Probab. Theory Relat. Fields (2015) 162:431–461
DOI 10.1007/s00440-014-0576-6

Reconstruction and estimation in the planted partition model

Elchanan Mossel · Joe Neeman · Allan Sly

Received: 21 August 2012 / Revised: 28 June 2014 / Published online: 20 July 2014

Abstract The planted partition model (also known as the stochastic blockmodel) is a classical cluster-exhibiting random graph model that has been extensively studied in statistics, physics, and computer science. In its simplest form, the planted partition model is a model for random graphs on n nodes with two equal-sized clusters, with an between-class edge probability of q and a within-class edge probability of p. Although most of the literature on this model has focused on the case of increasing degrees (ie. $pn, qn \to \infty$ as $n \to \infty$), the sparse case $p, q = O(1/n)$ is interesting both from a mathematical and an applied point of view. A striking conjecture of Decelle, Krzkala, Moore and Zdeborová based on deep, non-rigorous ideas from statistical physics gave a precise prediction for the algorithmic threshold of clustering in the sparse planted partition model. In particular, if $p = a/n$ and $q = b/n$, then Decelle et al. conjectured that it is possible to cluster in a way correlated with the true partition if $(a-b)^2 > 2(a+b)$, and impossible if $(a-b)^2 < 2(a+b)$. By comparison, the best-known rigorous result is that of Coja-Oghlan, who showed that clustering is possible if $(a-b)^2 > C(a+b)$ for some sufficiently large C. We prove half of their prediction, showing that it is indeed impossible to cluster if $(a-b)^2 < 2(a+b)$. Furthermore we

Supported by NSF grants DMS-1106999, CCF-1320105 and DOD ONR grant N000141110140.

Supported by Sloan Research Fellowship in mathematics and by NSF award DMS-1208339.

E. Mossel · J. Neeman (✉) · A. Sly
Department of Statistics, UC Berkeley, Berkeley, USA
e-mail: joeneeman@gmail.com

E. Mossel
Department of Computer Science, UC Berkeley, Berkeley, USA

A. Sly
Department of Mathematics, Australian National University, Canberra, Australia

show that it is impossible even to estimate the model parameters from the graph when $(a-b)^2 < 2(a+b)$; on the other hand, we provide a simple and efficient algorithm for estimating a and b when $(a-b)^2 > 2(a+b)$. Following Decelle et al, our work establishes a rigorous connection between the clustering problem, spin-glass models on the Bethe lattice and the so called reconstruction problem. This connection points to fascinating applications and open problems.

Mathematics Subject Classification (2010) Primary 05C80; Secondary 60J85 · 90B15 · 91D30

1 Introduction

1.1 The planted partition problem

The clustering problem in its general form is, given a (possibly weighted) graph, to divide its vertices into several strongly connected classes with relatively weak cross-class connections. This problem is fundamental in modern statistics, machine learning and data mining, but its applications range from population genetics [29], where it is used to find genetically similar sub-populations, to image processing [33,36], where it can be used to segment images or to group similar images, to the study of social networks [28], where it is used to find strongly connected groups of like-minded people.

The algorithms used for clustering are nearly as diverse as their applications. On one side are the hierarchical clustering algorithms [22] which build a hierarchy of larger and larger communities, by either recursive aggregation or division. On the other hand model-based statistical methods, including the celebrated EM algorithm [10], are used to fit cluster-exhibiting statistical models to the data. A third group of methods work by optimizing some sort of cost function, for example by finding a minimum cut [16,33] or by maximizing the Girvan–Newman modularity [1,27].

Despite the variety of available clustering algorithms, the theory of clustering contains some fascinating and fundamental algorithmic challenges. For example, the "min-bisection" problem—which asks for the smallest graph cut dividing a graph into two equal-sized pieces—is well-known to be NP-hard [14]. Going back to the 1980s, there has been much study of the average-case complexity of the min-bisection problem. For instance, the min-bisection problem is much easier if the minimum bisection is substantially smaller than most other bisections. This has led to interest in random graph models for which a typical sample has exactly one good minimum bisection. Perhaps the simplest such model is the "planted bisection" model, which is similar to the Erdös–Renyi model.

Definition 1 (*The planted bisection model*) For $n \in \mathbb{N}$ and $p, q \in (0,1)$, let $\mathcal{G}(n,p,q)$ denote the model of random, ±-labelled graphs in which each vertex u is assigned (independently and uniformly at random) a label $\sigma_u \in \{\pm\}$, and then each possible edge (u,v) is included with probability p if $\sigma_u = \sigma_v$ and with probability q if $\sigma_u \neq \sigma_v$.

If $p = q$, the planted partition model is just an Erdös–Renyi model, but if $p \gg q$ then a typical graph will have two well-defined clusters. Actually, the literature on the min-bisection problem usually assumes that the two classes have exactly the same size (instead of a random size), but this modification makes almost no difference in the context of this work.

The planted bisection model was not the earliest model to be studied in the context of min-bisection—Bui et al. [6] and Boppana [5] considered graphs chosen uniformly at random from all graphs with a given number of edges and a small minimum bisection. Dyer and Frieze [11] were the first to study the min-bisection problem on the planted bisection model; they showed that if $p > q$ are fixed as $n \to \infty$ then the minimum bisection is the one that separates the two classes, and it can be found in expected $O(n^3)$ time.

The result of Dyer and Frieze was improved by Jerrum and Sorkin [21], who reduced the running time to $O(n^{2+\epsilon})$ and allowed $p - q$ to shrink at the rate $n^{-1/6+\epsilon}$. More interesting than these improvements, however, was the fact that Jerrum and Sorkin's analysis applied to the popular and fast-in-practice Metropolis algorithm. Later, Condon and Karp [8] gave better theoretical guarantees with a linear-time algorithm that works for $p - q \geq \Omega(n^{-1/2+\epsilon})$.

With the exception of Boppana's work (which was for a different model), the aforementioned results applied only to relatively dense graphs. McSherry [25] showed that a spectral clustering algorithm works as long as $p - q \geq \Omega(\sqrt{q(\log n)/n})$. In particular, his result is meaningful as long as p and q are at least $\Omega((\log n)/n)$. These are essentially the sparsest possible graphs for which the minimum cut will agree with the planted bisection, but Coja-Oghlan [7] managed to obtain a result for even sparser graphs by studying a relaxed problem. Instead of trying to recover the minimum bisection, he showed that a spectral algorithm will find a bisection which is positively correlated with the planted bisection. His result applies as long as $p - q \geq \Omega(\sqrt{q/n})$, and so it is applicable even when p and q are $O(1/n)$; that is, it is relevant even for graphs with a constant average degree.

1.2 Block models in statistics

The statistical literature on clustering is more closely focused on real-world network data, with the planted bisection model (or "stochastic blockmodel," as it is known in the statistics community) used as an important test-case for theoretical results. Its study goes back to Holland et al. [18], who discussed parameter estimation and gave a Bayesian method for finding a good bisection, without theoretical guarantees. Snijders and Nowicki [35] studied several different statistical methods—including maximum likelihood estimation and the EM algorithm—for the planted bisection model with $p - q = \Omega(1)$. They then applied those methods to social networks data. More recently, Bickel and Chen [1] showed that maximizing the Girvan–Newman modularity—a popular measure of cluster strength—recovers the correct bisection, for the same range of parameters as the result of McSherry. They also demonstrated that their methods perform well on social and telephone network data. Spectral clustering, the method studied by Boppana and McSherry, has also appeared in the statistics literature: Rohe et al. [32] gave a theoretical analysis of spectral clustering under the planted bisection model and also applied the method to data from Facebook.

1.3 Sparse graphs and insights from statistical physics

The case of sparse graphs with constant average degree is well motivated from the perspective of real networks. Indeed, Leskovec et al. [24] collected and studied a vast collection of large network datasets, ranging from social networks like LinkedIn and MSN Messenger, to collaboration networks in movies and on the arXiv, to biological networks in yeast. Many of these networks had millions of nodes, but most had an average degree of no more than 20; for instance, the LinkedIn network they studied had approximately seven million nodes, but only 30 million edges. Similarly, the real-world networks considered by Strogatz [37]—which include coauthorship networks, power transmission networks and web link networks—also had small average degrees. Thus it is natural to consider the planted partition model with parameters p and q of order $O(1/n)$.

Although sparse graphs are natural for modelling many large networks, the planted partition model seems to be most difficult to analyze in the sparse setting. Despite the large amount of work studying this model, the only results we know of that apply in the sparse case $p, q = O(\frac{1}{n})$ are those of Coja-Oghlan. Recently, Decelle et al. [9] made some fascinating conjectures for the cluster identification problem in the sparse planted partition model. In what follows, we will set $p = a/n$ and $q = b/n$ for some fixed $a > b > 0$.

Conjecture 1 ([9]) *If $(a-b)^2 > 2(a+b)$ then the clustering problem in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ is solvable as $n \to \infty$, in the sense that one can a.a.s. find a bisection which is positively correlated with the planted bisection.*

To put Coja-Oghlan's work into the context of this conjecture, he showed that if $(a-b)^2 > C(a+b)$ for a large enough constant C, then the spectral method solves the clustering problem. Decelle et al.'s work is based on deep but non-rigorous ideas from statistical physics. In order to identify the best bisection, they use the sum-product algorithm (also known as belief propagation). Using the cavity method, they argue that the algorithm should work, a claim that is bolstered by compelling simulation results.

What makes Conjecture 1 even more interesting is the fact that it might represent a threshold for the solvability of the clustering problem.

Conjecture 2 ([9]) *If $(a-b)^2 < 2(a+b)$ then the clustering in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ problem is not solvable as $n \to \infty$, in the sense that not even a computationally unbounded algorithm can find a partition whose correlation with the planted bisection is bounded away from zero, with probability bounded away from zero, as $n \to \infty$.*

This second conjecture is based on a connection with the tree reconstruction problem (see [26] for a survey). Consider a multi-type branching process where there are two types of particles named $+$ and $-$. Each particle gives birth to Pois(a) (ie. a Poisson distribution with mean a) particles of the same type and Pois(b) particles of the complementary type. In the tree reconstruction problem, the goal is to recover the label of the root of the tree from the labels of level r where $r \to \infty$. This problem goes back to Kesten and Stigum [23] in the 1960s, who showed that if $(a-b)^2 > 2(a+b)$ then it is possible to recover the root value with non-trivial probability. The converse was not resolved until 2000, when Evans et al. [12] proved that if $(a-b)^2 \le 2(a+b)$ then it is impossible to recover the root with probability bounded above $1/2$ independent of r. This is equivalent to the reconstruction or extremality threshold for the Ising model on a branching process.

At the intuitive level the connection between clustering and tree reconstruction, follows from the fact that the neighborhood of a vertex in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ should look like a random labelled tree with high probability. Moreover, the distribution of that labelled tree should converge as $n \to \infty$ to the multi-type branching process defined above. We will make this connection formal later.

Decelle et al. also made a conjecture related to the the parameter estimation problem that was previously studied extensively in the statistics literature. Here the problem is to identify the parameters a and b; note that if the blocks can be recovered exactly, then the parameters may easily be estimated simply by counting edges between different blocks; this was noted, for example, in [1]. Nevertheless, parameter estimation is interesting on its own, for instance because it is a first step for some clustering algorithms. Hence, [18] and [35] both discussed ML and Bayesian methods for parameter estimation, although neither work gave theoretical guarantees.

As in the clustering problem, Decelle et al. provided a parameter-estimation algorithm based on belief propagation and they used physical ideas to argue that there is a threshold above which the parameters can be estimated, and below which they cannot.

Conjecture 3 ([9]) *If $(a-b)^2 > 2(a+b)$ then there is a consistent estimator for a and b under $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$. Conversely, if $(a-b)^2 < 2(a+b)$ then there is no consistent estimator.*

2 Our results

Our main contribution is to establish Conjectures 2 and 3. Recall that $a, b > 0$ are fixed as $n \to \infty$, and let $\mathbb{P}_n$ denote the probability with respect to $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$.

