbb{Z}$*.*

Notice that the fact that $\mathbf{D}_{n,p} = \Theta(\sqrt{\log n}))$ when $p < 1/2$ follows directly from Theorem 2. However, to obtain the distribution full of $\mathbf{D}_{n,p}$ is technically quite involved.

3 Generating functions and analytic combinatorics

By conditioning $G(n,p)$ we loose the independence of the $G(n,p)$ model. To overcome this fact we use *counting* techniques. Let $\mathcal{F}$ be a class (invariant under isomorphims) of labelled graphs, and let $\mathcal{F}_{n,m}$ be the graphs in $\mathcal{F}$ with n vertices and m edges. We denote by $n(G)$ number of vertices of G, and by $m(G)$ the number of edges. The exponential generating function (EGF) associated to $\mathcal{F}$ is
$$F(w,z) = \sum_{G\in\mathcal{F}} w^{m(G)} \frac{z^{n(G)}}{n(G)!},$$
so that $|\mathcal{F}_{n,m}| = n![w^m z^n]F(w,z)$. In particular, the EGF of the class of non-empty cliques is
$$C(w,z) = \sum_{n\geq 1} w^{\binom{n}{2}} \frac{z^n}{n!}$$
From now on we use freely the symbolic method, as described in [2]. In particular, since a cluster graph is a *set* of cliques, its EGF is $\exp(uC(w,z))$, where the variable u marks components.

It is easy to see that the distribution of random cluster graphs is equal to
$$\mathbb{P}(\mathbf{CG}_{n,p} = G) = \frac{\left(\frac{p}{1-p}\right)^{m(G)}}{B_n(p/1-p)},$$

where the *partition function* $B_n(w)$ is given by $B_n(w) = n![z^n]e^{C(w,z)}$. We notice that $B_n(1)$ is the n-th Bell number, counting partitions of a set of size n. From here one easily obtains the probability generating functions (PGF) of the main parameters. Recall that the PGF of an integer-valued non-negative random variable X is defined as

$$\mathrm{PGF}_X(u) = \mathbb{E}(e^X) = \sum_{k\geq 0} \mathbb{P}(X = k)u^k.$$

Proposition 4. *Let $\mathbf{M}_{n,p}$, $\mathbf{C}_{n,p}$ and $\mathbf{D}_{n,p}$ as in Section 2. Set $B_n(w) = n![z^n]e^{C(w,z)}$ as before, and and $w = \frac{p}{1-p}$. The probability generating functions of these random variables are equal to*

$$\mathrm{PGF}_{\mathbf{M}_{n,p}}(u) = \frac{B_n(u\,w)}{B_n(w)},$$
$$\mathrm{PGF}_{\mathbf{C}_{n,p}}(u) = \frac{[z^n]e^{uC(w,z)}}{[z^n]e^{C(w,z)}},$$
$$\mathrm{PGF}_{\mathbf{D}_{n,p}}(u) = \frac{[z^n]C_1(w,u\,z)e^{C(w,z)}}{u[z^n]C_1(w,z)e^{C(w,z)}}.$$

In order to obtain limit theorems we use the moment generating function (alternatively, the characteristic function $\mathbb{E}(e^{itX})$)

$$\mathbb{E}(e^{tX}) = \mathrm{PGF}_X(e^t).$$

Our main tool is Levy's continuity theorem:

Theorem 5. *Let X_n and Y be real valued random variables. If $\mathbb{E}(e^{tX_n})$ converges pointwise for t in a neighborhood of 0 to $\mathbb{E}(e^{tY})$, then X_n converges in law to Y.*

In particular, if there exists μ_n and σ_n such that, pointwise for s in a neighborhood of 0

$$\mathrm{PGF}_{X_n}(e^{s/\sigma_n}) \sim e^{s\mu_n/\sigma_n}e^{s^2/2} \qquad \text{as } n \to \infty$$

then the renormalized random variables $X_n^\star = \frac{X_n - \mu_n}{\sigma_n}$ converges to the standard normal distribution.

