

Community detection thresholds and the weak Ramanujan property

Laurent Massoulié
Microsoft Research–Inria Joint Centre
laurent.massoulie@inria.fr

November 14, 2013

Abstract

The present work is concerned with community detection, that is reconstruction of hidden components, in random graph models. Decelle et al. [1] conjectured the existence of a sharp threshold on model parameters for community detection in sparse random graphs drawn from the *stochastic block model*. Mossel, Neeman and Sly [2] established the negative part of the conjecture, proving impossibility of meaningful reconstruction below the threshold. The positive part of the conjecture remained elusive so far: results of Coja-Oghlan [3] imply that a particular spectral method applied to the graph’s adjacency matrix achieves non-trivial reconstruction, but this applies only when above the conjectured threshold by a possibly large constant.

In this work we solve the positive part of the conjecture. To that end we introduce a modified adjacency matrix B based on neighborhood expansion. Specifically B counts *simple*, or *self-avoiding* paths of a given length ℓ between pairs of nodes. We then prove that for logarithmic length ℓ , the leading eigenvectors of this modified matrix provide a non-trivial reconstruction of the underlying structure, thereby settling the conjecture. A key step in the proof consists in establishing a *weak Ramanujan property* of the constructed matrix B . Namely, the spectrum of B consists in two leading eigenvalues $\rho(B)$, λ_2 and $n-2$ eigenvalues of a lower order $O(n^\epsilon \sqrt{\rho(B)})$ for all $\epsilon > 0$, $\rho(B)$ denoting B ’s spectral radius.

Graphs are called Ramanujan when they are d -regular with second eigenvalue $\lambda \leq 2\sqrt{d-1}$. Random d -regular graphs were shown to have a second largest eigenvalue λ of $2\sqrt{d-1} + o(1)$ by Friedman [4], thus being *almost* Ramanujan. Erdős-Rényi graphs with average degree d at least logarithmic ($d = \Omega(\log n)$) were shown by Feige and Ofek [5] to have a second eigenvalue of $O(\sqrt{d})$, a slightly weaker version of the Ramanujan property.

However this spectrum separation property fails for sparse ($d = O(1)$) Erdős-Rényi graphs, whose spectrum is dominated by the presence of high-degree nodes. Our result thus shows that by constructing matrix B through neighborhood expansion, we regularize the original adjacency matrix to eventually recover a weak form of the Ramanujan property.

1 Introduction

1.1 Background

Community detection, like clustering, aims to identify groups of similar items from a global population. It is a useful primitive for performing recommendation, e.g. of contents or contacts to users of online social networks. The stochastic block model has been introduced by Holland et al. [6] to represent interactions between individuals. It consists of a random graph on n nodes, each node $i \in \mathcal{N} = \{1, \dots, n\}$ being assigned a type σ_i from some fixed set Σ . Conditionally on node types, edge (i, j) is present with probability $p(\sigma_i, \sigma_j)$ independently of other edges, for some matrix of probabilities $(p(\sigma, \sigma'))$.

It constitutes an adequate testbed for community detection. Indeed the performance of candidate detection schemes, captured by the fraction of nodes i for which estimated types $\hat{\sigma}_i$ and true types σ_i coincide, can be compared and analysed on instances of the stochastic block model. Such analyses can in turn suggest new schemes.

Recently Decelle et al. [1] conjectured the existence of a phase transition in the sparse regime where the graph's average degree is $O(1)$. Specifically, they predicted that for parameters below a certain threshold, no estimates $\hat{\sigma}_i$ of node types existed that would be positively correlated with true types σ_i , while above the threshold, belief propagation algorithms could determine estimates $\hat{\sigma}_i$ achieving such a positive correlation. Their conjecture is formulated on a simple symmetric instance of the stochastic block model featuring two node types $\{+1, -1\}$. The phenomenon appears more general though: Heimlicher et al. [7] extended the conjecture to the more general setup of labeled stochastic block models.

The study of this phenomenon is important for two reasons. First, by localizing precisely the transition point below which no useful signal is present in the observations, one thus characterizes how much subsampling of the original graph can be performed before all information is lost. Second, algorithms leading to estimates $\hat{\sigma}_i$ that achieve positive correlation all the way down to the transition are expected to constitute more robust approaches than alternatives which would fail before the transition. It is therefore important to determine such algorithms.

The negative part of the conjecture has been proven by Mossel, Neeman and Sly [2]. Essentially they established that existence of estimates $\hat{\sigma}_i$ positively correlated with true types σ_i would imply feasibility of a reconstruction problem on a random tree model describing the local statistics of the original random graph. However by results of Evans et al. [8] such reconstruction is infeasible below the conjectured transition point.

Until now, positive results in the sparse case did not apply down to the transition point. The best results to date (see [2]) relied on Coja-Oghlan [3], showing that spectral clustering applied to the adjacency matrix, suitably trimmed by removal of high degree nodes, yields positively correlated estimates. However this does not apply down to the conjectured threshold.

This limitation stems from the following fact. Spectral methods perform well on matrices enjoying a spectral separation property, namely the spectrum should comprise a few large eigenvalues whose associated eigenvectors reflect the sought structure and all other eigenvalues should be negligible. The prototype of such separation is the Ramanujan property, according to which d -regular graphs have the second eigenvalue λ no larger than $2\sqrt{d-1}$ in absolute value. Friedman [4] established that random d -regular graphs almost satisfy this, in that for them $|\lambda| \leq 2\sqrt{d-1} + o(1)$. Erdős-Rényi graphs with average degree d are such that $|\lambda| \leq O(\sqrt{d})$, *provided* $d = \Omega(\log n)$ (see Feige and Ofek [5]), but such Ramanujan-like separation is lost for smaller d . This lack of separation inherently limits the power of spectral methods in the sparse case.

1.2 Main results

We focus on the stochastic block model in Decelle et al. [1]. The graph is denoted \mathcal{G} , node types (or spins) σ_i are uniformly and i.i.d. drawn from $\{-1, +1\}$. An edge is present between any two nodes i, j with probability a/n if $\sigma_i = \sigma_j$, and b/n if $\sigma_i = -\sigma_j$, constants a and b being the model parameters. The conjectured transition point is specified by quantity $\tau = (a - b)^2/[2(a + b)]$: for $\tau < 1$ it is known that positively correlated detection is impossible; we set out to prove that it is feasible for $\tau > 1$.

We shall make use of the notations $\alpha := (a + b)/2$, $\beta := (a - b)/2$. The detectability condition $\tau > 1$ can be restated as

$$\beta^2 > \alpha. \quad (1)$$

As mentioned, Coja-Oghlan regularizes the adjacency matrix of the random graph by removing high degree nodes before applying spectral clustering. In contrast, we regularize the initial data through *path expansion*. Namely, we do not work directly with the adjacency matrix of the original graph. Instead we form matrix $B^{(\ell)}$, where $B_{ij}^{(\ell)}$ counts the number of self-avoiding paths of graph edges of length ℓ connecting i to j .

Our main result is then the following

Theorem 1.1. *Assume Condition (1) holds. Set the path length parameter ℓ such that $\ell \sim c \log(n)$ for a constant c such that $c \log(\alpha) < 1/4$. Let x be a normed eigenvector corresponding to the second largest eigenvalue of $B^{(\ell)}$. There exists t such that, defining the spin estimates $\hat{\sigma}_i$ as*

$$\hat{\sigma}_i = \begin{cases} +1 & \text{if } x_i \sqrt{n} \geq t, \\ -1 & \text{otherwise,} \end{cases} \quad (2)$$

the empirical overlap between the true and estimated spins defined as

$$ov(\sigma, \hat{\sigma}) := \frac{1}{n} \sum_{i \in \mathcal{N}} \sigma_i \hat{\sigma}_i \quad (3)$$

converges in probability to the set $\{-r, +r\}$ for some strictly positive constant $r > 0$ as $n \rightarrow \infty$.

It proves the positive part of Decelle et al.'s conjecture. It further identifies a specific spectral method based on the path-expanded matrix $B^{(\ell)}$. An auxiliary result consists in showing that matrix $B^{(\ell)}$ enjoys a spectral separation property, that is a weak version of the Ramanujan property. Namely, denoting by $\rho(B^{(\ell)})$ the spectral radius of $B^{(\ell)}$, we show that the third largest eigenvalue λ of matrix $B^{(\ell)}$ satisfies for all positive constant ϵ :

$$|\lambda| \leq n^\epsilon \sqrt{\rho(B^{(\ell)})}.$$

We note that computation of $B^{(\ell)}$ and hence of the $\hat{\sigma}_i$ can be done in polynomial time: as shown in Lemma 4.2 the ℓ -neighborhood of any i contains at most one cycle so that each $B_{ij}^{(\ell)}$ is readily evaluated by suitable breadth-first search.

1.3 Paper organization

Section 2 contains the intermediate results involved in the proof and how they combine to establish Theorem 1.1. Section 3 proves Theorem 2.2, which expresses matrix $B^{(\ell)}$ as an expansion in terms

of the matrices $B^{(m)}$, $m < \ell$, together with bounds on the spectral norm of the matrix coefficients involved. Section 4 contains the so-called “local analysis” of node neighborhoods. Specifically it gives controls on the vectors $B^{(m)}e$ and $B^{(m)}\sigma$, where e is the all-ones vector and σ is the vector of spins, establishing a quasi-deterministic growth pattern with respect to m . Section 5 concludes.