Theorem 1 *If $(a-b)^2 \le 2(a+b)$ then, for any fixed vertices u and v,*

$$\mathbb{P}_n(\sigma_u = +|G, \sigma_v = +) \to \frac{1}{2} \ a.a.s.$$

Remark 1 Theorem 1 is stronger than Conjecture 2 because it says that an even easier problem cannot be solved: if we take two random vertices of G, Theorem 1 says that no algorithm can tell whether or not they have the same label. This is an easier task than finding a bisection, because finding a bisection is equivalent to labeling *all* the vertices; we are only asking whether two of them have the same label or not. Theorem 1 is also stronger than the conjecture because it includes the case $(a-b)^2 = 2(a+b)$, for which Decelle et al. did not conjecture any particular behavior.

To prove Conjecture 3, we compare the planted partition model to an appropriate Erdös–Renyi model: let $\mathbb{P}_n = \mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ and take $\mathbb{P}'_n = \mathcal{G}(n, \frac{a+b}{2n})$ to be the Erdös–Renyi model that has the same average degree as $\mathbb{P}_n$.

Theorem 2 *If $(a-b)^2 < 2(a+b)$ then $\mathbb{P}_n$ and $\mathbb{P}'_n$ are mutually contiguous i.e., for a sequence of events A_n, $\mathbb{P}_n(A_n) \to 0$ if, and only if, $\mathbb{P}'_n(A_n) \to 0$.*

Moreover, if $(a-b)^2 < 2(a+b)$ then there is no consistent estimator for a and b.

Note that the second part of the Theorem 2 follows from the first part, since it implies that $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ and $\mathcal{G}(n, \frac{\alpha}{n}, \frac{\beta}{n})$ are contiguous as long as $a+b = \alpha+\beta > \frac{1}{2}\max\{(a-b)^2, (\alpha-\beta)^2\}$. Indeed one cannot even consistently distinguish the planted partition model from the corresponding Erdös–Renyi model! The question of contiguity was previously studied by Janson [20] in a more general setting: he studied contiguity for vectors of conditionally independent Bernoulli variables. In our setting, however, Theorem 2 is much stronger than the results of [20], which imply contiguity if $|a-b| = O(n^{-1/2})$.

The other half of Conjecture 3 follows from a converse to Theorem 2:

Theorem 3 *If $(a-b)^2 > 2(a+b)$, then $\mathbb{P}_n$ and $\mathbb{P}'_n$ are asymptotically orthogonal. Moreover, a consistent estimator for a, b can be obtained as follows: let X_k be the number of cycles of length k, and define*

$$\hat{d}_n = \frac{2|E|}{n}$$
$$\hat{f}_n = (2k_n X_{k_n} - \hat{d}_n^{k_n})^{1/k_n}$$

where $k_n = \lfloor \log^{1/4} n \rfloor$. Then $\hat{d}_n + \hat{f}_n$ is a consistent estimator for a and $\hat{d}_n - \hat{f}_n$ is a consistent estimator for b.

Finally, there is an efficient algorithm whose running time is polynomial in n to calculate $\hat{d}_n$ and $\hat{f}_n$.

2.1 Proof techniques

2.1.1 Short cycles

To establish Theorem 3 we count the number of short cycles in $G \sim \mathbb{P}_n$. It is well-known that the number of k-cycles in a graph drawn from $\mathbb{P}'_n$ is approximately Poisson-distributed with mean $\frac{1}{k}(\frac{a+b}{2})^k$. The proof of this fact can be modified as in [4] to show a Poisson limit for cycle counts in more general inhomogeneous graphs. For completeness, we include the proof of our special case showing that the number of k-cycles in $\mathbb{P}_n$ is approximately Poisson-distributed with mean $\frac{1}{k}\big((\frac{a+b}{2})^k + (\frac{a-b}{2})^k\big)$.

By comparing the first and second moments of Poisson random variables and taking k to increase slowly with n, one can distinguish between the cycle counts of $G \sim \mathbb{P}_n$ and $G \sim \mathbb{P}'_n$ as long as $(a-b)^2 > 2(a+b)$.

The first half of Conjecture 3 follows because the same comparison of first and second moments implies that counting cycles gives a consistent estimator for $a+b$ and $a-b$ (and hence also for a and b).

While there is in general no efficient algorithm for counting cycles in graphs, we show that with high probability the number of short cycles coincides with the number of non-backtracking walks of the same length which can be computed efficiently using matrix multiplication.

The proof of Theorem 3 is carried out in Sect. 3.

2.1.2 Non-reconstruction

As mentioned earlier, Theorem 1 intuitively follows from the fact that the neighborhood of a vertex in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ should look like a random labelled tree with high probability and the distribution of that labelled tree should converge as $n \to \infty$ to the multi-type branching process defined above. While this intuition is not too hard to justify for small neighborhoods (by proving there are no short cycles etc.) the global ramifications are more challenging to establish. This is because, conditioned on the graph structure, the model is neither an Ising model, nor a Markov random field! This is due to two effects:

- The fact that the two clusters are of the same (approximate) size. This amounts to a global conditioning on the number of $+/-$'s.
- The model is not even a Markov random field conditioned on the number of $+$ and $-$ vertices. This follows from the fact that for every two vertices u, v that do not form an edge, there is a different weight for $\sigma_u = \sigma_v$ and $\sigma_u \neq \sigma_v$. In other words, if $a > b$, then there is a slight repulsion (anti-ferromagnetic interaction) between vertices not joined by an edge.

In Sect. 4, we prove Theorem 1 by showing how to overcome the challenges above.

2.1.3 The second moment

A major effort is devoted to the proof Theorem 2. In the proof we show that the random variables $\frac{\mathbb{P}_n(G)}{\mathbb{P}'_n(G)}$ don't have much mass near 0 or ∞. Since the margin of $\mathbb{P}_n$ is somewhat complicated to work with, the first step is to enrich the distribution $\mathbb{P}'_n$ by adding random labels. Then we show that the random variables $Y_n := \frac{\mathbb{P}_n(G,\sigma)}{\mathbb{P}'_n(G,\sigma)}$ don't have mass near 0 or ∞. We derive an explicit formula for the second moment of Y_n in Lemma 9. In particular we show that

$$\mathbb{E}Y_n^2 = (1 + o(1))\frac{e^{-t/2-t^2/4}}{\sqrt{1-t}}, \quad t = \frac{(a-b)^2}{2(a+b)}$$

This already shows that the second moment is bounded if $(a-b)^2 < 2(a+b)$. However, in order to establish the existence of a density, we also need to show that Y_n is bounded away from zero asymptotically. In order to establish this, we utilize the small graph conditioning method by calculating joint moments of the number of cycles and Y_n. It is quite surprising that this calculation can be carried out in rather elegant manner, since many other applications of this method are much more technically involved.

3 Counting cycles

The main result of this section is that the number of k-cycles of $G \sim \mathbb{P}_n$ is approximately Poisson-distributed. We will then use this fact to show the first part of Theorem 2. The cycle counting result that we present is actually a special case of a result by Bollobás et al. [4], who show a Poisson limit for cycle counts in more general inhomogeneous graphs, which have continuous edge labels and a kernel that defines edge probabilities; our special case is recovered by taking the obvious two-valued kernel in [4, Theorem 17.1]. For completeness, we include the proof of our special case.

Actually, Theorem 2 only requires us to calculate the first two moments of the number of k-cycles, but the rest of the moments require essentially no extra work. Indeed, once we compute the first moment, the others will follow by appealing to a classical result of Bollobás [3].

Theorem 4 *Let $X_{k,n}$ be the number of k-cycles of G, where $G \sim \mathbb{P}_n$. If $k = O(\log^{1/4}(n))$ then*

$$X_{k,n} \xrightarrow{d} \mathrm{Pois}\left(\frac{1}{k2^{k+1}}\big((a+b)^k + (a-b)^k\big)\right).$$

Before we prove this, let us explain how it implies Theorem 3. From now on, we will write X_k instead of $X_{k,n}$.

Proof of Theorem 3 We start by proving the first statement of the theorem. Let's recall the standard fact (which we have mentioned before) that under $\mathbb{P}'_n$, $X_k \xrightarrow{d} \mathrm{Pois}\left(\frac{(a+b)^k}{k2^{k+1}}\right)$. With this and Theorem 4 in mind,

$$\mathbb{E}_{\mathbb{P}} X_k, \mathrm{Var}_{\mathbb{P}}\, X_k \to \frac{(a+b)^k + (a-b)^k}{k2^{k+1}}$$
$$\mathbb{E}_{\mathbb{P}'} X_k, \mathrm{Var}_{\mathbb{P}'}\, X_k \to \frac{(a+b)^k}{k2^{k+1}}.$$

Set $k = k(n) = \log^{1/4} n$ (although any sufficiently slowly increasing function of n would do). Choose ρ such that $\frac{a-b}{2} > \rho > \sqrt{\frac{a+b}{2}}$. Then $\mathrm{Var}_{\mathbb{P}}\, X_k$ and $\mathrm{Var}_{\mathbb{P}'}\, X_k$ are both $o(\rho^{2k})$ as $k \to \infty$. By Chebyshev's inequality, $X_k \le \mathbb{E}_{\mathbb{P}'} X_k + \rho^k$ $\mathbb{P}'$-a.a.s. and $X_k \ge \mathbb{E}_{\mathbb{P}} X_k - \rho^k$ $\mathbb{P}$-a.a.s. Since $\mathbb{E}_{\mathbb{P}} X_k - \mathbb{E}_{\mathbb{P}'} X_k = \frac{1}{2k}(\frac{a-b}{2})^k = \omega(\rho^k)$, it follows that $\mathbb{E}_{\mathbb{P}} X_k - \rho^k \ge \mathbb{E}_{\mathbb{P}'} X_k + \rho^k$ for large enough k. And so, if we set $A_n = \{X_{k(n)} \le \mathbb{E}_{\mathbb{P}'} X_{k(n)} + \rho^k\}$ then $\mathbb{P}'(A_n) \to 1$ and $\mathbb{P}(A_n) \to 0$.

We next show that Theorem 4 gives us an estimator for a and b that is consistent when $(a-b)^2 > 2(a+b)$. First of all, we have a consistent estimator $\hat{d}$ for $d := (a+b)/2$ by simply counting the number of edges. Thus, if we can estimate $f := (a-b)/2$ consistently then we can do the same for a and b. Our estimator for f is

$$\hat{f} = (2kX_k - \hat{d}^k)^{1/k},$$

where $\hat{d}$ is some estimator with $\hat{d} \to d$ $\mathbb{P}$-a.a.s. and $k = k(n)$ increases to infinity slowly enough so that $k(n) = o(\log^{1/4} n)$ and $\hat{d}^k - d^k \to 0$ $\mathbb{P}$-a.a.s. Take $\sqrt{\frac{a+b}{2}} < \rho < \frac{a-b}{2} = f$; by Chebyshev's inequality, $2kX_k - d^k \in [f^k - \rho^k, f^k + \rho^k]$ $\mathbb{P}$-a.a.s. Since $k = k(n) \to \infty$, $\rho^k = o(f^k)$. Thus, $2kX_k - d^k = (1 + o(1))f^k$ $\mathbb{P}$-a.a.s. Since $\hat{d}^k - d^k \to 0$ and $f > 1$, $2kX_k - \hat{d}^k = f^k + o(1) = (1 + o(1))f^k$ $\mathbb{P}$-a.a.s. and so $\hat{f}$ is a consistent estimator for f. Finally we take $\hat{a} = \hat{d} + \hat{f}$ and $\hat{b} = \hat{d} - \hat{f}$. □

We observe that the estimator in the preceeding proof can be computed in almost linear time (i.e. in time $O(n^{1+\epsilon})$ for any $\epsilon > 0$).

Proposition 1 *The estimator of the previous proof can be computed in expected time* $O(n(a + b)^k)$.

Proof Recall $\hat{f}$ and $\hat{d}$ from the proof of Theorem 3. Clearly, we can compute $\hat{d}$ in time which is linear in the number of edges. Thus, we need to show how to find X_k efficiently. The essential idea is to count non-backtracking loops instead of cycles; we say that a path $v_1, v_2, \ldots, v_k$ is a non-backtracking loop if $v_1 = v_k$ and $v_i \neq v_{i-2}$ for all i. Indeed, a straightforward first moment argument shows that with high probability, each neighborhood of radius $2k(n)$ contains at most one cycle. On this event, every non-backtracking loop of length k is either a k-cycle, or it is an m-cycle that was traversed m/k times for some m dividing k. Thus, if we can compute, for every m dividing k, the number of non-backtracking loops of length m then we can also compute the number of cycles of length k.