In order to apply the previous result we need to estimate the corresponding PGFs as $n \to \infty$. This is not an easy task, due mainly to the quadratic exponent $\binom{n}{2}$ in the expression for $C(w,z)$. In fact, to compute moments, we need more generally to estimate the derivatives of $C(w,z)$ with respect to z. This is the most technical part of our work, involving Cauchy integrals, saddle-point methods, and the so-called Hayman admissible functions [2], among other tools.

We observe that the size of the largest block in the $p = 1/2$ regime is known to be $\Theta(\log n)$. When $p < 1/2$ it should be $\Theta(\sqrt{\log n})$ due to concentration, but we have not worked out the details.

4 Further results

In this final section, we collect further results on random cluster graphs.

The critical window when $p \downarrow \frac{1}{2}$. We know that when $p > 1/2$ the random cluster gaph $\mathbf{CG}_{n,p}$ is almost surely a single clique. If we let $p = p(n) > 1/2$, we are interested in the scale at which $\mathbf{CG}_{n,p}$ becomes a single clique.

Proposition 6. *Let $q \in (0,1)$ and $p_n(q)$ defined by*

$$\mathbb{P}(\mathbf{C}_{n,p_n(q)} = 1) = q.$$

Then

$$p_n(q) = \frac{1}{2} + \frac{\log(n)}{2n} + O\left(\frac{\log\log n}{n}\right).$$

Notice that the precise value of q is not important, in fact it only appears in the error term.

In addition, we show that there exists no 'almost complete' regime. For instance, for any sequence $p_n \in [0,1]$ we have

$$\mathbb{P}(\mathbf{C}_{n,p_n(q)} = K_{n-1} \cup K_1) \to 0, \quad \text{as } n \to \infty,$$

and similarly for $\mathbf{C}_{n,p_n(q)} = K_{n-r} \cup \{ \text{ small cliques } \}$, for fixed $r > 0$.

The upercritical regime ($p > \frac{1}{2}$). In this regime we know that there is only one clique w.h.p. Our next result is an asymptotic expansion for $\mathbb{P}(\mathbf{C}_{n,p} = K_n)$. First notice that if $w = \frac{p}{1-p} > 1$ then $C(w,z) = \sum_{n\geq 1} w^{\binom{n}{2}} \frac{z^n}{n!}$ has zero radius of convergence. Using recent tools for estimating coefficients of divergent series [1] we show that

Proposition 7.

$$\mathbb{P}(\mathbf{CG}_{n,p} = K_n) = 1 + \sum_{m=1}^{R-1} w^{-mn} P_m(n) + O\left(w^{-Rn} n^R\right)$$

where $P_m(n)$ are certain polynomials and $R \geq 0$ is an integer

The first terms in the expansion are $\mathbb{P}(\mathbf{CG}_{n,p} = K_n) = 1 - nw \cdot w^{-n} + O(n^2 w^{-2n})$.

The sparse regime $p \to 0$. We focus on the case where p_n decreases like a monomial $p_n = n^{-\alpha + o(1)}$ for $\alpha > 0$. We prove that in this regime, the degree distribution concentrates around one or two values. We first show how α should be chosen to concentrate this distribution around a particular degree d:

Theorem 8. *Let $d \in \mathbb{N} \cup \{0\}$ and consider a limiting sequence $p_n = n^{-\frac{2}{(d+1)^2} + o(1)}$. Then*

$$\mathbb{P}(\mathbf{D}_n = d) \to 1.$$

Furthermore, for any other $d' \in \mathbb{N} \cup \{n\}$, the degree distribution satisfies

$$\mathbb{P}(\mathbf{D}_n = d') = n^{-\left(\frac{d'-d}{d+1}\right)^2 + o(1)}. \tag{1}$$

In the field of random graphs, the case $p_n = \lambda/n$ is one of the most interesting regimes, known as the *sparse regime*. The next lemma shows that in this regime, the degree distribution is concentrated around two values, rather than one:

Proposition 9. *Let $\lambda > 0$ and consider the sequence $p_n \sim \lambda/n$, then*

$$\mathbb{P}(\mathbf{D}_n = 0) \to \frac{\sqrt{4\lambda + 1} - 1}{2\lambda}, \quad \mathbb{P}(\mathbf{D}_n = 1) \to 1 - \frac{\sqrt{4\lambda + 1} - 1}{2\lambda},$$

In particular, the sequence $p_n \sim 1/n$ yields $\mathbb{P}(\mathbf{D}_n = 0) \to \rho^{-1}$, where $\rho = \frac{\sqrt{5}+1}{2}$ is the golden ratio.