2 Proof structure

Our key objective is to determine the spectral structure of $B^{(\ell)}$. Specifically we wish to establish

Theorem 2.1. *Assume (1) and $\ell = c \log n$ with $c \log(\alpha) < 1/4$.*

(i) *The leading eigenvalue of $B^{(\ell)}$ is up to logarithmic factors $\Theta(\alpha^\ell)$, with corresponding eigenvector asymptotically parallel to $B^{(\ell)}e$.*

(ii) *Its second eigenvalue is $\Omega(\beta^\ell)$ up to logarithmic factors, with corresponding eigenvector asymptotically parallel to $B^{(\ell)}\sigma$.*

(iii) *There is a random variable X with unit mean and variance $1/(\beta^2/\alpha - 1)$ such that for all x that is an atom of neither X 's nor $-X$'s distribution, the following convergence in probability holds for any normed vector y asymptotically aligned with $B^{(\ell)}\sigma$:*

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1} \left\{ y_i \geq \frac{x}{\sqrt{n \mathbf{E}(X^2)}} \right\} \rightarrow \frac{1}{2} [\mathbf{P}(X \geq x) - \mathbf{P}(-X \geq x)]. \quad (4)$$

(iv) *For any $\epsilon > 0$, all other eigenvalues are of order $n^\epsilon \sqrt{\alpha^\ell}$.*

Before we describe the steps used to establish this, let us verify how it implies Theorem 1.1. Note that since $\mathbf{E}(X) = 1$, writing

$$\mathbf{E}(X) = \int_0^\infty (\mathbf{P}(X \geq x) - \mathbf{P}(-X \geq x)) dx,$$

we see that inequality $\mathbf{P}(X \geq x) - \mathbf{P}(-X \geq x) > 0$ must hold on a set of x 's of positive Lebesgue measure. Since the points x at which the distribution of either X or $-X$ has an atom is at most countable, there thus exists an x at which neither distribution has an atom, and the desired inequality $\mathbf{P}(X \geq x) - \mathbf{P}(-X \geq x) > 0$ holds. Letting $t = x/\sqrt{\mathbf{E}(X^2)}$ and $r = \mathbf{P}(X \geq x) - \mathbf{P}(-X \geq x)$ we readily have by (4) that the empirical overlap in (3) must converge to $\{-r, +r\}$.

Theorem 2.1 will follow from the combination of two analyses. Let \bar{A} denote the expectation of the graph's adjacency matrix conditional on the spin vector σ , that is

$$\bar{A} = \frac{a}{n} \left[\frac{1}{2}(ee' + \sigma\sigma') - I \right] + \frac{b}{2n}(ee' - \sigma\sigma'). \quad (5)$$

The first analysis establishes the following

Theorem 2.2. *Matrix $B^{(\ell)}$ verifies the identity*

$$B^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \bar{A} B^{(m-1)}) - \sum_{m=1}^{\ell} \Gamma^{\ell,m}, \quad (6)$$

for matrices $\Delta^{(\ell)}$, $\Gamma^{\ell,m}$ such that for $\ell = O(\log n)$ and any fixed $\epsilon > 0$, with high probability

$$\rho(\Delta^{(\ell)}) \leq n^\epsilon \alpha^{\ell/2}, \quad (7)$$

$$\rho(\Gamma^{\ell,m}) \leq n^{\epsilon-1} \alpha^{(\ell+m)/2}, \quad m = 1, \dots, \ell. \quad (8)$$

A *local* analysis is then needed to establish properties of the ℓ -neighborhoods of nodes in graph \mathcal{G} . The key quantities in this analysis are the following

$$\begin{aligned} S_t(i) &= |\{j : d_{\mathcal{G}}(i, j) = t\}|, \\ D_t(i) &= \sum_j \mathbf{1}_{d_{\mathcal{G}}(i, j) = t} \sigma_j. \end{aligned} \quad (9)$$

They are close (in a sense made precise in Section 4) to the corresponding quantities $(B^{(t)}e)_i$, $(B^{(t)}\sigma)_i$, and are easier to analyze. In particular, they enjoy a *quasi-deterministic growth* property:

Theorem 2.3. *Under Condition (1) for some suitable constants C and ϵ , with probability $1 - O(n^{-\epsilon})$ the following holds for all $i \in \mathcal{N}$ and $\ell = O(\log n)$:*

$$\begin{aligned} S_t(i) &\leq C \log(n) \alpha^t, \quad t = 1, \dots, \ell, \\ |D_t(i)| &\leq C \log(n) \beta^t, \quad t = 1, \dots, \ell. \end{aligned} \quad (10)$$

$$\begin{aligned} S_t(i) &= \alpha^{t-\ell} S_{\ell}(i) + O(\log(n) + \sqrt{\log(n) \alpha^t}), \quad t = 1, \dots, \ell, \\ D_t(i) &= \beta^{t-\ell} D_{\ell}(i) + O(\log(n) + \sqrt{\log(n) \alpha^t}), \quad t = 1, \dots, \ell. \end{aligned} \quad (11)$$

This, combined with Theorem 2.2, yields the key intermediate step:

Theorem 2.4. *Let $\ell = c \log(n)$, where constant c is such that $c \log(\alpha) < 1/4$. Under Condition (1) the matrix $B^{(\ell)}$ counting simple paths satisfies the following weak Ramanujan property*

$$\sup_{|x|=1, x' B^{(\ell)} e = x' B^{(\ell)} \sigma = 0} |B^{(\ell)} x| \leq n^{\epsilon} \alpha^{\ell/2}. \quad (12)$$

Another key ingredient consists in coupling the neighborhoods of nodes in graph \mathcal{G} with a random tree process, and performing a martingale analysis of this tree process. This is done in Section 4.3. It establishes (see Theorem 4.2) that the vector $(\beta^{-\ell} D_{\ell}(i))$ is close in some sense to a vector $(\sigma_i D_i)$ where the D_i are i.i.d., distributed as the limit of a martingale. This limiting martingale distribution is precisely that of variable X in the statement of Theorem 2.1.

3 Matrix expansion and spectral radii bounds

Our aim in this section is to establish Theorem 2.2. Denoting ξ_{ij} the indicator of edge (i, j) 's presence in \mathcal{G} we can write

$$B_{ij}^{(\ell)} = \sum_{i_0, i_1, \dots, i_{\ell} \in \mathcal{N}} \mathbf{1}_{i_0=i} \mathbf{1}_{i_{\ell}=j} \mathbf{1}_{|\{i_0, \dots, i_{\ell}\}|=\ell+1} \prod_{t=1}^{\ell} \xi_{i_{t-1} i_t}. \quad (13)$$

Denote by P_{ij} the set of all so-called *self-avoiding*, or *simple* paths $i_0^{\ell} := \{i_0, \dots, i_{\ell}\}$ from i to j appearing in the above sum. Let

$$\Delta_{ij}^{(\ell)} := \sum_{i_0^{\ell} \in P_{ij}} \prod_{t=1}^{\ell} (A - \bar{A})_{i_{t-1} i_t} \quad (14)$$

where \bar{A} is as in (5). We then have the expansion:

$$\Delta_{ij}^{(\ell)} = B_{ij}^{(\ell)} - \sum_{m=1}^{\ell} \sum_{i_0^{\ell} \in P_{ij}} \prod_{t=1}^{\ell-m} (A - \bar{A})_{i_{t-1} i_t} \bar{A}_{i_{\ell-m} i_{\ell-m+1}} \prod_{t=\ell-m+2}^{\ell} A_{i_{t-1} i_t}. \quad (15)$$

Introduce the set Q_{ij}^m of paths i_0^ℓ defined by:

$$i_0^\ell \in Q_{ij}^m \Leftrightarrow i_0 = i \ \& \ i_\ell = j \ \& \ |\{i_0, \dots, i_{\ell-m}\}| = \ell - m + 1 \ \& \ |\{i_{\ell-m+1}, \dots, i_\ell\}| = m.$$

Paths in Q_{ij}^m are thus concatenations of simple paths $i_0^{\ell-m}$ and $i_{\ell-m+1}^\ell$. Note that $P_{ij} \subset Q_{ij}^m$. Let R_{ij}^m denote the set difference $Q_{ij}^m \setminus P_{ij}$. It then consists of paths i_0^ℓ such that both $i_0^{\ell-m}$ and $i_{\ell-m+1}^\ell$ are simple, and further verify that the intersection of the corresponding sets is not empty.