To count the number of non-backtracking loops of length m, note that we can recursively count non-backtracking paths as follows: if $N^m_{u,v}$ is the number of non-backtracking paths from u to v and d_v is the degree of v then $N^{m+1}_{u,v} = \sum_{w \sim v} N^m_{u,w} - (d_v - 1)N^{m-1}_{u,v}$. Now, with high probability there are at most $n(a+b)^m$ choices of u, v with $N^m_{u,v} \neq 0$ (this may be checked by a simple first moment argument, because for any u, $\mathbb{E}\#\{v : N^m_{u,v} \neq 0\} \leq ((a + b)/2)^m\}$); hence $(N^m_{u,v})_{u,v \in V(G)}$ may be computed in time $O(n(a + b)^m)$. □

We remark that a slight adaptation of the preceding two proofs would show that if one is only interested in a constant accuracy, instead of asymptotic consistency, then it suffices to take k to be a sufficiently large constant, and the running time will be linear in n.

Now we will prove Theorem 4 using the method of moments. Recall, therefore, that if $Y \sim \text{Pois}(\lambda)$ then $\mathbb{E}Y_{[m]} = \lambda^m$, where $Y_{[m]}$ denotes the falling factorial $Y(Y - 1) \cdots (Y - m + 1)$. It will therefore be our goal to show that $\mathbb{E}(X_k)_{[m]} \to \left(\frac{(a+b)^k + (a-b)^k}{k2^{k+1}}\right)^m$. It turns out that this follows almost entirely from the corresponding proof for the Erdös–Renyi model. The only additional work we need to do is in the case $m = 1$.

Lemma 1 *If* $k = o(\sqrt{n})$ *then*

$$\mathbb{E}_{\mathbb{P}} X_k = \binom{n}{k} \frac{(k-1)!}{2} (2n)^{-k} \Big((a+b)^k + (a-b)^k\Big) \sim \frac{1}{k2^{k+1}} \Big((a+b)^k + (a-b)^k\Big).$$

Proof Let $v_0, \dots, v_{k-1}$ be distinct vertices. Let Y be the indicator that $v_0 \dots v_{k-1}$ is a cycle in G. Then $\mathbb{E}_{\mathbb{P}} X_k = \binom{n}{k} \frac{(k-1)!}{2} \mathbb{E}_{\mathbb{P}} Y$, so let us compute $\mathbb{E}_{\mathbb{P}} Y$. Define N to be the number of times in the cycle $v_1 \dots v_k$ that $\sigma_{v_i} \neq \sigma_{v_{i+1}}$ (with addition taken modulo k). Then

$$\mathbb{E}_{\mathbb{P}} Y = \sum_{m=0}^{k} \mathbb{P}(N = m) \mathbb{P}((v_1 \cdots v_k) \in G | N = m) = n^{-k} \sum_{m=0}^{k} \mathbb{P}(N = m) a^{k-m} b^m.$$

On the other hand, we can easily compute $P(N = m)$: for each $i = 0, \dots, k-2$, there is probability $\frac{1}{2}$ to have $\sigma_{v_i} = \sigma_{v_{i+1}}$, and these events are mutually indepedent. But whether $\sigma_{v_{k-1}} = \sigma_{v_0}$ is completely determined by the other events since there must be an even number of $i \in \{0, \dots, k-1\}$ such that $\sigma_{v_i} \neq \sigma_{v_{i+1}}$. Thus,

$$\begin{aligned} \mathbb{P}(N = m) &= \Pr\Big(\text{Binom}\left(k-1, \frac{1}{2}\right) \in \{m-1, m\} \Big) \\ &= 2^{-k+1} \left(\binom{k-1}{m-1} + \binom{k-1}{m} \right) = 2^{-k+1} \binom{k}{m} \end{aligned}$$

for even m, and zero for odd m. Hence,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}} Y &= n^{-k} 2^{-k+1} \sum_{m \text{ even}} a^{k-m} b^m \binom{k}{m} \\ &= n^{-k} 2^{-k} \Big((a+b)^k + (a-b)^k \Big). \end{aligned}$$

The second part of the claim amounts to saying that $n_{[k]} \sim n^k$, which is trivial when $k = o(\sqrt{n})$. □

Proof of Theorem 4 Let $\mu = \frac{1}{k 2^k} \big((a+b)^k + (a-b)^k \big)$; our goal, as discussed before Lemma 1, is to show that $\mathbb{E}(X_k)_{[m]} \to \mu^m$. Note that $(X_k)_{[m]}$ is the number of ordered m-tuples of k-cycles in G. We will divide these m-tuples into two sets: A is the set of m-tuples for which all of the k-cycles are disjoint, while B is the set of m-tuples in which at least one pair of cycles is not disjoint.

Now, take $(C_1, \dots, C_m) \in A$. Since the C_i are disjoint, they appear independently in G. By the proof of Lemma 1, the probability that cycles $C_1, \dots, C_m$ are all present is

$$n^{-km} 2^{-km} \Big((a+b)^k + (a-b)^k \Big)^m.$$

Since there are $\binom{n}{km} \frac{(km)!}{k^m}$ elements of A, it follows that the expected number of vertex-disjoint m-tuples of k-cycles is

$$\binom{n}{km} \frac{(km)!}{k^m} n^{-km} 2^{-km} \Big((a+b)^k + (a-b)^k \Big)^m \sim \mu^m.$$

It remains to show, therefore, that the expected number of non-vertex-disjoint m-tuples converges to zero. Let Y be the number of non-vertex-disjoint m-tuples,

$$Y = \sum_{(C_1,\dots,C_m)\in B} \prod_{i=1}^{m} 1_{\{C_i\subset G\}}.$$

Then the distribution of Y under $\mathbb{P}$ is stochastically dominated by the distribution of Y under the Erdös–Renyi model $\mathcal{G}(n, \frac{\max\{a,b\}}{n})$. It's well-known (see, eg. [3], Chapter 4) that as long as $k = O(\log^{1/4} n)$, $\mathbb{E}Y \to 0$ under $\mathcal{G}(n, \frac{c}{n})$ for any c; hence $\mathbb{E}Y \to 0$ under $\mathbb{P}$ also. □

4 Non-reconstruction

The goal of this section is to prove Theorem 1. As we said in the introduction, the proof of Theorem 1 uses a connection between $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ and Markov processes on trees. Before we go any further, therefore, we should define a Markov process on a tree and state the result that we will use.

Let T be an infinite rooted tree with root ρ. Given a number $0 \le \epsilon < 1$, we will define a random labelling $\tau \in \{\pm\}^T$. First, we draw τ_ρ uniformly in $\{\pm\}$. Then, conditionally independently given τ_ρ, we take every child u of ρ and set $\tau_u = \tau_\rho$ with probability $1-\epsilon$ and $\tau_u = -\tau_\rho$ otherwise. We can continue this construction recursively to obtain a labelling τ for which every vertex, independently, has probability $1-\epsilon$ of having the same label as its parent.

Back in 1966, Kesten and Stigum [23] asked (although they used somewhat different terminology) whether the label of ρ could be deduced from the labels of vertices at level R of the tree (where R is very large). There are many equivalent ways of stating the question. The interested reader should see the survey [26], because we will only mention two of them.

Let $T_R = \{u \in T : d(u,\rho) \le R\}$ and define $\partial T_R = \{u \in T : d(u,\rho) = R\}$. We will write τ_{T_R} for the configuration τ restricted to T_R.

Theorem 5 *Suppose T is a Galton–Watson tree where the offspring distribution has mean $d > 1$. Then*

$$\lim_{R\to\infty} Pr(\tau_\rho = + | \tau_{\partial T_R}) = \frac{1}{2} \ a.s.$$

if, and only if $d(1-2\epsilon)^2 \le 1$.

In particular, if $d(1-2\epsilon)^2 \le 1$ then $\tau_{\partial T_R}$ contains no information about τ_ρ. Theorem 5 was established by several authors over the course of more than 30 years. The non-reconstruction regime (ie. the case $d(1-2\epsilon)^2 \le 1$) is the harder one, and that part of Theorem 5 was first proved for d-ary trees in [2], and for Galton–Watson trees in [12]. This latter work actually proves the result for more general trees in terms of their branching number.

We will be interested in trees T whose offspring distribution is $\mathrm{Pois}(\frac{a+b}{2})$ and we will take $1-\epsilon = \frac{a}{a+b}$. Some simple arithmetic applied to Theorem 5 then shows that reconstruction of the root's label is impossible whenever $(a-b)^2 \leq 2(a+b)$. Not coincidentally, this is the same threshold that appears in Theorem 1.

4.1 Coupling of balls in G to the broadcast process on trees

The first step in applying Theorem 5 to our problem is to observe that a neighborhood of $(G, \sigma) \sim \mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ looks like (T, τ). Indeed, fix $\rho \in G$ and let G_R be the induced subgraph on $\{u \in G : d(u, \rho) \leq R\}$.

Proposition 2 *Let $R = R(n) = \lfloor \frac{1}{10 \log(2(a+b))} \log n \rfloor$. There exists a coupling between (G, σ) and (T, τ) such that $(G_R, \sigma_{G_R}) = (T_R, \tau_{T_R})$ a.a.s.*

For the rest of this section, we will take $R = \lfloor \frac{1}{10 \log(2(a+b))} \log n \rfloor$.

The proof of this proposition essentially follows from the fact that (T, τ) can be constructed from a sequence of independent Poisson variables, while (G_R, σ_{G_R}) can be constructed from a sequence of binomial variables, with approximately the same means. This argument is therefore quite similar to the analogous argument for $\mathcal{G}(n, \frac{a}{n})$; readers who are already familiar with this sort of argument may therefore wish to skip to Sect. 4.2.

For a vertex $v \in T$, let Y_v be the number of children of v; let $Y_v^=$ be the number of children whose label is τ_v and let $Y_v^{\neq} = Y_v - Y_v^=$. By Poisson thinning, $Y_v^= \sim \mathrm{Pois}(a/2)$, $Y_v^{\neq} \sim \mathrm{Pois}(b/2)$ and they are independent. Note that (T, τ) can be entirely reconstructed from the label of the root and the two sequences $(Y_i^=)$, $(Y_i^{\neq})$.

We can almost do the same thing for G_R, but it is a little more complicated. We will write $V = V(G)$ and $V_R = V(G) \backslash V(G_R)$. For every subset $W \subset V$, denote by W^+ and W^- the subsets of W that have the corresponding label. For example, $V_R^+ = \{v \in V_R : \sigma_v = +\}$. For a vertex $v \in \partial G_R$, let X_v be the number of neighbors that v has in V_r; then let $X_v^=$ be the number of those neighbors whose label is σ_v and set $X_v^{\neq} = X_v - X_v^=$. Then $X_v^= \sim \mathrm{Binom}(|V_r^{\sigma_v}|, a)$, $X_v^{\neq} \sim \mathrm{Binom}(|V_r^{-\sigma_v}|, b)$ and they are independent. Note, however, that they do not contain enough information to reconstruct G_R: it's possible to have $u, v \in \partial G_r$ which share a child in V_r, but this cannot be determined from X_u and X_v. Fortunately, such events are very rare and so we can exclude them. In fact, this process of carefully excluding bad events is all that needs to be done to prove Proposition 2.

In order that we can exclude their complements, let us give names to all of our good events. For any r, let A_r be the event that no vertex in V_{r-1} has more than one neighbor in G_{r-1}. Let B_r be the event that there are no edges within ∂G_r. Clearly, if A_r and B_r hold for all $r = 1, \ldots, R$ then G_R is a tree. In fact, it's easy to see that A_r and B_r are the only events that prevent $\{X_v^=, X_v^{\neq}\}_{v \in G}$ from determining (G_R, σ_{G_R}).