Conditioning to other classes of graphs. For fixed $p \in (0,1)$, let $F(n,p)$ the random graph $G(n,p)$ conditioned to be a forest. $F(n,p)$ behaves like a random uniform forest, in the sense that the number of edges is linear and asymptotically Gaussian, and the number of components is asymptotically Poisson distributed; only the constants depend on p and thereis no phase transition. The same is true conditioning on being planar, or related classes of graphs.

In order to get a situation like for random cluster graphs, we believe that one should need to condition on classes of graphs admitting superlinear number of edges.

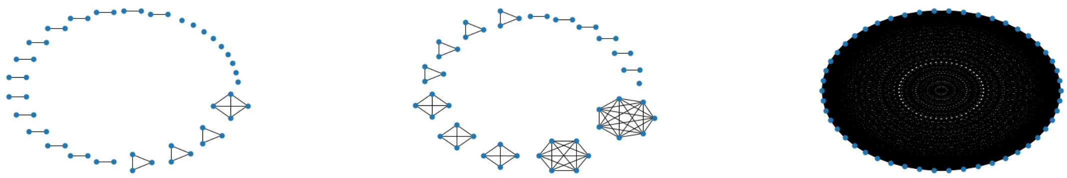
Sampling. How can we sample a random cluster graph $\mathbf{CG}_{n,p}$? Certainly not sampling with rejection, since the event $G(n, p)$ being a cluster graph is extremely rare. Instead we sample first the size of one clique and the rest by induction. Let $\mathbf{S}_{n,p}$ be the size of the clique containing vertex 1. Then we have

Proposition 10.

$$\mathbb{P}(\mathbf{S}_{n,p} = s) = \binom{n}{s-1}\left(\frac{p}{1-p}\right)^{\binom{s}{2}} \frac{B_{n-s}(p/(1-p))}{B_n(p/(1-p))},$$

where $B_n(w)$ is as in Section 3.

Once we sample the size s of the first clique according to the previous distribution, we can sample recursively on the remaining $n - s$ vertices. Below we show examples of this procedure for (from left to right) $p = 0.25, p = 0.51$ and $p = 0.53$.



Application to community detection. We come finally to the original motivation for our research. *Community detection* aims at partitioning the nodes of a network into *communities*: sets of vertices that are more strongly connected to each other than to the remainder of the network. A popular approach is to optimize a quantity known as *text*modularity over the set of partitions. A resolution parameter controls the granularity of the obtained clustering

Given a graph G and cluster graphCG representing a potential partition, and a resolution parameter γ, the modularity is defined as

$$M(G, CG, \gamma) = \frac{1}{m(G)}\left(m(G \cap CG) - \gamma \cdot m(CG)\right)$$

The main goal is to understand modularity better and how to choose γ. For that one can use $\mathbf{CG}_{n,p}$ as a model for a prior distribution. When the communities have sizes close to $\log n$, setting $p = 1/2$ will likely lead to detecting communities of the desired granularity. But when the communities are significantly smaller than $\log n$, one should choose $p > 1/2$. Preliminary inevstigations indicate that the choice of

$$p_n = \frac{1}{2} + \frac{\log n}{2n} + O\left(\frac{\log\log n}{n}\right)$$

leads to significantly better community-detection performance.

References

[1] S. Dovgal, K. Nurligareev. *Asymptotics for graphically divergent series: dense digraphs and 2-SAT formulae*, arXiv:2310.05282 (2023).

[2] P. Flajolet, R. Sedgewick. *Analytic Combinatorics*. Cambridge U. Press, 2009.

[3] S. Fortunato, M. Barthélemy. Resolution limit in community detection. *Proc. Nat. Acad. Sci.* 104.1 (2007), 36–41.

[4] V. N. Sachkov. *Probabilistic methods in combinatorial analysis*. Cambridge U. Press, 1997.