Define matrix $\Gamma^{\ell,m}$ as

$$\Gamma_{ij}^{\ell,m} := \sum_{i_0^\ell \in R_{ij}^m} \prod_{t=1}^{\ell-m} (A - \bar{A})_{i_{t-1}i_t} \bar{A}_{i_{\ell-m}i_{\ell-m+1}} \prod_{t=\ell-m+2}^{\ell} A_{i_{t-1}i_t}. \quad (16)$$

With these notations at hand, one obtains from (15):

$$\begin{aligned} \Delta_{ij}^{(\ell)} &= B_{ij}^{(\ell)} - \sum_{m=1}^{\ell} \sum_{i_0^\ell \in Q_{ij}^m} \prod_{t=1}^{\ell-m} (A - \bar{A})_{i_{t-1}i_t} \bar{A}_{i_{\ell-m}i_{\ell-m+1}} \prod_{t=\ell-m+2}^{\ell} A_{i_{t-1}i_t} \\ &\quad + \sum_{m=1}^{\ell} \sum_{i_0^\ell \in R_{ij}^m} \prod_{t=1}^{\ell-m} (A - \bar{A})_{i_{t-1}i_t} \bar{A}_{i_{\ell-m}i_{\ell-m+1}} \prod_{t=\ell-m+2}^{\ell} A_{i_{t-1}i_t} \\ &= B_{ij}^{(\ell)} - \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \bar{A} B^{(m-1)})_{ij} + \sum_{m=1}^{\ell} \Gamma_{ij}^{\ell,m}, \end{aligned} \quad (17)$$

where we noticed that the summation over paths in Q_{ij}^m of the corresponding products yields the (ij) -entry of the product matrix $\Delta^{(\ell-m)} \bar{A} B^{(m-1)}$. This is precisely expansion (6).

We then have the following

Proposition 3.1. *For all integers $k, \ell \geq 1$, it holds that*

$$\mathbf{E} \left[\rho(\Delta^{(\ell)})^{2k} \right] \leq \sum_{v=\ell+1}^{k\ell+1} \sum_{e=v-1}^{k\ell} n^v [(v+1)^2(\ell+1)]^{2k(1+e-v+1)} \left(\frac{\alpha}{n} \right)^{v-1} \left[\frac{\max(a, b)}{n} \right]^{e-v+1}. \quad (18)$$

The proof uses the trace method, bounding $\rho(\Delta^{(\ell)})^{2k}$ by the trace of $(\Delta^{(\ell)})^{2k}$ and a specific encoding of circuits involved in this trace computation. The fact that circuits consist in concatenations of simple paths of length ℓ is then leveraged to control which codes of circuits have to be considered in the trace bound. The details are provided in the Appendix.

Inequality (7) readily follows from Proposition 3.1. Indeed for $\ell = O(\log(n))$ and fixed $\epsilon > 0$, choose an integer $k > 0$ such that $\epsilon > 1/(2k)$. By (18), it holds that

$$\mathbf{E}(\rho(\Delta^{(\ell)})^{2k}) \leq (1 + o(1)) n \alpha^{k\ell} [(k\ell + 2)^2(\ell + 1)]^{2k}.$$

Thus

$$\begin{aligned} \mathbf{P}(\rho \geq n^\epsilon \alpha^{\ell/2}) &\leq \frac{\mathbf{E}(\rho^{2k})}{n^{2k\epsilon} \alpha^{k\ell}} \\ &\leq (1 + o(1)) \frac{n \alpha^{k\ell} [(k\ell + 2)^2(\ell + 1)]^{2k}}{n^{2k\epsilon} \alpha^{k\ell}} \\ &\leq (1 - o(1)) n^{1-2k\epsilon} [(k\ell + 2)^2(\ell + 1)]^{2k} \\ &= o(1), \end{aligned}$$

since we chose k so that $2k\epsilon > 1$ and the last term is polylogarithmic in n . This establishes (7).

We now establish a bound on the spectral radius of the matrix $\Gamma^{\ell,m}$ previously introduced. Specifically, we have

Proposition 3.2. *For all $k, \ell \geq 1$ and $m \in \{1, \dots, \ell\}$ we have the following*

$$\mathbf{E}((\rho(\Gamma^{\ell,m})^{2k}) \leq \sum_{v=m \vee (\ell-m+1)}^{1+k(\ell+m)} \sum_{e=v-1}^{k(\ell+m)} \left(\frac{a \vee b}{n} \right)^{2k+e-v+1} v^{2k} [(v+1)^2(\ell+1)]^{4k(1+e-v+1)} n^v \left(\frac{\alpha}{n} \right)^{v-1}. \quad (19)$$

The proof, postponed to the appendix, again uses the trace method, a specific encoding of circuits involved in the trace bound, and then leverages the constraints on circuits appearing in this bound. It readily implies inequality (8). Indeed for $\ell = O(\log(n))$, and any fixed $\epsilon > 0$, choose $k > 0$ such that $\epsilon > 1/(2k)$. By (19) it holds that

$$\mathbf{E}(\rho(\Gamma^{\ell,m})^{2k}) \leq (1 + o(1)) n \alpha^{k(\ell+m)} \left[\frac{a \vee b}{n} (k(\ell+m) + 2)^5 (\ell+1)^2 \right]^{2k}.$$

Thus

$$\begin{aligned} \mathbf{P}(\rho \geq n^{\epsilon-1} \alpha^{(\ell+m)/2}) &\leq \frac{\mathbf{E}(\rho(\Gamma^{\ell,m})^{2k})}{n^{2k(\epsilon-1)} \alpha^{k(\ell+m)}} \\ &\leq (1 + o(1)) n^{1-2k\epsilon} \left[(a \vee b) (k(\ell+m) + 2)^5 (\ell+1)^2 \right]^{2k}, \end{aligned}$$

and this last bound decays to zero as a power of n by the condition $2k\epsilon > 1$ and the fact that the last term in the product is polylogarithmic in n . This completes the proof of Theorem 2.2.

4 Local Analysis: structure of expanded neighborhoods

This section is devoted to the structure of the local neighborhoods of nodes. We start with general bounds. We then relate vectors of interest $B^{(\ell)}e$ and $B^{(\ell)}\sigma$ to the neighborhood structures. The martingale analysis of neighborhood structures follows.

4.1 Preliminaries

For any $k \geq 0$, the number of nodes with spin \pm at distance k (respectively $\leq k$) of node i is denoted $U_k^\pm(i)$ (respectively, $U_{\leq k}^\pm(i)$). We thus have

$$S_t(i) = U_t^+(i) + U_t^-(i), \quad D_t(i) = U_t^+(i) - U_t^-(i). \quad (20)$$

We shall omit indices i when considering quantities related to a fixed node i . In the remainder of the section we condition on the spins σ of all nodes. We denote n_\pm as the number of nodes with spin \pm .

For fixed $i \in \mathcal{N}$ it is readily seen that, conditionally on $\mathcal{F}_{k-1} := \sigma(U_t^+, U_t^-, t \leq k-1)$, we have:

$$\begin{aligned} U_k^+ &\sim \text{Bin} \left(n_+ - U_{\leq k-1}^+, 1 - (1 - a/n)^{U_{k-1}^+} (1 - b/n)^{U_{k-1}^-} \right), \\ U_k^- &\sim \text{Bin} \left(n_- - U_{\leq k-1}^-, 1 - (1 - a/n)^{U_{k-1}^-} (1 - b/n)^{U_{k-1}^+} \right). \end{aligned} \quad (21)$$

Theorem (2.3) is established based on these characterizations by extensive use of Chernoff bounds for binomial variables. Its proof is deferred to the Appendix.

The next technical result establishes approximate independence of neighborhoods of distinct nodes. It is instrumental in Section 4.3 e.g. in establishing weak laws of large numbers on the fraction of nodes satisfying a given property.

Lemma 4.1. *Consider any two fixed nodes i, j with $i \neq j$. Let $\ell = c \log(n)$ where constant c is such that $c \log(\alpha) < 1/2$. Then the variation distance between the joint law of their neighborhood processes $\mathcal{L}((U_k^\pm(i))_{k \leq \ell}, (U_k^\pm(j))_{k \leq \ell})$ and the law with the same marginals and independence between them, denoted $\mathcal{L}((U_k^\pm(i))_{k \leq \ell}) \otimes \mathcal{L}((U_k^\pm(j))_{k \leq \ell})$, goes to zero as a negative power of n as $n \rightarrow \infty$.*

Proof. Take two independent realizations of the processes $(U_k^\pm(i))_{k \leq \ell}$ and $(U_k^\pm(j))_{k \leq \ell}$. Use them to perform a joint construction of the two processes as follows. Having constructed the sets $\mathcal{U}_t^\pm(i) \subset \mathcal{N}$, $\mathcal{U}_t^\pm(j) \subset \mathcal{N}$ for $t = 1, \dots, k-1$ and assuming the i -sets and the j -sets have not yet met, we construct them at step k as follows. To construct $\mathcal{U}_k^\pm(i)$ we select a size $U_k^\pm(i)$ subset uniformly at random from $\mathcal{N}^\pm \setminus \mathcal{U}_{\leq k-1}^\pm(i)$. We do similarly for j . The construction can proceed based on the independent inputs so long as the resulting i -sets and j -sets do not intersect. However on $\cap_{t \leq k} \{S_t(i) \vee S_t(j) \leq C \log(n)\}$, the expected size of the intersection will be upper-bounded by $O(\log^2(n) \alpha^{2k}/n) = O(\log^2(n) n^{-2\epsilon})$, where $c \log(\alpha) = 1/2 - \epsilon$. The controls in the proof of Theorem 2.3 ensure that the probability of $\cap_{t \leq k} \{S_t(i) \vee S_t(j) \leq C \log(n)\}$ is $1 - O(n^{-\epsilon})$ and the result follows. \square

We now state a lemma on the presence of cycles in the ℓ -neighborhoods of nodes. It will be instrumental in bounding the discrepancy between vectors $B^{(\ell)}e$ (resp. $B^{(\ell)}\sigma$) and $\{S_\ell(i)\}$ (resp. $\{D_\ell(i)\}$). Its proof, deferred to the Appendix, relies on the previous coupling Lemma 4.1.