Lemma 2 *If*

1. *$(T_{r-1}, \tau_{T_{r-1}}) = (G_{r-1}, \sigma_{G_{r-1}})$;*
2. *$X_u^= = Y_u^=$ and $X_u^{\neq} = Y_u^{\neq}$ for every $u \in \partial G_{r-1}$; and*

3. *A_r and B_r hold*

then $(T_r, \tau_{T_r}) = (G_r, \sigma_{G_r})$.

Proof The proof is essentially obvious from the construction of X_u and Y_u, but we will be pedantic about it anyway. The statement $(T_{r-1}, \tau_{T_{r-1}}) = (G_{r-1}\sigma_{G_{r-1}})$ means that there is some graph homomorphism $\phi : G_{r-1} \to T_{r-1}$ such that $\sigma_u = \tau_{\phi(u)}$. If $u \in \partial G_{r-1}$ and $X_u^{=} = Y_{\phi(u)}^{=}$ and $X_u^{\neq} = Y_{\phi(u)}^{\neq}$ then we can extend ϕ to $G_{r-1} \cup \mathcal{N}(u)$ while preserving the fact that $\sigma_v = \tau_{\phi(v)}$ for all v. On the event A_r, this extension can be made simultaneously for all $u \in \partial G_{r-1}$, while the event B_r ensures that this extension remains a homomorphism. Thus, we have constructed a label-preserving homomorphism from (G_r, σ_{G_r}) to (T_r, τ_{T_r}), which is the same as saying that these two labelled graphs are equal.

From now on, we will not mention homomorphisms; we will just identify u with $\phi(u)$. □

In order to complete our coupling, we need to identify one more kind of good event. Let C_r be the event

$$C_r = \{|\partial G_s| \leq 2^s (a+b)^s \log n \text{ for all } s \leq r+1\}.$$

The events C_r are useful because they guarantee that V_r is large enough for the desired binomial-Poisson approximation to hold. The utility of C_r is demonstrated by the next two lemmas.

Lemma 3 *For all $r \leq R$,*

$$\mathbb{P}(C_r | C_{r-1}, \sigma) \geq 1 - n^{-\log(4/e)}.$$

Moreover, $|G_r| = O(n^{1/8})$ on C_{r-1}.

Lemma 4 *For any r,*

$$\mathbb{P}(A_r | C_{r-1}, \sigma) \geq 1 - O(n^{-3/4})$$
$$\mathbb{P}(B_r | C_{r-1}, \sigma) \geq 1 - O(n^{-3/4}).$$

Proof of Lemma 3 First of all, X_v is stochastically dominated by $\mathrm{Binom}(n, \frac{a+b}{n})$ for any v. On C_{r-1}, $|\partial G_r| \leq 2^r (a+b)^r \log n$ and so $|\partial G_{r+1}|$ is stochastically dominated by

$$Z \sim \mathrm{Binom}\Big(2^r (a+b)^r n \log n, \frac{a+b}{n}\Big).$$

Thus,

$$\begin{aligned}\mathbb{P}(\neg C_r | C_{r-1}, \sigma) &= \mathbb{P}\big(|\partial G_{r+1}| > 2^{r+1}(a+b)^{r+1} \log n \big| C_{r-1}, \sigma\big) \\ &\leq \mathbb{P}(Z \geq 2\mathbb{E}Z) \leq \Big(\frac{e}{4}\Big)^{\mathbb{E}Z}\end{aligned}$$

by a multiplicative version of Chernoff's inequality. But

$$\mathbb{E}Z = 2^r(a+b)^{r+1}\log n \geq \log n,$$

which proves the first part of the lemma.

For the second part, on C_{r-1}

$$|G_r| = \sum_{r=1}^{R} |\partial G_r| \leq \sum_{r=1}^{R} 2^r(a+b)^r \log n \leq (2(a+b))^{R+1}\log n = O(n^{1/8}).$$

□

Proof of Lemma 4 For the first claim, fix $u, v \in \partial G_r$. For any $w \in V_r$, the probability that (u, w) and (v, w) both appear is $O(n^{-2})$. Now, $|V_r| \leq n$ and Lemma 3 implies that $|\partial G_r|^2 = O(n^{1/4})$. Hence the result follows from a union bound over all triples u, v, w.

For the second part, the probability of having an edge between any particular $u, v \in \partial G_r$ is $O(n^{-1})$. Lemma 3 implies that $|\partial G_r|^2 = O(n^{1/4})$ and so the result follows from a union bound over all pairs u, v. □

The final ingredient we need is a bound on the total variation distance between binomial and Poisson random variables.

Lemma 5 *If m and n are positive integers then*

$$\left\| \operatorname{Binom}\left(m, \frac{c}{n}\right) - \operatorname{Pois}(c)\right\|_{TV} = O\Big(\frac{\max\{1, |m-n|\}}{n}\Big).$$

Proof of Lemma 4 Assume that $m \leq 2n$, or else the result is trivial. A classical result of Hodges and Le Cam [17] shows that

$$\left\| \operatorname{Binom}\left(m, \frac{c}{n}\right) - \operatorname{Pois}\left(\frac{mc}{n}\right)\right\|_{TV} \leq \frac{c^2 m}{n^2} = O(n^{-1}).$$

With the triangle inequality in mind, we need only show that $\operatorname{Pois}(cm/n)$ is close to $\operatorname{Pois}(c)$. This follows from a direct computation: if $\lambda < \mu$ then $\| \operatorname{Pois}(\lambda) - \operatorname{Pois}(\mu) \|_{TV}$ is just

$$\sum_{k\geq 0} \frac{|e^{-\mu}\mu^k - e^{-\lambda}\lambda^k|}{k!} \leq |e^{-\mu} - e^{-\lambda}| \sum_{k\geq 0} \frac{\mu^k}{k!} + e^{-\lambda} \sum_{k\geq 0} \frac{|\mu^k - \lambda^k|}{k!}.$$

Now the first term is $e^{\mu-\lambda} - 1$ and we can bound $\mu^k - \lambda^k \leq k(\mu-\lambda)\mu^{k-1}$ by the mean value theorem. Thus,

$$\| \operatorname{Pois}(\lambda) - \operatorname{Pois}(\mu) \|_{TV} \leq e^{\mu-\lambda} - 1 + e^{\mu-\lambda}(\mu - \lambda) = O(\mu - \lambda).$$

The claim follows from setting $\mu = c$ and $\lambda = \frac{cm}{n}$. □

Finally, we are ready to prove Proposition 2.

Proof of Proposition 2 Let $\tilde{\Omega}$ be the event that $\Big| |V^+| - |V^-| \Big| \leq n^{3/4}$. By Hoeffding's inequality, $\mathbb{P}(\tilde{\Omega}) \to 1$ exponentially fast.

Fix r and suppose that C_{r-1} and $\tilde{\Omega}$ hold, and that $(T_r, \tau_r) = (G_r, \sigma_r)$. Then for each $u \in \partial G_r$, $X_u^=$ is distributed as $\mathrm{Binom}(|V_r^{\sigma_u}|, a/n)$. Now,

$$\frac{n}{2} + n^{3/4} \geq |V^{\sigma_u}| \geq |V_r^{\sigma_u}| \geq |V^{\sigma_u}| - |G_{r-1}| \geq \frac{n}{2} - n^{3/4} - O(n^{1/8})$$

and so Lemma 5 implies that we can couple $X_u^=$ with $Y_u^=$ such that $\mathbb{P}(X_u^= \neq Y_u^=) = O(n^{-1/4})$ (and similarly for $X_u^{\neq}$ and $Y_u^{\neq}$). Since $|\partial G_{r-1}| = O(n^{1/8})$ by Lemma 3, the union bound implies that we can find a coupling such that with probability at least $1 - O(n^{-1/8})$, $X_u^= = Y_u^=$ and $X_u^{\neq} = Y_u^{\neq}$ for every $u \in \partial G_{r-1}$. Moreover, Lemmas 3 and 4 imply A_r, B_r and C_r hold simultaneously with probability at least $1 - n^{-\log(4/e)} - O(n^{-3/4})$. Putting these all together, we see that the hypothesis of Lemma 2 holds with probability at least $1 - O(n^{-1/8})$. Thus,

$$\mathbb{P}\Big((G_{r+1}, \sigma_{r+1}) = (T_{r+1}, \tau_{r+1}), C_r \Big| (G_r, \sigma_r) = (T_r, \tau_r), C_{r-1}\Big) \geq 1 - O(n^{-1/8}).$$

But $\mathbb{P}(C_0) = 1$ and we can certainly couple (G_1, σ_1) with (T_1, τ_1). Therefore, with a union bound over $r = 1, \ldots, R$, we see that $(G_R, \sigma_R) = (T_R, \tau_R)$ a.a.s. □

4.2 No long range correlations in G

We have shown that a neighborhood in G looks like a Galton–Watson tree with a Markov process on it. In this section, we will apply this fact to prove Theorem 1. In the statement of Theorem 1, we claimed that $\mathbb{E}(\sigma_\rho | G, \sigma_v) \to 0$, but this is clearly equivalent to $\mathrm{Var}(\sigma_\rho | G, \sigma_v) \to 1$. This latter statement is the one that we will prove, because the conditional variance has a nice monotonicity property.

The idea behind the proof of Theorem 1 is to condition on the labels of ∂G_R, which can only make reconstruction easier. Then we can remove the conditioning on σ_v, because $\sigma_{\partial G_R}$ gives much more information anyway. Since Theorem 5 and Proposition 2 imply that σ_v cannot be reconstructed from $\sigma_{\partial G_R}$, we conclude that it cannot be reconstructed from σ_v either.

The goal of this section is to prove that once we have conditioned on $\sigma_{\partial G_R}$, we can remove the conditioning on σ_v. If $\sigma | G$ were distributed according to a Markov random field, this would be trivial because conditioning on $\sigma_{\partial G_R}$ would turn σ_v and σ_ρ independent. For our model, unfortunately, there are weak long-range interactions. However, these interactions are sufficiently weak that we can get an asymptotic independence result for separated sets as long as one of them takes up most of the graph.

In what follows, we say that $X = o(a(n))$ a.a.s. if for every $\epsilon > 0$, $\Pr(|X| \geq \epsilon a(n)) \to 0$ as $n \to \infty$, and we say that $X = O(a(n))$ a.a.s. if

$$\limsup_{K \to \infty} \limsup_{n \to \infty} \Pr(|X| \geq K a(n)) = 0.$$

Lemma 6 *Let $A = A(G)$, $B = B(G)$, $C = C(G) \subset V$ be a (random) partition of V such that B separates A and C in G. If $|A \cup B| = o(\sqrt{n})$ for a.a.e. G*

$$\mathbb{P}(\sigma_A | \sigma_{B \cup C}, G) = (1 + o(1))\mathbb{P}(\sigma_A | \sigma_B, G)$$

for a.a.e. G and σ.

Note that Lemma 6 is only true for a.a.e. σ. In particular, the lemma does not hold for σ that are very unbalanced (eg. $\sigma = +^V$).

Before proving Lemma 6, let us show how it and Proposition 2 imply Theorem 1.

Proof of Theorem 1 By the monotonicity of conditional variances,

$$\mathrm{Var}(\sigma_\rho | G, \sigma_v, \sigma_{\partial G_R}) \leq \mathrm{Var}(\sigma_\rho | G, \sigma_v).$$

Since $|G_R| = o(\sqrt{n})$ a.a.s. and $v \not\in G_R$ a.a.s, it follows from Lemma 6 that σ_v and σ_ρ are a.a.s. conditionally independent given $\sigma_{\partial G_R}$ and G. Thus, $\mathrm{Var}(\sigma_\rho | G, \sigma_v, \sigma_{\partial G_R}) \to \mathrm{Var}(\sigma_\rho | G, \sigma_{\partial G_R})$. Finally, Proposition 2 implies that

$$|\,\mathrm{Var}(\sigma_\rho | G, \sigma_{\partial G_R}) - \mathrm{Var}(\tau_\rho | T, \tau_{\partial T_R})| \to 0;$$

since Theorem 5 says that $\mathrm{Var}(\tau_\rho | T, \tau_{\partial T_R})$ converges to 1 a.a.s., it follows that $\mathrm{Var}(\sigma_\rho | G, \sigma_{\partial G_R}) \to 1$ a.a.s. also. □

Proof of Lemma 6 As in the analogous proof for a Markov random field, we factorize $\mathbb{P}(G, \sigma)$ into parts depending on A, B and C. We then show that the part which measures the interaction between A and C is negligible. The rest of the proof is then quite similar to the Markov random fields case.