Lemma 4.2. *Assume $\ell = c \log(n)$ with $c \log(\alpha) < 1/2$. Then with high probability the number of nodes i whose ℓ -neighborhood contains one cycle is $O(\log^4(n) \alpha^{2\ell})$. Assume further that $c \log(\alpha) < 1/4$. Then with high probability no node i has more than one cycle-edge in its ℓ -neighborhood.*

4.2 From neighborhood variables S_t and D_t to path matrix $B^{(\ell)}$

We first state how to transport the deterministic growth controls (11) of Theorem 2.3 to vectors $B^{(m-1)}e$ and $B^{(m-1)}\sigma$, a key step in the proof of Theorem 2.4. One has the following

Lemma 4.3. *Let \mathcal{B} denote the set of nodes i whose ℓ -neighborhood contains a cycle. Then for $m \leq \ell$, $\ell = c \log n$ with $c \log \alpha < 1/4$:*

$$i \notin \mathcal{B} \Rightarrow \begin{cases} (B^{(m-1)}e)_i &= S_{m-1}(i) &= \alpha^{m-1-\ell} (B^{(\ell)}e)_i + O(\log(n)) + O(\sqrt{\log(n) \alpha^{m-1}}), \\ (B^{(m-1)}\sigma)_i &= D_{m-1}(i) &= \beta^{m-1-\ell} (B^{(\ell)}\sigma)_i + O(\log(n)) + O(\sqrt{\log(n) \alpha^{m-1}}), \end{cases} \quad (22)$$

$$i \in \mathcal{B} \Rightarrow |(B^{(m)}\sigma)_i| \leq |(B^{(m)}e)_i| \leq 2 \sum_{t=0}^m S_i(t) = O(\log(n) \alpha^m). \quad (23)$$

Proof is in the Appendix, together with that of the following Corollary:

Corollary 4.1. *For all $m \in \{1, \dots, \ell\}$ it holds with high probability that*

$$\sup_{|x|=1, x' B^{(\ell)}e = x' B^{(\ell)}\sigma = 0} |e' B^{(m-1)}x| = O\left(\log^3(n) \alpha^{\ell+m-1} + \sqrt{n}[\log(n) + \sqrt{\log(n) \alpha^{m-1}}]\right), \quad (24)$$

$$\sup_{|x|=1, x' B^{(\ell)}e = x' B^{(\ell)}\sigma = 0} |\sigma' B^{(m-1)}x| = O\left(\log^3(n) \alpha^{\ell+m-1} + \sqrt{n}[\log(n) + \sqrt{\log(n) \alpha^{m-1}}]\right). \quad (25)$$

We are now ready to prove Theorem 2.4:

Proof. (of Theorem 2.4). Using identity (6), write for unit norm x :

$$|B^{(\ell)}x| \leq \rho(\Delta^{(\ell)}) + \sum_{m=1}^{\ell} \rho(\Delta^{(\ell-m)})|\bar{A}B^{(m-1)}x| + \sum_{m=1}^{\ell} \rho(\Gamma^{\ell,m}).$$

We can ignore the terms $\rho(\Delta^{(\ell)})$ and $\rho(\Gamma^{\ell,m})$, known to be less than $n^\epsilon \alpha^{\ell/2}$ from (7) and (8). Recalling the expression (5) of \bar{A} , one has:

$$|\bar{A}B^{(m-1)}x| \leq \frac{a}{n}|B^{(m-1)}x| + O(n^{-1/2}[|\sigma'B^{(m-1)}x| + |e'B^{(m-1)}x|]).$$

Using the bounds (24,25), the right-hand side is no larger than

$$\frac{a}{n}|B^{(m-1)}x| + O\left(n^{-1/2}\left(\log^3(n)\alpha^{\ell+m-1} + \sqrt{n}[\log(n) + \sqrt{\log(n)\alpha^{m-1}}]\right)\right).$$

By the previous inequalities (10,22,23) and the row sum bound, we have that

$$\rho(B^{(m-1)}) = O(\log(n)\alpha^{m-1}).$$

This thus yields

$$\begin{aligned} |\bar{A}B^{(m-1)}x| &\leq O\left(\frac{\log(n)\alpha^{m-1}}{n} + \frac{\log^3(n)\alpha^{\ell+m-1}}{n^{1/2}} + \log(n) + \sqrt{\log(n)\alpha^{m-1}}\right) \\ &= O(\log(n) + \sqrt{\log(n)\alpha^{m-1}}). \end{aligned}$$

We thus have

$$\begin{aligned} |B^{(\ell)}x| &\leq n^\epsilon \alpha^{\ell/2} + \sum_{m=1}^{\ell} \rho(\Delta^{(\ell-m)})O(\log(n) + \sqrt{\log(n)\alpha^{m-1}}) \\ &\leq n^\epsilon \alpha^{\ell/2} + \sum_{m=1}^{\ell} n^\epsilon \alpha^{(\ell-m)/2} O(\log(n) + \sqrt{\log(n)\alpha^{m-1}}) \\ &\leq n^\epsilon \alpha^{\ell/2} O(1 + 2\ell \log(n)). \end{aligned}$$

The result readily follows. \square

We now state two Lemmas which will allow to establish Theorem 4.1.

Lemma 4.4. *The following evaluations hold whp for $\ell = c \log n$ with $c \log \alpha < 1/4$:*

$$\begin{aligned} |B^{(\ell)}e - \{S_\ell(i)\}_{i \in \mathcal{N}}| &= o(|B^{(\ell)}e|), \\ |B^{(\ell)}\sigma - \{D_\ell(i)\}_{i \in \mathcal{N}}| &= o(|B^{(\ell)}\sigma|), \\ \langle B^{(\ell)}e, B^{(\ell)}\sigma \rangle &= o(|B^{(\ell)}e| \times |B^{(\ell)}\sigma|) \end{aligned} \tag{26}$$

Lemma 4.5. *The following inequalities hold for $\ell = c \log n$ with $c \log \alpha < 1/4$:*

$$\Omega(\alpha^\ell)|B^{(\ell)}e| \leq |B^{(\ell)}B^{(\ell)}e| \leq O(\log^2 n \alpha^\ell)|B^{(\ell)}e| \tag{27}$$

$$\Omega(\beta^\ell)|B^{(\ell)}\sigma| \leq |B^{(\ell)}B^{(\ell)}\sigma| \leq O(\log^4(n)\beta^\ell)|B^{(\ell)}\sigma|. \tag{28}$$

Using these, we now establish the following

Theorem 4.1. *For $\ell = c \log n$ with $c \log \alpha < 1/4$, the two leading eigenvectors of $B^{(\ell)}$ are asymptotically aligned with vectors $\{S_\ell(i)\}$, $\{D_\ell(i)\}$, with corresponding eigenvalues of order up to logarithmic terms $\Theta(\alpha^\ell)$ and $\Omega(\beta^\ell)$. All other eigenvalues are $O(n^\epsilon \sqrt{\alpha}^\ell)$ for any fixed $\epsilon > 0$.*

Proof. Estimates (27–28) and the weak Ramanujan property of Theorem 2.4 imply that the leading eigenvector is aligned with $B^{(\ell)}e$ and has eigenvalue α^ℓ up to logarithmic terms. The second eigenvector is necessarily asymptotically in the span of $\{B^{(\ell)}e, B^{(\ell)}\sigma\}$ and with eigenvalue $\Omega(\beta^\ell)$. By asymptotic orthonormality of vectors $B^{(\ell)}e$ and $B^{(\ell)}\sigma$ and their asymptotic alignment with $\{S_\ell(i)\}$, $\{D_\ell(i)\}$ respectively, the conclusion regarding the first two eigen-elements follows. The bound on the magnitude of other eigenvalues follows from Theorem 2.4 and the Courant-Fisher theorem. \square

4.3 Coupling with Poisson tree growth process

Introduce the stochastic process $\{V_t^\pm\}_{t \geq 0}$ defined by

$$\begin{aligned} V_0^+ &= 1, \quad V_0^- = 0, \\ V_t^+, V_t^- &\text{independent conditionally on } \mathcal{G}_{t-1}, \\ \mathcal{L}(V_t^\pm | \mathcal{G}_{t-1}) &= \text{Poi}((a/2)V_{t-1}^\pm + (b/2)V_{t-1}^\mp) \end{aligned} \quad (29)$$

where $\mathcal{G}_{t-1} = \sigma(V^{\pm k}, k \leq t-1)$. We then have the following

Lemma 4.6. *Let $i \in \mathcal{N}$ be fixed with spin $\sigma_i = \sigma$. For a constant $c > 0$ such that $c \log(\alpha) < 1/2$, and $\ell = c \log(n)$, the following holds. The variation distance between $(U_t^\pm(i))_{t \leq \ell}$ and $(V_t^{\pm\sigma})_{t \leq \ell}$ goes to zero as a negative power of n as $n \rightarrow \infty$.*

The proof given in the Appendix relies on the Stein-Chen method for Poisson approximation. Define now the processes

$$\begin{aligned} M_t &= \alpha^{-t}(V_t^+ + V_t^-), \\ \Delta_t &= \beta^{-t}(V_t^+ - V_t^-), \end{aligned} \quad (30)$$

where V_t^\pm is as defined in (29). We then have the following

Lemma 4.7. *The two processes $\{M_t\}, \{\Delta_t\}$ are \mathcal{G}_t -martingales. Process $\{M_t\}$ is uniformly integrable under Condition $\alpha > 1$. Under Condition $\beta^2 > \alpha$, process $\{\Delta_t\}$ is also uniformly integrable.*

Corollary 4.2. *Under $\alpha < \beta^2$ the martingale $\{\Delta_t\}$ converges almost surely to a unit mean random variable Δ_∞ . Moreover this random variable has a finite variance $1/(\beta^2/\alpha - 1)$ to which the variance of Δ_t converges. It further holds that $\mathbf{E}|\Delta_t^2 - \Delta_\infty^2| \rightarrow 0$ as $t \rightarrow \infty$.*

Together these properties allow to establish the following

Theorem 4.2. *One has the following convergence in probability*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \beta^{-2\ell} D_\ell^2(i) = \mathbf{E}(\Delta_\infty^2). \quad (31)$$

For all $\tau \in \mathbb{R}$ that is a point of continuity of the distribution of both Δ_∞ and $-\Delta_\infty$, one has the following convergence in probability for both signs \pm

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i \in \mathcal{N}: \sigma_i = \pm} \mathbf{1}_{y_i \geq \tau / \sqrt{n \mathbf{E}(\Delta_\infty^2)}} = \frac{1}{2} \mathbf{P}(\pm \Delta_\infty \geq \tau). \quad (32)$$

Let $y \in \mathbb{R}^n$ be the normed vector defined as

$$y_i = \frac{D_\ell(i)}{\sqrt{\sum_{j=1}^n D_\ell(j)^2}}, \quad i = 1, \dots, n. \quad (33)$$

Let x be a vector in \mathbb{R}^n such that we have the convergence in probability

$$\lim_{n \rightarrow \infty} \|x - y\| = 0. \quad (34)$$

Define the spin estimates

$$\hat{\sigma}_i = \begin{cases} + & \text{if } x_i \geq t/\sqrt{n\mathbf{E}(\Delta_\infty^2)}, \\ - & \text{otherwise.} \end{cases} \quad (35)$$

For each t that is an atom of neither Δ_∞ 's or $-\Delta_\infty$'s distribution, the following convergence in probability holds

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \sigma_i \hat{\sigma}_i = \frac{1}{2} (\mathbf{P}(\Delta_\infty \geq t) - \mathbf{P}(-\Delta_\infty \geq t)). \quad (36)$$

To convey the main ideas of the proof (deferred to the Appendix), we now indicate how to establish a property similar to (32), namely for a continuous bounded function g we establish convergence in probability

$$\frac{1}{n} \sum_{i \in \mathcal{N}' \mid \sigma_i = \pm} g(\beta^{-\ell} D_\ell(i)) \rightarrow \frac{1}{2} \mathbf{E}g(\pm \Delta_\infty). \quad (37)$$

The expectation of the sum in the left-hand side reads

$$\frac{n_\pm}{n} \left[\mathbf{E}g(\pm \Delta_\ell) \mathbf{1}_{\text{coupling}} + O(|g|_\infty)(1 - \mathbf{P}(\text{coupling})) \right]$$

where the event $\{\text{coupling}\}$ refers to coupling as in Lemma 4.6. By this Lemma, the martingale convergence property of Δ_ℓ , the fact that $n_\pm/n \rightarrow 1/2$ and continuity of g , this expectation converges to $(1/2)\mathbf{E}g(\pm \Delta_\infty)$. Now the expectation of the square of the empirical average in (37) reads

$$\frac{1}{n^2} \left[n_\pm O(|g|_\infty^2) + 2 \binom{n_\pm}{2} \mathbf{E}g(\beta^{-\ell} D_\ell(i))g(\beta^{-\ell} D_\ell(j)) \right]$$

where $i \neq j$ are two fixed nodes with spin \pm . By the coupling lemma 4.1 it holds that

$$\mathbf{E}g(\beta^{-\ell} D_\ell(i))g(\beta^{-\ell} D_\ell(j)) = \left(\mathbf{E}g(\beta^{-\ell} D_\ell(i)) \right)^2 + O(n^{-\epsilon})|g|_\infty^2.$$

It follows that the variance of the empirical average in (37) goes to zero as $n \rightarrow \infty$. Its announced convergence in probability to $(1/2)\mathbf{E}g(\pm \Delta_\infty)$ then follows by Tchebitchev's inequality.

Theorems 4.1 and 4.2 readily imply Theorem 2.1.

5 Conclusions

The methods developed here may find further applications, e.g. to prove the more general conjecture by Heimlicher et al. [7] of a phase transition in the labeled stochastic block model. More generally one might ask what is the range of applicability of our path expansion approach to “fix” spectral methods by recovering Ramanujan-like spectral separation properties. It is likely that a similar regularization would occur by considering matrix \hat{B} defined by $\hat{B}_{ij} = \mathbf{1}_{d_G(i,j)=\ell}$ but we have not been able to prove this yet.

acknowledgements: The author gratefully acknowledges stimulating discussions on the topic with Marc Lelarge and Charles Bordenave.

References

- [1] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova, “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications,” *Physics Review E*, vol. 84:066106, 2011.
- [2] E. Mossel, J. Neeman, and A. Sly, “Stochastic block models and reconstruction,” Feb. 2012, available at: <http://arxiv.org/abs/1202.1499>.
- [3] A. Coja-oghlan, “Graph partitioning via adaptive spectral techniques,” *Comb. Probab. Comput.*, vol. 19, no. 2, pp. 227–284, 2010. [Online]. Available: <http://dx.doi.org/10.1017/S0963548309990514>
- [4] J. Friedman, “A proof of alon’s second eigenvalue conjecture and related problem,” *Mem. Amer. Math. Soc.*, no. 910., 2008.
- [5] U. Feige and E. Ofek, “Spectral techniques applied to sparse random graphs,” *Random Struct. Algorithms*, vol. 27, no. 2, pp. 251–275, Sept. 2005. [Online]. Available: <http://dx.doi.org/10.1002/rsa.v27:2>
- [6] S. L. Paul W. Holland, Kathryn Blackmond Laskey, “Stochastic blockmodels: First steps,” *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.
- [7] S. Heimlicher, M. Lelarge, and L. Massoulié, “Community detection in the labelled stochastic block model,” Nov. 2012, available at: <http://arxiv.org/abs/1209.2910>.
- [8] W. Evans, C. Kenyon, Y. Peres, and L. Schulman, “Broadcasting on trees and the ising model,” pp. 410–433, 2000.
- [9] D. Williams, *Probability with martingales*. Cambridge University Press, 1991.

A Proof of Proposition 3.1

Proof. We control the spectral radius of $\Delta^{(\ell)}$ using the trace method. Specifically chose $k > 0$. It holds that

$$\mathbf{E}(\rho(\Delta^{(\ell)})^{2k}) \leq \mathbf{E}\text{Tr}(\Delta^{(\ell)})^{2k}. \quad (38)$$

Note that $\text{Tr}(\Delta^{(\ell)})^{2k}$ is the sum over circuits of length $2k$ of the products of the entries $\Delta_e^{(\ell)}$ over the edges e in the circuit. Moreover, given the definition of $\Delta^{(\ell)}$, these correspond to products of entries $A_e - \bar{A}_e$ over edges e of circuits of length $2k\ell$ satisfying the property that consecutive length ℓ -paths are simple.

We bound the expectation of the corresponding sum as follows. Let v (respectively, e) be the number of nodes (respectively, edges) traversed by a particular circuit. The quantity $c = e - v + 1$ is the so-called “tree excess”, counting the number of edges that are traversed while not being part of the tree consisting of edges whose first traversal strictly augments the number of spanned nodes.

We represent the corresponding circuit as follows.

We number nodes by the order in which they are met by the circuit, starting with node 1.

We break each length ℓ -simple path into consecutive sequences consisting of

- a path using only edges already used in the circuit, and lying on the tree of new node discoveries
- a path of discoveries of new nodes
- a cycle edge connecting the end of the two previous steps to a node already spanned. Such a cycle edge may have already been traversed by the circuit one or several times.

Given the tree previously spanned, and the current position on it, the first part of the sequence is characterized by the node label of its end: indeed, since on this subsequence we enforce the condition that the paths are simple, back-tracking is forbidden. Hence there is only one path on the tree going from the origin to the destination. We thus represent the first part by the number of the destination node if this part is non-empty, and by zero otherwise.

The second part of the sequence is simply represented by its length, which is constrained to lie in $\{0, \dots, \ell\}$. Indeed, it cannot exceed ℓ , as we consider sequences that lie within a length ℓ -simple path.

Finally, the third part of the sequence is simply characterized by the number characterizing its end point, and by zero if this part is not present. We must allow for this case, as when we break up a length ℓ -simple path into constituting such sequences, the last such sequence may not end up by traversal of such a redundant edge.