Define

$$\psi_{uv}(G, \sigma) = \begin{cases} \frac{a}{n} & \text{if } (u, v) \in E(G) \text{ and } \sigma_u = \sigma_v \\ \frac{b}{n} & \text{if } (u, v) \in E(G) \text{ and } \sigma_u \neq \sigma_v \\ 1 - \frac{a}{n} & \text{if } (u, v) \not\in E(G) \text{ and } \sigma_u = \sigma_v \\ 1 - \frac{b}{n} & \text{if } (u, v) \not\in E(G) \text{ and } \sigma_u \neq \sigma_v. \end{cases}$$

For arbitrary subsets $U_1, U_2 \subset V$, define

$$Q_{U_1,U_2} = Q_{U_1,U_2}(G, \sigma) = \prod_{u \in U_1, v \in U_2} \psi_{uv}(G, \sigma).$$

(If U_1 and U_2 overlap, the product ranges over all unordered pairs (u, v) with $u \neq v$; that is, if (u, v) is in the product then (v, u) is not.) Then

$$2^n \mathbb{P}(G, \sigma) = \mathbb{P}(G | \sigma) = Q_{A \cup B, A \cup B} Q_{B \cup C, C} Q_{A,C}. \tag{1}$$

First, we will show that $Q_{A,C}$ is essentially independent of σ. Take a deterministic sequence α_n with $\alpha_n/\sqrt{n} \to \infty$ but $\alpha_n|A| = o(n)$ a.a.s. Define $s_A(\sigma) = \sum_{v\in A}\sigma_v$ and $s_C(\sigma) = \sum_{v\in C}\sigma_v$ and let

$$\begin{aligned}\Omega &= \{\tau \in \{\pm\}^V : |s_C(\tau)| \le \alpha_n\}\\ \Omega_U &= \Omega_U(\sigma) = \{\tau \in \{\pm\}^V : \tau_U = \sigma_U \text{ and } |s_C(\tau)| \le \alpha_n\}.\end{aligned}$$

By the definition of α_n, if $\tau \in \Omega$ then $|s_A(\tau)s_C(\tau)| \le |A|\alpha_n = o(n)$ a.a.s. Thus, $\tau \in \Omega$ implies

$$\begin{aligned}Q_{A,C}(G,\tau) &= \prod_{u\in A, v\in C} \psi_{uv}(G,\tau)\\ &= \left(1-\frac{a}{n}\right)^{(|A||C|+s_A(\tau)s_C(\tau))/2}\left(1-\frac{b}{n}\right)^{(|A||C|-s_A(\tau)s_C(\tau))/2}\\ &= (1+o(1))\left(1-\frac{a}{n}\right)^{|A||C|/2}\left(1-\frac{b}{n}\right)^{|A||C|/2} \text{ a.a.s.}\end{aligned}\tag{2}$$

where we have used the fact that $u \in A$, $v \in C$ implies that $(u,v) \notin E(G)$, and thus ψ_{uv} is either $1-\frac{a}{n}$ or $1-\frac{b}{n}$. Moreover, $1-\frac{a}{n}$ appears once for every pair $(u,v) \in A \times C$ where $\tau_u = \tau_v$. The number of such pairs is $|A_+||C_+| + |A_-||C_-|$ where $A_+ = \{u \in A : \tau_u = +\}$ (and similarly for C_+, etc.); it's easy to check, then, that $2(|A_+||C_+| + |A_-||C_-|) = |A||C| + s_A s_C$, which explains the exponents in (2).

Note that the right hand side of (2) depends on G (through $A(G)$ and $C(G)$) but not on τ. Writing $2^{-n}K(G)$ for the right hand side of (2), (1) implies that if $\tau \in \Omega$ then

$$\mathbb{P}(G,\tau) = (1+o(1))K(G)Q_{A\cup B,A\cup B}(G,\tau)Q_{B\cup C,C}(G,\tau)\tag{3}$$

for a.a.e. G. Moreover, $\alpha_n/\sqrt{n} \to \infty$ implies that $\sigma \in \Omega$ for a.a.e. σ, and so for any $U = U(G)$, $\mathbb{P}(\sigma_U, G) = (1+o(1))\mathbb{P}(\sigma_U, \sigma \in \Omega, G)$ a.a.s; therefore,

$$\begin{aligned}\mathbb{P}(\sigma_U, G) &= (1+o(1))\mathbb{P}(\sigma_U, G)1_{\{\sigma\in\Omega\}}\\ &= (1+o(1))\sum_{\tau\in\Omega_U(\sigma)}\mathbb{P}(\tau, G)\\ &= (1+o(1))K(G)\sum_{\tau\in\Omega_U(\sigma)} Q_{A\cup B,A\cup B}(G,\tau)Q_{B\cup C,C}(G,\tau)\end{aligned}\tag{4}$$

for a.a.e. G and σ. (Note that the $o(1)$ term in (3) depends only on G, so there is no problem in pulling it out of the sum.) Applying (4) twice, with $U = A\cup B$ and $U = B$,

$$\begin{aligned}\mathbb{P}(\sigma_A|\sigma_B, G) &= \frac{\mathbb{P}(\sigma_{A\cup B}, G)}{\mathbb{P}(\sigma_B, G)}\\ &= (1+o(1))\frac{\sum_{\tau\in\Omega_{A\cup B}} Q_{A\cup B,A\cup B}(G,\tau)Q_{B\cup C,C}(G,\tau)}{\sum_{\tau\in\Omega_B} Q_{A\cup B,A\cup B}(G,\tau)Q_{B\cup C,C}(G,\tau)}.\end{aligned}\tag{5}$$

Note that $Q_{U_1,U_2}(\tau)$ depends on τ only through $\tau_{U_1 \cup U_2}$. In particular, in the numerator of (5), $Q_{A\cup B, A\cup B}(G, \tau)$ doesn't depend on τ since we only sum over τ with $\tau_{A\cup B} = \sigma_{A\cup B}$. Hence, the right hand side of (5) is just

$$(1 + o(1)) \frac{Q_{A\cup B, A\cup B}(G, \sigma) \sum_{\tau \in \Omega_{A\cup B}} Q_{B\cup C, C}(G, \tau)}{\left(\sum_{\tau \in \Omega_{B\cup C}} Q_{A\cup B, A\cup B}(G, \tau)\right)\left(\sum_{\tau \in \Omega_{A\cup B}} Q_{B\cup C, C}(G, \tau)\right)}, \qquad (6)$$

where we could factorize the denominator because with τ_B fixed, $Q_{A\cup B, A\cup B}$ depends only on τ_A, while $Q_{B\cup C, C}$ depends only on τ_C. Cancelling the common terms, then multiplying top and bottom by $Q_{B\cup C, C}(G, \sigma)$, we have

$$\begin{aligned}
(6) &= (1 + o(1)) \frac{Q_{A\cup B, A\cup B}(G, \sigma)}{\sum_{\tau \in \Omega_{B\cup C}} Q_{A\cup B, A\cup B}(G, \tau)} \\
&= (1 + o(1)) \frac{Q_{A\cup B, A\cup B}(G, \sigma) Q_{B\cup C, C}(G, \sigma)}{\sum_{\tau \in \Omega_{B\cup C}} Q_{A\cup B, A\cup B}(G, \tau) Q_{B\cup C, C}(G, \tau)} \\
&= (1 + o(1)) \frac{\mathbb{P}(G, \sigma)}{\mathbb{P}(G, \sigma_{B\cup C})} \\
&= (1 + o(1)) \mathbb{P}(\sigma_A | \sigma_{B\cup C}, G) \text{ a.a.s.}
\end{aligned}$$

where the penultimate line used (4) for the denominator and (3) (plus the fact that $\sigma \in \Omega$ a.a.s.) for the numerator. On the other hand, recall from (5) that (6) $= (1 + o(1))\mathbb{P}(\sigma_A | \sigma_B, G)$ a.a.s. □

5 The second moment argument

In this section, we will prove Theorem 2. The general direction of this proof was already described in the introduction, but let's begin here with a slightly more detailed overview. Recall that $\mathbb{P}'_n$ denotes the Erdös–Renyi model $\mathcal{G}(n, \frac{a+b}{2n})$. The first thing we will do is to extend $\mathbb{P}'_n$ to be a distribution on labelled graphs. In order to do this, we only need to describe the conditional distribution of the label given the graph. We will take

$$\mathbb{P}'_n(\sigma | G) = \frac{\mathbb{P}_n(G|\sigma)}{Z_n(G)},$$

where $Z_n(G)$ is the normalization constant for which this is a probability. Now, our goal is to show that $\frac{\mathbb{P}_n(G,\sigma)}{\mathbb{P}'_n(G,\sigma)}$ is well-behaved; with our definition of $\mathbb{P}'_n(\sigma|G)$, we have

$$\frac{\mathbb{P}_n(G, \sigma)}{\mathbb{P}'_n(G, \sigma)} = \frac{\mathbb{P}_n(\sigma) Z_n(G)}{\mathbb{P}'_n(G)} = 2^{-n} \frac{Z_n(G)}{\mathbb{P}'_n(G)}.$$

Thus, Theorem 2 reduces to the study of the partition function $Z_n(G)$. To do this, we will use the small subgraph conditioning method. This method was developed by Robinson and Wormald [30,31] in order to prove that most d-regular graphs are

Hamiltonian. Janson [19] then showed that the method can be used to prove contiguity, and it has since been applied in many different settings (see the survey [38] for a more detailed discussion). Essentially, the method is useful for studying a sequence $Y_n(G_n)$ of random variables which are not concentrated around their means, but which become concentrated when we condition on the number of short cycles that G_n has. Fortunately for us, this method has been developed into an easily applicable tool, the application of which only requires the calculation of some joint moments. The formulation below comes from [38], Theorem 4.1.

Theorem 6 *Fix two sequences of probability distributions $\mathbb{P}'_n$ and $\mathbb{P}_n$ on a common sequence of discrete measure spaces, and let $Y_n = \frac{\mathbb{P}_n}{\mathbb{P}'_n}$ be the density of $\mathbb{P}_n$ with respect to $\mathbb{P}_n$. Let $\lambda_k > 0$ and $\delta_k \geq -1$ be real numbers. For each n, suppose that there are random variables $X_k = X_k(n) \in \mathbb{N}$ for $k \geq 3$ such that*

(a) *For each fixed $m \geq 1$, $\{X_k(n)\}_{k=3}^m$ converge jointly under $\mathbb{P}'_n$ to independent Poisson variables with means λ_k;*
(b) *For every $j_1, \dots, j_m \in \mathbb{N}$,*

$$\frac{\mathbb{E}_{\mathbb{P}'_n}\left(Y_n[X_3(n)]_{j_1} \cdots [X_m(n)]_{j_m}\right)}{\mathbb{E}_{\mathbb{P}'_n} Y_n} \to \prod_{k=3}^{m} (\lambda_k(1+\delta_k))^{j_k};$$

(c)

$$\sum_{k\geq 3} \lambda_k \delta_k^2 < \infty;$$

(d)

$$\frac{\mathbb{E}_{\mathbb{P}'_n} Y_n^2}{(\mathbb{E}_{\mathbb{P}'_n} Y_n)^2} \to \exp\left(\sum_{k\geq 3} \lambda_k \delta_k^2\right).$$

Then $\mathbb{P}'_n$ and $\mathbb{P}_n$ are contiguous.

In our application of Theorem 6 the discussion at the beginning of this section implies that $Y_n = Y_n(G) = 2^{-n} \frac{Z_n(G)}{\mathbb{P}'_n(G)}$. We will take $X_k(n)$ to be the number of k-cycles in G_n. Thus, condition (a) in Theorem 6 is already well-known, with $\lambda_k = \frac{1}{2k}\left(\frac{a+b}{2}\right)^k$. This leaves us with three conditions to check. We will start with (d), but before we do so, let us fix some notation.