Let us now use this representation to bound the number of corresponding sequences. An individual sequence is represented by a triplet (p, q, r) with $p \in \{0, \dots, v\}$, $q \in \{0, \dots, \ell\}$, and $r \in \{0, \dots, v\}$. Note further that each such sequence corresponds to either the end of an individual length ℓ -simple path, or the traversal of a redundant edge. The number of such edges is $c = e - v + 1$, and each edge can be traversed at most $2k$ times by the constraint that circuits are formed from length ℓ simple paths. Thus the number of valid circuits corresponding to v and e is at most

$$[(v+1)^2(\ell+1)]^{2k(1+e-v+1)}.$$

For a given number of nodes v and edges e , the number of corresponding nodes in $\{1, \dots, n\}$ is upper-bounded by n^v . For a given edge present with multiplicity $m \in \{1, \dots, 2k\}$, the corresponding

expectation is zero if $m = 1$, and for $m \geq 2$, we have

$$\mathbf{E}((\xi_{ij} - \mathbf{E}(\xi_{ij}|\sigma))^m|\sigma) \leq \frac{a(\sigma_i, \sigma_j)}{n}.$$

Here $a(\sigma_i, \sigma_j)$ equals a if $\sigma_i = \sigma_j$ and b otherwise. For the cyclic edges we use the upper bound $\max(a, b)/n$. There are $e - v + 1$ such factors. We are left with a tree with $v - 1$ edges, for which upon averaging over σ we get a contribution $(\alpha/n)^{v-1}$. Now the number of nodes v on any configuration whose contribution in expectation does not vanish must lie between $\ell + 1$ and $k\ell + 1$: indeed each node discovery costs one edge, but this edge must be doubled for the contribution not to vanish. Since there are in total $2k\ell$ edges, at most $k\ell$ nodes can be discovered in addition to the original node of the circuit. The number of distinct edges is similarly bounded by $k\ell$ in any configuration with non-vanishing expectation. This gives the bound

$$\mathbf{E}(\rho(\Delta^{(\ell)})^{2k}) \leq \sum_{v=\ell+1}^{k\ell+1} \sum_{e=v-1}^{k\ell} [(v+1)^2(\ell+1)]^{2k(1+e-v+1)} n^v \left(\frac{\alpha}{n}\right)^{v-1} \left(\frac{a \vee b}{n}\right)^{e-v+1}$$

as announced. \square

B Proof of Proposition 3.2

Proof. By the trace method again, we have that $\rho(\Gamma^{\ell, m})^{2k}$ is upper-bounded by the trace of $[\Gamma^{\ell, m}(\Gamma^{\ell, m})']^k$. The latter corresponds to sum over circuits of length $2\ell k$ of products of terms that can be either $A_e - \bar{A}_e$, \bar{A}_e or A_e . The constraints are that a length ℓ chunk of the circuit is the concatenation of two simple paths of length $m - 1$ and $\ell - m$, and that the two of them have a non-empty intersection.

We represent such contributions as follows. We let v denote the number of nodes traversed by the circuit, and by e the number of edges traversed by the circuit, while ignoring edges that are weighed by an \bar{A} -term. Note that by the constraint that the concatenated simple parts of each length ℓ -chunk intersect, the corresponding graph is necessarily connected. We adopt the following representation of the corresponding circuits.

Nodes are again denoted by the order in which they are first met, starting with node 1. We represent each simple path that constitutes the circuit by sequences of three phases as before. Note that there are now $4k$ such simple paths: each length ℓ chunk of the original circuit is broken into an $m - 1$ - and an $\ell - m$ -path. We adopt the same representation as before, except that we must now also incorporate the label of the starting point after traversal of an \bar{A} -edge.

Thus we have the upper bound on the number of valid circuit labels with v nodes and e edges:

$$v^{2k} [(v+1)^2(\ell+1)]^{4k(1+e-v+1)}.$$

Let us bound the values that v and e can take. Necessarily, $v \geq \max(m, \ell - m + 1)$: indeed, each length ℓ chunk comprises simple paths of length $m - 1$ and $\ell - m$. Moreover, there are overall $2k(\ell - 1)$ edges (recall that we discount the \bar{A} -edges). Out of these, $2k(\ell - m)$ must be doubled for the expectation not to vanish. There are thus at most $1 + k(\ell + m)$ nodes v in total, and at most $k(\ell + m)$ distinct edges in total.

We thus obtain the upper bound

$$\mathbf{E}[(\rho(\Gamma^{\ell, m})^{2k}] \leq \sum_{v=m \vee (\ell-m+1)}^{1+k(\ell+m)} \sum_{e=v-1}^{k(\ell+m)} \left(\frac{a \vee b}{n}\right)^{2k} v^{2k} [(v+1)^2(\ell+1)]^{4k(1+e-v+1)} n^v \left(\frac{\alpha}{n}\right)^{v-1} \left(\frac{a \vee b}{n}\right)^{e-v+1}.$$

□

C Proof of Theorem 2.3

The following inequality is easily verified to hold for any non-negative U, V, a, b, n such that $a/n, b/n \leq 1$, and will be instrumental in the sequel:

$$\frac{aU + bV}{n} - \frac{1}{2} \left(\frac{aU + bV}{n} \right)^2 \leq 1 - (1 - a/n)^U (1 - b/n)^V \leq \frac{aU + bV}{n}. \quad (39)$$

Next lemma is the key ingredient to establish Theorem 2.3.

Lemma C.1. *The following properties hold with high probability for all $i \in \mathcal{N}$ and all $t \leq \ell$, with $\ell = C \log(n)$.*

Let $T = \inf\{t \leq \ell : S_t \geq K \log(n)\}$ for some constant K . Then $S_T = \Theta(\log(n))$.

Let $\epsilon_t := \epsilon \alpha^{-(t-T)/2}$ for some constant $\epsilon > 0$. Then for all $t, t' \in \{T, \dots, \ell\}$, $t > t'$, the vector $U_t = (U_t^+, U_t^-)$ verifies the coordinate-wise bounds:

$$U_t \in \left[\prod_{s=t'+1}^t (1 - \epsilon_s) M^{t-t'+1} U_{t'}, \prod_{s=t'+1}^t (1 + \epsilon_s) M^{t-t'+1} U_{t'} \right], \quad (40)$$

where M denotes the matrix $(a/2 \ b/2, b/2 \ a/2)$.

Proof. Recall that conditionally on \mathcal{F}_{t-1} the random variables U_t^+ and U_t^- are independent, distributed according to

$$U_t^\pm \sim \text{Bin} \left(n^\pm - U_{<t}^\pm, 1 - (1 - a/n)^{U_{t-1}^\pm} (1 - b/n)^{U_{t-1}^\mp} \right).$$

Let T be the first instant t for which $U_t \geq K \log(n)$, for some K to be specified.

By definition of T , necessarily $U_{T-1} < K \log(n)$. Thus

$$U_T^\pm \leq \text{Bin}(n^\pm, (a \vee b) \frac{K \log(n)}{n}).$$

The mean of the Binomial distribution in the right-hand side of the above is equivalent to $(a \vee b)(1/2)K \log(n)$ and less than $\kappa \log(n)$ for $\kappa = (a \vee b)K$. Hence by Chernoff's inequality, for $h(x) := x \log(x) - x + 1$,

$$\mathbf{P}(U_T^\pm \leq K'/2 \log(n) | \mathcal{F}_{T-1}) \leq e^{-\kappa \log(n) h(K'/2\kappa)}.$$

Take K' so that $\kappa h(K'/2\kappa) > 2$. The right-hand side of the above is then no larger than n^{-2} .

Thus properties (10) clearly hold for $t \leq T$. We now establish that they hold with sufficiently large probability for larger t .

Conditional on \mathcal{F}_T , the binomial distribution of U_{T+1}^\pm has mean

$$[n^\pm - U_{<T+1}^\pm] \times [1 - (1 - a/n)^{U_T^\pm} (1 - b/n)^{U_T^\mp}].$$

Using the inequalities (39) we obtain that this mean lies in the interval

$$[(a \wedge b) \frac{1}{3} K \log(n), (a \vee b) K' \log(n)].$$

For a given $\epsilon > 0$, we can choose K sufficiently large so that

$$(a \wedge b) \frac{1}{3} K h(1 + \epsilon) > 2.$$

It follows that U_{T+1}^\pm admits a relative deviation from its conditional mean by ϵ with probability at most n^{-2} .

We now define

$$\epsilon_t = \epsilon \alpha^{-(t-T)/2},$$

and consider the events $\mathcal{A}_t := \{U_t^\pm \in [1 - \epsilon_t, 1 + \epsilon_t]^{\frac{aU_{t-1}^\pm + bU_{t-1}^\mp}{2}}\}$. Conditionally on $\mathcal{A}_T, \dots, \mathcal{A}_t$, the vector $U_t = (U_t^+, U_t^-)$ verifies the announced inequality (40). Given that α is the spectral radius of M , it follows from this condition that $U_t^\pm \geq (1 - O(\epsilon)) \alpha^{t-T} K'' \log(n)$. We then check that Chernoff's bound applies to show that the condition holds at step t with high enough probability. It suffices to ensure that

$$U_t^\pm \tilde{h}(\epsilon_t) \geq 2 \log(n),$$

where we take $\tilde{h}(u) := \min[(1+u) \log(1+u) - u, (1-u) \log(1-u) + u]$. However as we just saw the left-hand side of this expression is lower-bounded by

$$(1 - O(\epsilon)) \alpha^{t-T} K'' \log(n) \tilde{h}(\epsilon_t) \geq (1 - O(\epsilon)) \alpha^{t-T} K'' \log(n) K''' \epsilon_t^2,$$

where we took a second-order expansion of \tilde{h} around 0. The condition is therefore met as soon as $(1 - O(\epsilon)) K'' K''' \epsilon^2 \geq 2$. For K large enough this holds. \square

Proof. (of Theorem 2.3). For $t \leq \ell$, if $t \leq T$, we necessarily have that $S_t, |D_t| = O(\log n)$. Consider then $t > T$. Note that matrix M is such that

$$M^k = \frac{1}{2} \begin{pmatrix} \alpha^k + \beta^k & \alpha^k - \beta^k \\ \alpha^k - \beta^k & \alpha^k + \beta^k \end{pmatrix}.$$

Using (40), we readily have for $t, t' \leq T$, with $t > t'$:

$$\begin{aligned} S_t &\leq \prod_{s=t'+1}^t (1 + \epsilon_s) (1, 1) M^{t-t'} U_{t'} \\ &= \prod_{s=t'+1}^t (1 + \epsilon_s) \alpha^{t-t'} S_{t'}. \end{aligned}$$

A similar lower bound holds with $-\epsilon_s$ in place of $+\epsilon_s$. Setting $t' = T$ in the upper bound, since $S_T = O(\log(n))$, the upper bound (10) follows for S_t , as $\prod_{s=T+1}^t (1 + \epsilon_s) = O(1)$.

Note now that

$$\max \left(\prod_{s=t'+1}^t (1 + \epsilon_s) - 1, 1 - \prod_{s=t'+1}^t (1 - \epsilon_s) \right) = O(\epsilon_{t'}) = O(\alpha^{-t'/2}).$$

It readily follows that (11) holds for S_t .

Consider now D_t . Using (40) again, we have

$$D_{t-1} - \alpha \epsilon_t S_t \leq D_t \leq \beta D_{t-1} + \alpha \epsilon_t S_t.$$

Iterating, we obtain

$$|D_t - \beta^{t-t'} D_{t'}| \leq \sum_{s=t'+1}^t \alpha \beta^{t-s} \epsilon_s S_s. \quad (41)$$

Since $S_s = O(\log(n)\alpha^{s-T})$, $|D_T| = O(\log(N))$ and $\epsilon_s = O(\alpha^{-(s-T)/2})$, we obtain for $t' = T$:

$$|D_t| = O(\log(n)\beta^t + \sum_{s=T+1}^t \beta^{t-s} \log(n)\alpha^{(s-T)/2}) = O(\log(n)\beta^t),$$

where we have used the assumption that $\beta^2 > \alpha$ to bound $\sum_{u>0} \beta^{-u}\alpha^{u/2}$. Property (10) thus holds for D_t .

Finally, the right-hand side of (41) is of order

$$\sum_{s=t'+1}^t \beta^{t-s} \alpha^{(s-T)/2} \log(n) = O(\log(n)\beta^{t-t'}\alpha^{t'/2}).$$

Thus setting $t = \ell$, for $\ell > t' \geq T$ we have

$$D_{t'} = \beta^{t'-\ell} D_\ell + O(\log(n)\alpha^{t'/2}).$$

Since for $t' < T$ we readily have $D_{t'} = O(\log(n))$ by definition of T , property (11) follows for D_t . \square

D Proof of Lemma 4.2

Proof. There are two ways for creating cycles within the distance k -neighborhood of i : an edge may be present between two nodes at distance $k-1$ of i , or two nodes at distance $k-1$ may be connected to the same node at distance k of i . The number of edges of the first type is stochastically dominated by $\text{Bin}(S_{k-1}^2, a \vee b/n)$. Its expected number conditionally on $\Omega_{k-1}(i)$, defined as

$$\Omega_{k-1}(i) = \{S_{k-1}(i) \leq C(\log n)\alpha^{k-1}\}$$

is at most $O(\log^2(n)\alpha^{2\ell}/n)$. Thus by the union bound the probability that there is such an edge in the ℓ -neighborhood of i is at most:

$$\ell \times O(\log^2(n)\alpha^{2\ell}/n) + \sum_{k=1}^{\ell} (1 - \mathbf{P}(\Omega_k(i))) = O(\log^3(n)\alpha^{2\ell}/n).$$

As for the second type of cycles, its number is stochastically dominated by

$$\text{Bin}(n, (a \vee b/n)^2 S_{k-1}^2).$$

On $\Omega_{k-1}(i)$ its conditional expectation is $O(\log^2(n)\alpha^{2\ell})$.

By the same argument, the probability that there are two cycle-edges within the ℓ -neighborhood of i is upper-bounded by $O(\log^6(n)\alpha^{4\ell}/n^2)$. By the union bound we readily have that with high probability no node has two cycle-edges within its ℓ -neighborhood as soon as $\log^6(n)\alpha^{4\ell} \ll n$, which holds for $\ell = c \log(n)$ with $c \log(\alpha) < 1/4$.

Let Z_i denote the event that the ℓ -neighborhood of i contains a cycle. On the event

$$\cap_{i \in \mathcal{N}} \cap_{k \leq \ell} \Omega_k(i),$$

the ℓ -neighborhoods of an arbitrary pair of distinct nodes i, j are disjoint with probability $1 - O(\log^2(n)\alpha^{2\ell}/n)$, conditionally upon which the probability that they both have a cycle in their

neighborhood is upper-bounded by $O(\log^6(n)\alpha^{4\ell}/n^2)$. Conditionally on the event that their neighborhoods meet, the expectation of the product $Z_i Z_j$ is still upper-bounded by $O(\log^3(n)\alpha^{2\ell}/n)$.

Eventually Markov's inequality yields

$$\begin{aligned} \mathbf{P}(\sum_i Z_i \geq m \log^3(n)\alpha^{2\ell}) &\leq \frac{\mathbf{E}(\sum_i Z_i)^2}{m^2 \log^6(n)\alpha^{4\ell}} \\ &\leq \frac{n\mathbf{E}(Z_1) + n^2\mathbf{E}(Z_1 Z_2)}{m^2 \log^6(n)\alpha^{4\ell}} \\ &\leq \frac{O(\log^3(n)\alpha^{2\ell}) + n^2[O(\log^6(n)\alpha^{4\ell}/n^2) + (\log^2(n)\alpha^{2\ell}/n)(\log^3(n)\alpha^{2\ell}/n)]}{m^2 \log^6(n)\alpha^{4\ell}} \\ &= O\left(\frac{1}{m^2}\right). \end{aligned}$$

Taking $m = \log(n)$ (say), then with high probability $\sum_i Z_i = O(\log^4(n)\alpha^{2\ell})$. \square

E Proof of Lemma 4.3

Consider first those indices $i \notin \mathcal{B}$ whose ℓ -neighborhood is a tree. For any k and any $m \leq \ell$, $B_{ik}^{(m)}$ can take values only 0 or 1 (there is at most one simple path connecting i to k), and $B_{ik}^{(m)} = 1$ if and only if $d_{\mathcal{G}}(i, k) = m$. For such i , one therefore has the following identities:

$$i \notin \mathcal{B} \Rightarrow \begin{cases} (B^{(m)}e)_i = S_i(m), \\ (B^{(m)}\sigma)_i = D_i(m), \end{cases} \quad (42)$$

Relations (22) readily follows from Theorem 2.3. Let us now consider an index $i \in \mathcal{B}$ whose ℓ -neighborhood is not a tree. We know from Lemma 4.2 that such nodes have in their ℓ -neighborhood only one cycle. Clearly only nodes at distance at most ℓ of i can be counted in $(B^{(\ell)}e)_i$, and they can be counted at most twice because the neighborhood contains only one cycle. Control (23) readily follows.

F Proof of Corollary 4.1

Proof. Let \mathcal{B} denote the set of nodes i such that their ℓ -neighborhood contains a cycle. Let x be a normed vector such that $x'B^{(\ell)}e = 0$. We then have

$$\begin{aligned} |e'B^{(m-1)}x| &= |\sum_{i \in \mathcal{N}} x_i (B^{(m-1)}e)_i| \\ &\leq |\sum_{i \in \mathcal{B}} x_i (B^{(m-1)}e)_i| \\ &\quad + |\sum_{i \in \mathcal{B}} x_i \alpha^{m-1-\ell} (B^{(\ell)}e)_i| \\ &\quad + |\sum_{i \in \mathcal{N}} x_i [\alpha^{m-1-\ell} (B^{(\ell)}e)_i + O(\log(n) + \sqrt{\log(n)\alpha^{m-1}})]|. \end{aligned} \quad (43)$$

Using the bound (23) for $i \in \mathcal{B}$, we can bound the first summation, using Cauchy-Schwarz's inequality by

$$\begin{aligned} |\sum_{i \in \mathcal{B}} x_i (B^{(m-1)}e)_i| &\leq O(\log(n)\alpha^{m-1})\sqrt{|\mathcal{B}|} \\ &\leq O(\log^3(n)\alpha^{\ell+m-1}), \end{aligned}$$

where we have used the bound on the size of \mathcal{B} derived in Lemma 4.2. The second summation in (43) is similarly bounded. As for the third summation, using the fact that $e'B^{(\ell)}x = 0$, it is upper-bounded by

$$|\sum_{i \in \mathcal{N}} x_i O(\log(n) + \sqrt{\log(n)\alpha^{m-1}})|.$$

By Cauchy-Schwarz again, this is no larger than

$$O\left(\sqrt{n}(\log(n) + \sqrt{\log(n)\alpha^{m-1}})\right).$$

The announced bound (24) on $|e'B^{(m-1)}x|$ follows. Similarly, the bound (25) on $|\sigma'B^{(m-1)}x|$ follows by using property $\sigma'B^{(\ell)}x = 0$ instead of property $e'B^{(\ell)}x = 0$. \square

G Proof of Lemma 4.6

Proof. We assume that $\sigma_i = +$, the case $\sigma = -$ being similar. Introduce the events

$$\Omega_k = \{S_k \leq C \log(n)\alpha^k\}, \quad k \geq 1, \quad (44)$$

where constant C is as in Theorem 2.3. As established in the proof of Theorem 2.3, the probability of each Ω_k is $1 - o(n^{-2})$.