Let σ and τ be two labellings in $\{\pm\}^n$. We will also omit the subscript n in $\mathbb{P}_n$ and $\mathbb{P}'_n$, and when we write $\prod_{(u,v)}$, we mean that u and v range over all unordered pairs of distinct vertices $u, v \in G$. Let t (for "threshold") be defined by $t = \frac{(a-b)^2}{2(a+b)}$.

For the rest of this section, $G \sim \mathbb{P}'$. Therefore we will drop the $\mathbb{P}'$ from $\mathbb{E}_{\mathbb{P}'}$ and just write $\mathbb{E}$.

5.1 The first two moments of Y_n

Since $Y_n = \frac{\mathbb{P}(G,\sigma)}{\mathbb{P}'(G,\sigma)}$, $\mathbb{E}Y_n = 1$ trivially. Let's do a short computation to double-check it, though, because it will be useful later. Define

$$W_{uv} = W_{uv}(G,\sigma) = \begin{cases} \frac{2a}{a+b} & \text{if} \quad \sigma_u = \sigma_v, \quad (u,v) \in E \\ \frac{2b}{a+b} & \text{if} \quad \sigma_u \neq \sigma_v, \quad (u,v) \in E \\ \frac{n-a}{n-(a+b)/2} & \text{if} \quad \sigma_u = \sigma_v, \quad (u,v) \notin E \\ \frac{n-b}{n-(a+b)/2} & \text{if} \quad \sigma_u \neq \sigma_v, \quad (u,v) \notin E \end{cases}$$

and define V_{uv} by the same formula, but with σ replaced by τ. Then

$$Y_n = 2^{-n} \sum_{\sigma \in \{\pm\}^n} \prod_{(u,v)} W_{uv}$$

and

$$Y_n^2 = 2^{-2n} \sum_{\sigma,\tau \in \{\pm\}^n} \prod_{(u,v)} W_{uv} V_{uv}.$$

Since $\{W_{uv}\}_{(u,v)}$ are independent given σ, it follows that

$$\mathbb{E}Y_n = 2^{-n} \sum_{\sigma \in \{\pm\}^n} \prod_{(u,v)} \mathbb{E}W_{uv} \tag{7}$$

and

$$\mathbb{E}Y_n^2 = 2^{-2n} \sum_{\sigma,\tau \in \{\pm\}^n} \prod_{(u,v)} \mathbb{E}W_{uv} V_{uv}. \tag{8}$$

Thus, to compute $\mathbb{E}Y_n$, we should compute $\mathbb{E}W_{uv}$, while computing $\mathbb{E}Y_n^2$ involves computing $\mathbb{E}W_{uv}V_{uv}$.

Lemma 7 *For any fixed σ,*

$$\mathbb{E}W_{uv}(G,\sigma) = 1.$$

Proof of Lemma 4 Suppose $\sigma_u = \sigma_v$. Then $\mathbb{P}'((u,v) \in E) = \frac{a+b}{2n}$, so

$$\mathbb{E}W_{uv} = \frac{2a}{a+b} \cdot \frac{a+b}{2n} + \frac{n-a}{n-(a+b)/2} \cdot \left(1 - \frac{a+b}{2n}\right) = \frac{a}{n} + 1 - \frac{a}{n} = 1.$$

The case for $\sigma_u \neq \sigma_v$ is similar. □

Notwithstanding that computing $\mathbb{E}Y_n$ is trivial anyway, Lemma 7 and (7) together imply that $\mathbb{E}Y_n = 1$. Let us now move on to the second moment.

Lemma 8 *If* $\sigma_u\sigma_v\tau_u\tau_v = +$ *then*

$$\mathbb{E}W_{uv}V_{uv} = 1 + \frac{1}{n}\cdot\frac{(a-b)^2}{2(a+b)} + \frac{(a-b)^2}{4n^2} + O(n^{-3}).$$

If $\sigma_u\sigma_v\tau_u\tau_v = -$ *then*

$$\mathbb{E}W_{uv}V_{uv} = 1 - \frac{1}{n}\cdot\frac{(a-b)^2}{2(a+b)} - \frac{(a-b)^2}{4n^2} + O(n^{-3}).$$

Proof Suppose $\sigma_u\sigma_v = \tau_u\tau_v = +1$. Then

$$\begin{aligned}\mathbb{E}W_{uv}V_{uv} &= \left(\frac{2a}{a+b}\right)^2\cdot\frac{a+b}{2n} + \left(\frac{n-a}{n-(a+b)/2}\right)^2\cdot\left(1-\frac{a+b}{2n}\right)\\ &= \frac{2a^2}{n(a+b)} + \frac{(1-\frac{a}{n})^2}{1-\frac{a+b}{2n}}\\ &= \frac{2a^2}{n(a+b)} + \left(1-\frac{a}{n}\right)^2\left(1+\frac{a+b}{2n}+\frac{(a+b)^2}{4n^2}+O(n^{-3})\right)\\ &= 1+\frac{1}{n}\cdot\frac{(a-b)^2}{2(a+b)} + \frac{(a-b)^2}{4n^2} + O(n^{-3}).\end{aligned}$$

The computation for $\sigma_u\sigma_v = \tau_u\tau_v = -1$ is analogous.

Now assume $\sigma_u\sigma_v = +1$ while $\tau_u\tau_v = -1$. By a very similar computation,

$$\begin{aligned}\mathbb{E}W_{uv}V_{uv} &= \frac{4ab}{(a+b)^2}\cdot\frac{a+b}{2n} + \frac{(1-\frac{a}{n})(1-\frac{b}{n})}{(1-\frac{a+b}{2n})^2}\left(1-\frac{a+b}{2n}\right)\\ &= 1-\frac{1}{n}\cdot\frac{(a-b)^2}{2(a+b)} - \frac{(a-b)^2}{4n^2} + O(n^{-3}).\end{aligned}$$

The computation for $\sigma_u\sigma_v = -1$, $\tau_u\tau_v = +1$ is analogous. □

Given what we said just before Lemma 7, we can now compute $\mathbb{E}Y_n^2$ just by looking at the number of (u, v) where $\sigma_u\sigma_v\tau_u\tau_v = \pm 1$. To make this easier, we introduce another parameter, $\rho = \rho(\sigma,\tau) = \frac{1}{n}\sum_i \sigma_i\tau_i$. Writing $s_\pm$ for the number of $\{u, v\}$ with $u \neq v$ for which $\sigma_u\sigma_v\tau_u\tau_v = \pm$ we get:

$$\rho^2 = n^{-1} + 2n^{-2}\sum_{u\neq v}\sigma_u\sigma_v\tau_u\tau_v = n^{-1} + 2n^{-2}(s_+ - s_-)$$

Since we also have $2n^{-2}(s_+ + s_-) = 1 - n^{-1}$, we obtain

$$s_+ = (1+\rho^2)\frac{n^2}{4} - \frac{n}{2},\quad s_- = (1-\rho^2)\frac{n^2}{4}.$$

Lemma 9

$$\mathbb{E}Y_n^2 = (1+o(1))\frac{e^{-t/2-t^2/4}}{\sqrt{1-t}}.$$

Before we proceed to the proof, recall (or check, by writing out the Taylor series of the logarithm) that

$$\left(1+\frac{x}{n}\right)^{n^2} = (1+o(1))e^{nx-\frac{1}{2}x^2}$$

as $n \to \infty$.

Proof Define $\gamma_n = \frac{t}{n} + \frac{(a-b)^2}{4n^2}$; note that

$$\begin{aligned}
(1+\gamma_n)^{n^2} &= (1+o(1))\exp\left(\frac{(a-b)^2}{4} + tn - \frac{t^2}{2}\right) \\
(1-\gamma_n)^{n^2} &= (1+o(1))\exp\left(-\frac{(a-b)^2}{4} - tn - \frac{t^2}{2}\right) \\
(1+\gamma_n)^{n} &= (1+o(1))\exp(t).
\end{aligned}$$

Then, by Lemma 8,

$$\begin{aligned}
2^{2n}\mathbb{E}Y_n^2 &= \sum_{\sigma,\tau}\prod_{(u,v)}\mathbb{E}W_{uv}V_{uv} \\
&= \sum_{\sigma,\tau}(1+\gamma_n+O(n^{-3}))^{s_+}(1-\gamma_n+O(n^{-3}))^{s_-} \\
&= (1+o(1))e^{-t/2}\sum_{\sigma,\tau}(1+\gamma_n)^{(1+\rho^2)n^2/4}(1-\gamma_n)^{(1-\rho^2)n^2/4} \\
&= (1+o(1))e^{-t/2-t^2/4}\sum_{\sigma,\tau}\exp\left(\frac{\rho^2}{2}\left(\frac{(a-b)^2}{4}+tn\right)\right).
\end{aligned}$$

Computing the last term would be easy if $\rho\sqrt{n}$ were normally distributed. Instead, it is binomially distributed, which—unsurprisingly—is just as good. To show it, though, will require a slight digression.

Lemma 10 *If $\xi_i \in \{\pm\}$ are taken uniformly and independently at random and $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$ then*

$$\mathbb{E}\exp(sZ_n^2/2) \to \frac{1}{\sqrt{1-s}}$$

whenever $s < 1$.

Proof Since $z \mapsto \exp(sz^2/2)$ is a continuous function, the central limit theorem implies that $\exp(sZ_n^2/2) \xrightarrow{d} \exp(sZ^2/2)$, where $Z \sim \mathcal{N}(0,1)$. Now, $\mathbb{E}\exp(sZ^2/2) = \frac{1}{\sqrt{1-s}}$ and so the proof is complete if we can show that the sequence $\exp(sZ_n^2/2)$ is uniformly integrable. But this follows from Hoeffding's inequality:

$$\Pr(\exp(sZ_n^2/2) \geq M) = \Pr\left(|Z_n| \geq \sqrt{\frac{2\log M}{s}}\right) \leq M^{-1/s},$$

which is integrable near ∞ (uniformly in n) whenever $s < 1$. □

To finish the proof of Lemma 9, take Z_n as in Lemma 10 and note that

$$2^{-2n}\sum_{\sigma,\tau}\exp\left(\frac{\rho^2}{2}\left(\frac{(a-b)^2}{4}+tn\right)\right) = \mathbb{E}\exp\left(\frac{t(1+o(1))}{2}Z_n^2\right) \to \frac{1}{\sqrt{1-t}}.$$

5.2 Dependence on the number of short cycles

Our next task is to check condition (b) in Theorem 6. Note, therefore, that $[X_3]_{j_3}\cdots[X_m]_{j_m}$ is the number of ways to have an ordered tuple containing j_3 3-cycles of G, j_4 4-cycles of G, and so on. Therefore, if we can compute $\mathbb{E}Y_n 1_H$ where 1_H indicates that any particular union of cycles occurs in G_n, then we can compute $\mathbb{E}Y_n[X_3]_{m_3}\cdots[X_m]_{j_m}$. Computing $\mathbb{E}Y_n 1_H$ is the main task of this section; we will do it in three steps. First, we will get a general formula for $\mathbb{E}Y_n 1_H$ in terms of H. We will apply this general formula in the case that H is a single cycle and get a much simpler formula back. Finally, we will extend this to the case when H is a union of vertex-disjoint cycles.

As promised, we begin the program with a general formula for $\mathbb{E}1_H Y_n$. Let H be a graph on some subset of $[n]$, with $|V(H)| = m$. With some slight abuse of notation, We write 1_H for the random variable that is 1 when $H \subset G$, and $\mathbb{P}'(H)$ for the probability that $H \subset G$.

Lemma 11

$$\mathbb{E}1_H Y_n = 2^{-m}\mathbb{P}'(H)\sum_{\sigma\in\{\pm1\}^m}\prod_{(u,v)\in E(H)} w_{uv}(\sigma),$$

where

$$w_{uv}(\sigma) = \begin{cases} \frac{2a}{a+b} & \textit{if } (u,v) \in S(\sigma) \\ \frac{2b}{a+b} & \textit{otherwise.} \end{cases}$$

Proof of Lemma 4 We break up $\sigma \in \{\pm1\}^n$ into $(\sigma_1,\sigma_2) \in \{\pm1\}^{V(H)}\times\{\pm1\}^{V(G)\setminus V(H)}$ and sum over the two parts separately. Note that if $(u,v) \in E(H)$ then $W_{uv}(G,\sigma)$

depends on σ only through σ_1. Let $D(H) = E(G)\backslash E(H)$, so that $(u, v) \in D(H)$ implies that W_{uv} and 1_H are independent. Then

$$\begin{aligned}\mathbb{E}1_H Y_n &= 2^{-n}\sum_{\sigma_1}\sum_{\sigma_2}\mathbb{E}1_H\prod_{(u,v)} W_{uv}(G,\sigma)\\ &= 2^{-n}\sum_{\sigma_1}\Bigg(\Bigg(\mathbb{E}1_H\prod_{(u,v)\in E(H)} W_{uv}\Bigg)\sum_{\sigma_2}\prod_{(u,v)\in D(H)}\mathbb{E}W_{uv}\Bigg)\\ &= 2^{-m}\sum_{\sigma_1}\Bigg(\mathbb{E}1_H\prod_{(u,v)\in E(H)} W_{uv}\Bigg),\end{aligned}$$

because if $(u, v) \in D(H)$ then, for every σ, Lemma 7 says that $\mathbb{E}W_{uv}(G, \sigma) = 1$. To complete the proof, note that if $(u, v) \in E(H)$ then for any σ, $W_{uv}(G, \sigma) \equiv w_{uv}(\sigma)$ on the event $H \subseteq G$. □

The next step is to compute the right hand side of Lemma 11 in the case that H is a cycle. This computation is very similar to the one in Lemma 1, when we computed the expected number of k-cycles in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$. Essentially, we want to compute the expected "weight" of a cycle, where the weight of each edge depends only on whether its endpoints have the same label or not.

Lemma 12 *If H is a k-cycle then*

$$\sum_{\sigma\in\{\pm 1\}^H}\prod_{(u,v)\in E(H)} w_{uv}(\sigma) = 2^k\left(1+\left(\frac{a-b}{a+b}\right)^k\right).$$

Proof of Lemma 4 Let $e_1, \ldots, e_k$ be the edges of H. Provided that we renormalize, we can replace the sum over σ by an expectation, where σ is taken uniformly in $\{\pm 1\}^H$. Now, let N be the number of edges of H whose endpoints have different labels. As discussed in the proof of Lemma 1, $\Pr(N = j) = 2^{-k+1}\binom{k}{j}$ for even j, and zero otherwise. Then

$$\begin{aligned}\mathbb{E}_\sigma\prod_{(u,v)\in E(H)} w_{uv}(\sigma) &= \mathbb{E}_\sigma\left(\frac{2a}{a+b}\right)^{k-N}\left(\frac{2b}{a+b}\right)^N\\ &= \frac{2}{(a+b)^k}\sum_{j \text{ even}}\binom{k}{j}a^{k-j}b^j\\ &= 1+\left(\frac{a-b}{a+b}\right)^k.\end{aligned}$$

□

Extending this calculation to vertex-disjoint unions of cycles is quite easy: suppose H is the union of cycles H_i. Since $w_{uv}(\sigma)$ only depends on σ_u and σ_v, we can just split up the sum over $\sigma \in \{\pm\}^H$ into a product of sums, where each sum ranges over $\{\pm\}^{H_i}$. Then applying Lemma 12 to each H_i yields a formula for H.

Lemma 13 *Define*

$$\delta_k = \left(\frac{a-b}{a+b}\right)^k.$$

If $H = \bigcup_i H_i$ is a vertex-disjoint union of graphs and each H_i is a k_i-cycle, then

$$\sum_{\sigma \in \{\pm 1\}^H} \prod_{(u,v)\in E(H)} w_{uv}(H,\sigma) = 2^{|H|} \prod_i (1+\delta_{k_i}).$$

We we need one last ingredient, which we hinted at earlier, before we can show condition (b) of Theorem 6. We only know how to exactly compute $\mathbb{E} Y_n 1_H$ when H is a disjoint union of cycles. Now, most tuples of cycles are disjoint, but in order to dismiss the contributions from the non-disjoint unions, we need some bound on $\mathbb{E} Y_n 1_H$ that holds for all H:

Lemma 14 *For any H,*

$$\sum_{\sigma \in \{\pm 1\}^H} \prod_{(u,v)\in E(H)} w_{uv}(\sigma) \le 2^{|H|+|E(H)|}.$$

Proof

$$w_{uv}(\sigma) \le \frac{2\max\{a,b\}}{a+b} \le 2$$

for any i, j, H and σ. □

Finally, we are ready to put these ingredients together and prove condition (b) of Theorem 6. For the rest of the section, take $\delta_k = (\frac{a-b}{a+b})^k$ as it was in Lemma 13. Also, recall that $\lambda_k = \frac{1}{2k}\left(\frac{a+b}{2}\right)^k$ is the limit of $\mathbb{E} X_k$ as $n \to \infty$.

Lemma 15 *Let X_k be the number of k-cycles in G. For any $j_3, \ldots, j_m \in \mathbb{N}$,*

$$\mathbb{E} Y_n \prod_{k=3}^m [X_k]_{j_k} \to \prod_{k=3}^m (\lambda_k(1+\delta_k))^{j_k}.$$

Proof Set $M = \sum_k k m_k$. First of all,

$$[X_k]_j = \sum_{H_1,\ldots,H_j} \prod_i 1_{H_i}$$

where the sum ranges over all j-tuples of distinct k-cycles, and 1_H indicates the event that the subgraph H appears in G. Thus,

$$\prod_{k=3}^m [X_k]_{j_k} = \sum_{(H_{ki})} \prod_{k=3}^m \prod_{i=1}^{j_k} 1_{H_{ki}} = \sum_{(H_{ki})} 1_{\{\bigcup H_{ki}\}},$$

where the sum ranges over all M-tuples of cycles $(H_{ki})_{k \le m, i \le j_k}$ for which each H_{ki} is an k-cycle, and every cycle is distinct. Let $\mathcal{H}$ be the set of such tuples; let $A \subset \mathcal{H}$ be the set of such tuples for which the cycles are vertex-disjoint, and let $B = \mathcal{H} \backslash A$. Thus, if $H = \bigcup H_{ki}$ for $(H_{ki}) \in A$, then

$$\mathbb{E} Y_n 1_H = \prod_k (1 + \delta_k)^{j_k} \mathbb{P}'(H)$$

by Lemmas 11 and 13. Note also that standard counting arguments (see, for example, [3], Chapter 4) imply that $|A| \mathbb{P}'(H) \to \prod_k \lambda_k^{j_k}$.

On the other hand, if $(H_{ki}) \in B$ then $H := \bigcup_{ki} H_{ki}$ has at most $M - 1$ vertices, M edges, and its number of edges is strictly larger than its number of vertices. Thus, $\mathbb{P}'(H)\binom{n}{|H|} \to 0$, so Lemmas 11 and 14 imply that

$$\sum_{H' \sim H} \mathbb{E} Y_n 1_H \le \mathbb{P}'(H) |H|! \binom{n}{|H|} 2^M \to 0,$$

where the sum ranges over all ways to make an isomorphic copy of H on n vertices. Since there are only a bounded number of isomorphism classes in

$$\left\{ \bigcup_{ki} H_{ki} : (H_{ki}) \in B \right\},$$

it follows that $\sum_H \mathbb{E} Y_n 1_H \to 0$, where the sum ranges over all unions of non-disjoint tuples in $\mathcal{H}$. Thus,

$$\begin{aligned} \mathbb{E} Y_n \prod_{k=3}^{m} [X_k]_{j_k} &= \mathbb{E} Y_n \left(\sum_{(H_{ki}) \in A} 1_{\bigcup H_{ki}} + \sum_{(H_{ki}) \notin B} 1_{\bigcup H_{ki}} \right) \\ &= |A| \mathbb{P}'(H) \prod_k (1 + \delta_k)^{j_k} + o(1) \\ &\to \prod_k (\lambda_k (1 + \delta_k))^{j_k}. \end{aligned}$$

To complete the proof of Theorem 2, note that $\delta_k^2 \lambda_k = \frac{t^k}{2k}$. Thus, $\sum_{k \ge 3} \delta_k^2 \lambda_k = \frac{1}{2}(\log(1 - t) - t - t^2/2)$. When $t < 1$, this (with Lemma 9) proves conditions (c) and (d) of Theorem 6. Since condition (a) is classical and condition (b) is given by Lemma 15, the conclusion of Theorem 6 implies the first statement in Theorem 2.

We finally apply the first half of Theorem 2 to show that no estimator can be consistent when $(a - b)^2 < 2(a + b)$. In fact, if $\hat{a}$ and $\hat{b}$ are estimators for a and b which converge in probability, then their limit when $(a - b)^2 < 2(a + b)$ depends only on $a + b$. To see this, let α, β be another choice of parameters with $(\alpha - \beta)^2 < 2(\alpha + \beta)$ and $\alpha + \beta = a + b$; let $\mathbb{Q}_n = \mathcal{G}_n(\alpha, \beta)$; take a^* to be the in-probability limit of $\hat{a}$ under $\mathbb{P}_n$ and α^* to be its limit under $\mathbb{Q}_n$. For an arbitrary $\epsilon > 0$, let A_n be the event

$|\hat{a} - a^*| > \epsilon$; thus, $\mathbb{P}_n(A_n) \to 0$. By the first part of Theorem 2, $\mathbb{P}'_n(A_n) \to 0$ also. Since $\alpha + \beta = a + b$, we can apply the first part of Theorem 2 to $\mathbb{Q}_n$, implying that $\mathbb{Q}_n(A_n) \to 0$ and so $\alpha^* = a^*$. That is, $\hat{a}$ converges to the same limit under $\mathbb{Q}_n$ and $\mathbb{P}_n$.

6 Conjectures and open problems

6.1 Regular models

We briefly discuss how can one define a regular version of the model and what we expect from the behavior of such a model. A regular model should satisfy the following properties:

- The graph G is a.s. a simple d-regular graph.
- For each vertex u among the d neighbors it is connected to, it is connected to $\mathrm{Binom}(d, 1-\epsilon)$ vertices v with $\sigma_v = \sigma_u$.
- Choices at different vertices are (almost) independent.

As is often the case with random regular graphs, the construction is not completely trivial. Here are two possible constructions:

- Let $\{X_v : v \in V\}$ be a collection of independent $\mathrm{Binom}(d, 1-\epsilon)$ variables, conditioned on

$$\sum_{v:\sigma_v=+} X_v = \sum_{v:\sigma_v=-} X_v \quad \text{is even.}$$

  Now the $(+,+)$ edges are defined by sampling a uniform random graph on $\{v : \sigma_v = +\}$ with degree distribution given by $\{X_v : \sigma_v = +\}$, while the $(-,-)$ edges are defined by sampling a uniform random graph on $\{v : \sigma_v = -\}$ with degree distribution given by $\{X_v : \sigma_v = -\}$. To construct the $(+,-)$ edges we take a uniformly random bipartite graph with left degrees given by $\{d - X_v : \sigma_v = +\}$ and right degrees given by $\{d - X_v : \sigma_v = -\}$.
- The second construction uses a variant of the configuration model. We generate the graph by generating d independent matchings. The probability of each matching is proportional to $(1-\epsilon)^{n_=}\epsilon^{n_{\neq}}$, where $n_=$ is the number of edges (u, v) with $\sigma_u = \sigma_v$ points and $n_{\neq}$ is the number of edges (u, v) with $\sigma_u \neq \sigma_v$.

We conjecture that the results of the paper should extend to the models above where the quantity $(a-b)^2/2(a+b)$ is now replaced by $(d-1)\theta^2$, where $\theta = 1 - 2\epsilon$. Friedman's proof of Alon's conjecture [13] gives a very accurate information regarding the spectrum of uniformly random d-regular graphs. We propose the following related conjecture.

Conjecture 4 *Assume $(d-1)\theta^2 > 1$. Then there exist an $\delta > 0$, s.t. with high probability, the second eigenvalue of the graph generated $\lambda_2(G)$ satisfies $\lambda_2(G) > 2\sqrt{d-1} + \delta$. Moreover, all other eigenvalues of G are smaller than $2\sqrt{d-1}$, and the eigenvector associated to $\lambda_2(G)$ is correlated with the true partition.*

By comparison, the results of [13] imply that for all $\delta > 0$ with high probability, if G is a uniformly random d-regular graph then $\lambda_2(G) < 2\sqrt{d-1} + \delta$. Thus the result above provides a simple spectral algorithm to distinguish between the standard random d-regular model and the biased d-regular model when $(d-1)\theta^2 > 1$. Moreover, our conjecture also says that a spectral algorithm can be used to solve the clustering problem.

Below we sketch a proof for part of Conjecture 4. Specifically, we will show that if $(d-1)\theta^2 > 1$ then there is an approximate eigenvalue-eigenvector pair (λ, f) (in the sense that $Af \approx \lambda f$ where A is the adjacencency matrix of G) where $\lambda > 2\sqrt{d-1}+\delta$ and f is correlated with the true partition. The more difficult part of the conjecture would be to show that all other eigenvalues are smaller than $2\sqrt{d-1}$. If this were true, it would imply that $\lambda_2(G) \approx \lambda$ and that the eigenvector of $\lambda_2(G)$ is close to f.

Proof We will assume that G satisfies the following two properties:

- The process around each vertex looks like the Ising model on a d regular tree.
- Given two different vertices u, v, the process in neighborhoods of u and v are asymptotically independent.

Let r be a large constant and let $f(v) = \sum\{\sigma_w : d(w, v) = r\}$. Then $\sum_v f(v) = 0$ and it is therefore orthogonal to the leading eigenvector. Let A be the adjacency matrix of the graph. We claim that $\|Af - \lambda f\|_2$ is much smaller than $\|f\|_2$, where $\lambda = \theta^{-1} + (d-1)\theta$. Note that $\lambda > 2\sqrt{d-1}$ if and only if $|\theta| > (d-1)^{-1/2}$.

Assuming that the neighborhood of v is a d-regular tree,

$$(Af)(v) = \sum_{w:d(v,w)=r+1} \sigma_w + (d-1) \sum_{w:d(v,w)=r-1} \sigma_w$$

and so we can write $(Af)(v) - \lambda f(v)$ as

$$\begin{aligned} Af(v) - \lambda f(v) = &\left(\sum_{w:d(v,w)=r+1} \sigma_w - \theta(d-1) \sum_{w:d(v,w)=r} \sigma_w \right) \\ &-\theta^{-1} \left(\sum_{w:d(v,w)=r} \sigma_w - \theta(d-1) \sum_{w:d(v,w)=r-1} \sigma_w \right) \end{aligned} \tag{9}$$

We can re-arrange the first sum as

$$\sum_{\{w:d(v,w)=r\}} \sum_{\{w' \sim w, d(w',v)=r+1\}} \sigma_w - \theta\sigma_{w'}.$$

Noting that all the summands are independent given $\{\sigma_w : d(v, w) = r\}$, we see that the above sum has expectation zero and variance of the order $C(d-1)^r$ for some constant C. Applying a similar decomposition (but at level $r-1$) to the second sum in (9), we get

$$\mathbb{E}[(Af(v) - \lambda f(v))^2] \le C(d-1)^r.$$

Summing over all v, we conclude that

$$\mathbb{E}[\|Af - \lambda f\|_2^2] \leq Cn(d-1)^r.$$

On the other hand, from [15] it follows that for each v individually

$$\mathbb{E}[f(v)^2] \geq C'((d-1)\theta)^{2r},$$

for some absolute constant C'. Since the value of $f(v)$ and $f(w)$ for $v \neq w$ are essentially independent, it follows that with high probability $\|f\|_2^2 > C'n((d-1)\theta)^{2r}$. Taking r sufficiently large we see that $\|Af - \lambda f\|_2 \leq \delta(r)\|f\|_2$ with high probability where $\delta(r) \to 0$ as $r \to \infty$. □

6.2 Open problems for the non-regular model

Of the conjectures that we mentioned in the introduction, Conjecture 1 remains open. However, there are variations and extensions of Conjectures 1–3 that may be even more interesting. For example, we could ask whether Conjecture 1 can be realized by one of several popular and efficient algorithms.

Conjecture 5 1. *If $(a-b)^2 > 2(a+b)$ then the clustering problem in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ can be solved by a spectral algorithm.*
2. *If $(a-b)^2 > 2(a+b)$ then the clustering problem in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ can be solved by the belief propogation algorithm of* [9].
3. *If $(a-b)^2 > 2(a+b)$ then the clustering problem in $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ can be solved by simulating an Ising model on G, conditioned to be almost balanced.*

Of these conjectures, part 1 is closely related to the work of Coja-Oghlan [7], while part 3 would substantially extend the result of Dyer and Frieze [11].

Another way to extend Conjectures 1–3 would be to increase the number of clusters from two to k. The model $\mathcal{G}(n, p, q)$ is well-studied for more than two clusters, in which case it is known as the "planted partition" model. In fact, many of the results that we cited in the introduction extend to $k > 2$ also. However, the work of [9] suggests that the case of larger k is rather more delicate than the case $k = 2$, and that it contains interesting connections to complexity theory. The following conjecture comes from their work, and it is based on a connection to phase transitions in the Potts model on trees:

Conjecture 6 *For any k, there exists $c(k)$ such that if $a > b$ then:*

1. *If $\frac{(a-b)^2}{a+(k-1)b} < c(k)$ then the clustering problem cannot be solved;*
2. *If $c(k) < \frac{(a-b)^2}{a+(k-1)b} < k$ then the clustering problem is solvable, but not in polynomial time;*
3. *If $\frac{(a-b)^2}{a+(k-1)b} > k$ then the clustering problem can be solved in polynomial time.*

When $k \leq 4$, $c(k) = k$ and so case 2 does not occur. When $k \geq 5$, $c(k) < k$.

Part of the difficulty in studying Conjecture 6 can be seen from work of the third author [34]. His work contains the best known non-reconstruction results for the Potts model on trees, but the results for $k > 2$ are less precise and more difficult to prove than what is known for $k = 2$.

Decelle et al. also state a version of Conjecture 6 in the case $a < b$. Although this case is not naturally connected to clustering, it has close connections to random Boolean satisfiability problems and to spin glasses. In particular, they conjecture that when $a < b$, case 2 above becomes much larger.

Acknowledgments A.S. would like to thank Christian Borgs for suggesting the problem and Lenka Zdeborová for useful discussions. Part of this work was done while A.S. was at Microsoft Research, Redmond. The authors would also like to thank Lenka Zdeborová for comments on a draft of this work, and the two anonymous reviewers for further helpful comments.

References

1. Bickel, P.J., Chen, A.: A nonparametric view of network models and Newman–Girvan and other modularities. Proc. Natl. Acad. Sci. **106**(50), 21068–21073 (2009)
2. Bleher, P.M., Ruiz, J., Zagrebnov, V.A.: On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. J. Stat. Phys. **79**(1), 473–482 (1995)
3. Bollobás, B.: Random Graphs, 2nd edn. Cambridge University Press, Cambridge (2001)
4. Bollobás, B., Janson, S., Riordan, O.: The phase transition in inhomogeneous random graphs. Random Struct. Alg. **31**(1), 3–122 (2007)
5. Boppana, R.B.: Eigenvalues and graph bisection: an average-case analysis. In: 28th Annual Symposium on Foundations of Computer Science, pp. 280–285. IEEE (1987)
6. Bui, T.N., Chaudhuri, S., Leighton, F.T., Sipser, M.: Graph bisection algorithms with good average case behavior. Combinatorica **7**(2), 171–191 (1987)
7. Coja-Oghlan, A.: Graph partitioning via adaptive spectral techniques. Combinat. Prob. Comput. **19**(02), 227–284 (2010)
8. Condon, A., Karp, R.M.: Algorithms for graph partitioning on the planted partition model. Random Struct. Alg. **18**(2), 116–140 (2001)
9. Decelle, A., Krzakala, F., Moore, C., Zdeborová, L.: Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Phys. Rev. E **84**, 066106 (Dec 2011)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodological), **39**(1), 1–38 (1977)
11. Dyer, M.E., Frieze, A.M.: The solution of some random NP-hard problems in polynomial expected time. J. Alg. **10**(4), 451–489 (1989)
12. Evans, W., Kenyon, C., Peres, Y., Schulman, L.J.: Broadcasting on trees and the Ising model. Ann. Appl. Prob. **10**(2), 410–433 (2000)
13. Friedman, J.: A proof of Alon's second eigenvalue conjecture and related problems. Mem. Am. Math. Soc. **195**(910), viii+100 (2008)
14. Garey, M.R., Johnson, D.S., Stockmeyer, L.: Some simplified NP-complete graph problems. Theor. Computer Sci. **1**(3), 237–267 (1976)
15. Häggström, O., Mossel, E.: Nearest-neighbor walks with low predictability profile and percolation in $2 + \epsilon$ dimensions. Ann. Prob. **26**(3), 1212–1231 (1998)
16. Hartuv, E., Shamir, R.: A clustering algorithm based on graph connectivity. Inf. Process. Lett. **76**(4), 175–181 (2000)
17. Hodges, J.L., Le Cam, L.: The Poisson approximation to the Poisson binomial distribution. Ann. Math. Stat. **31**(3), 737–740 (1960)
18. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)
19. Janson, S.: Random regular graphs: asymptotic distributions and contiguity. Combinat. Prob. Comput. **4**(04), 369–405 (1995)

20. Janson, S.: Asymptotic equivalence and contiguity of some random graphs. Random Struct. Alg. **36**(1), 26–45 (2010)
21. Jerrum, M., Sorkin, G.B.: The Metropolis algorithm for graph bisection. Discrete Appl. Math. **82**(1–3), 155–175 (1998)
22. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)
23. Kesten, H., Stigum, B.P.: A limit theorem for multidimensional Galton–Watson processes. Ann. Math. Stat. **37**(5), 1211–1223 (1966)
24. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical properties of community structure in large social and information networks. In: Proceeding of the 17th International Conference on World Wide Web, pp. 695–704. ACM (2008)
25. McSherry, F.: Spectral partitioning of random graphs. In: 42nd IEEE Symposium on Foundations of Computer Science, pp. 529–537. IEEE (2001)
26. Mossel, E.: Survey—information flow on trees. DIMACS Ser. Discrete Math. Theor. Computer Sci. **63**, 155–170 (2004)
27. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
28. Newman, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. Proc. Natl. Acad. Sci. USA **99**(Suppl 1), 2566 (2002)
29. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics **155**(2), 945–959 (2000)
30. Robinson, R.W., Wormald, N.C.: Almost all cubic graphs are Hamiltonian. Random Struct. Alg. **3**(2), 117–125 (1992)
31. Robinson, R.W., Wormald, N.C.: Almost all regular graphs are Hamiltonian. Random Struct. Alg. **5**(2), 363–374 (1994)
32. Rohe, K., Chatterjee, S., Yu, B.: Spectral clustering and the high-dimensional stochastic blockmodel. Ann. Stat. **39**(4), 1878–1915 (2011)
33. Shi, J., Malik, J.: Normalized cuts and image segmentation. Pattern Anal. Mach. Intell. IEEE Trans. **22**(8), 888–905 (2000)
34. Sly, A.: Reconstruction for the Potts model. Ann. Prob. **39**(4), 1365–1406 (2011)
35. Snijders, T.A.B., Nowicki, K.: Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classif. **14**(1), 75–100 (1997)
36. Sonka, M., Hlavac, V., Boyle, R.: Image processing: analysis and machine vision, 4th edn. Cengage Learning, Stamford (2015)
37. Strogatz, S.H.: Exploring complex networks. Nature **410**(6825), 268–276 (2001)
38. Wormald, N.C.: Models of random regular graphs. In: Lamb, J.D., Preece, D.A. (eds.) Surveys in Combinatorics 1999. London Mathematical Society Lecture Note Series, vol. 267. Cambridge University Press, Cambridge (1999)