Let us evaluate, conditionally on \mathcal{F}_{k-1} and on Ω_{k-1} the variation distance between (U_k^+, U_k^-) and a pair of (conditionally on \mathcal{F}_{k-1}) independent random variables with respective distributions

$$W_k^+ \sim \text{Poi}\left(\frac{aU_{k-1}^+ + bU_{k-1}^-}{2}\right), \quad W_k^- \sim \text{Poi}\left(\frac{aU_{k-1}^- + bU_{k-1}^+}{2}\right).$$

The Stein-Chen method enables to bound the variation distance between a $\text{Bin}(n, \lambda/n)$ and a $\text{Poi}(\lambda)$ random variables by $n \min(1, \lambda^{-1})(\lambda/n)^2 \leq \lambda/n$. Furthermore, two Poisson random variables with respective parameters λ, λ' have variation distance at most $|\lambda - \lambda'|$. This entails the bounds

$$\begin{aligned} d_{\text{var}}(U_k^+, W_k^+) &\leq \left[1 - (1 - a/n)^{U_{k-1}^+} (1 - b/n)^{U_{k-1}^-}\right] \\ &\quad + \left|(n_+ - U_{\leq k-1}^+)[1 - (1 - a/n)^{U_{k-1}^+} (1 - b/n)^{U_{k-1}^-}] - \frac{aU_{k-1}^+ + bU_{k-1}^-}{2}\right|. \end{aligned}$$

We now use (39) to obtain

$$\begin{aligned} d_{\text{var}}(U_k^+, W_k^+) &\leq \frac{aU_{k-1}^+ + bU_{k-1}^-}{n} \\ &\quad + |n_+ - U_{\leq k-1}^+ - n/2|[1 - (1 - a/n)^{U_{k-1}^+} (1 - b/n)^{U_{k-1}^-}] \\ &\quad + \frac{1}{4n}(aU_{k-1}^+ + bU_{k-1}^-)^2. \end{aligned}$$

Let us now specify constant c such that $c \log(\alpha) < 1/2$, i.e. there is $\epsilon > 0$ such that $c \log(\alpha) \leq 1/2 - \epsilon$. For $k \leq \ell = c \log(n)$, on Ω_{k-1} , it holds that $S_{k-1} \leq C \log(n)n^{1/2-\epsilon}$. This, together with the bound $|n/2 - n_+| \leq O(\log(n)n^{1/2})$ ensures the upper bound

$$d_{\text{var}}(U_k^+, W_k^+) \leq O(\log(n)n^{-\epsilon}).$$

The same bound is readily established for the variation distance $d_{\text{var}}(U_k^-, W_k^-)$. These bounds readily establish by induction that the variation distance between the two sequences $(U_k^\pm)_{k \leq \ell}$ and $(V_k^\pm)_{k \leq \ell}$ are upper-bounded by $O(\log^2(n)n^{-\epsilon})$, establishing the Lemma. \square

H Proof of Lemma 4.7

Proof. Write

$$\begin{aligned}\mathbf{E}(V_{t+1}^\pm | \mathcal{G}_t) &= \frac{a}{2} V_t^\pm + \frac{b}{2} V_t^\mp \\ &= \frac{a}{2} \frac{\alpha^t M_t \pm \beta^t \Delta_t}{2} + \frac{b}{2} \frac{\alpha^t M_t \mp \beta^t \Delta_t}{2} \\ &= \alpha^{t+1} M_t \pm \beta^{t+1} \Delta_t.\end{aligned}$$

It readily follows that both processes $\{M_t\}$, $\{\Delta_t\}$ are martingales. To establish uniform integrability we shall show that both processes have uniformly bounded variance. To that end we use the conditional variance formula

$$\text{Var}(X) = \text{Var}(\mathbf{E}(X|\mathcal{F})) + \mathbf{E}(\text{Var}(X|\mathcal{F})),$$

and the fact that the variance of a Poisson random variable equals its mean. Thus

$$\mathbf{E}(V_t^\pm | \mathcal{G}_{t-1}) = \text{Var}(V_t^\pm | \mathcal{G}_{t-1}) = \frac{aV_{t-1}^\pm + bV_{t-1}^\mp}{2}.$$

This yields by the conditional variance formula

$$\begin{aligned}\text{Var}(M_t) &= \text{Var}(M_{t-1}) + \mathbf{E}(\alpha^{-t} M_{t-1}) \\ &= \text{Var}(M_{t-1}) + \alpha^{-t}.\end{aligned}$$

Since $\text{Var}(M_0) = 0$, it follows by induction that

$$\text{Var}(M_t) = \frac{1 - \alpha^{-t}}{\alpha - 1}, \quad t > 0.$$

The latter is uniformly bounded for $\alpha > 1$ hence the uniform integrability of $\{M_t\}$ under this condition.

Write now

$$\begin{aligned}\text{Var}(\Delta_t) &= \text{Var}(\Delta_{t-1}) + \mathbf{E}(\beta^{-2t} \text{Var}(V_t^+ - V_t^- | \mathcal{G}_{t-1})) \\ &= \text{Var}(\Delta_{t-1}) + \mathbf{E}(\beta^{-2t} \alpha^t M_{t-1}) \\ &= \text{Var}(\Delta_{t-1}) + \beta^{-2t} \alpha^t.\end{aligned}$$

It thus follows by $\text{Var}(\Delta_0) = 0$ and induction that

$$\text{Var}(\Delta_t) = \frac{1 - (\alpha/\beta^2)^t}{\beta^2/\alpha - 1}, \quad t > 0,$$

thus establishing uniform integrability of martingale $\{\Delta_t\}$. □

I Proof of Corollary 4.2

Proof. Convergence almost surely and in L_1 is guaranteed under uniform integrability by the martingale convergence theorem ([9]). Finiteness of the limiting variable's variance under uniform bounds on the variance is also standard; it follows from Fatou's lemma. Convergence of the variances is established as follows. The limiting variable satisfies a distributional equation given by

$$\Delta = \beta^{-1} \left(\sum_{i=1}^{\text{Poi}(a/2)} \Delta_i - \sum_{i=1}^{\text{Poi}(b/2)} \Delta'_i \right) \quad (45)$$

where the Δ_i, Δ'_i are i.i.d. and distributed as Δ . The only solution for the variance of Δ , apart from the degenerate solution 0, is then readily seen to be $1/(\beta^2/\alpha - 1)$, which is indeed the limit of the variance of Δ_t . The L_1 -convergence of Δ'_t to Δ_∞^2 is then a direct consequence of Scheffé's lemma. \square

J Proof of Theorem 4.2

Proof. Note that with probability of order $1 - O(n^{-\epsilon})$ for fixed positive ϵ , $\sigma(i)\beta^{-\ell}D_\ell(i)$ coincides with Δ_ℓ by the coupling lemma 4.6. When this coupling fails, by the bounds established in Theorem 2.3, it holds that $\beta^{-\ell}D_\ell(i)$ is $O(\log(n))$. This entails that the left-hand side of (31) verifies

$$\mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n \beta^{-2\ell} D_\ell^2(i) \right) = O(\log^2(n))n^{-\epsilon} + \mathbf{E}(\Delta_\ell^2 \mathbf{1}_{\text{coupling}}).$$

Write

$$|\mathbf{E}(\Delta_\ell^2 \mathbf{1}_{\text{coupling}}) - \mathbf{E}(\Delta_\infty^2)| \leq \mathbf{E}|\Delta_\ell^2 - \Delta_\infty^2| + \mathbf{E}(\Delta_\infty^2 \mathbf{1}_{\text{no coupling}}).$$

By Corollary 4.2, we have that the first term in the right-hand side goes to zero with ℓ ; the second term goes to zero with the probability that coupling fails (e.g. using Hardy-Littlewood-Polya's rearrangement inequalities). Thus the expectation converges to $\mathbf{E}\Delta_\infty^2$.

Let us now consider the second moment of the empirical sum:

$$\mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n \beta^{-2\ell} D_\ell^2(i) \right)^2.$$

We break it into two terms, the first being

$$\frac{1}{n^2} \mathbf{E} \sum_{i=1}^n \beta^{-4\ell} D_\ell^4(i).$$

Using Lemma 4.6 and Theorem 2.3, using similar arguments as before we can bound this term by

$$\frac{1}{n} O(\log^2(n)) [\mathbf{E}(\Delta_\infty^2) + o(1)]$$

which clearly goes to zero as $n \rightarrow \infty$.

The second term is given by

$$\frac{2}{n^2} \sum_{i < j} \beta^{-4\ell} \mathbf{E}(D_\ell^2(i) D_\ell^2(j)).$$

For given $i < j$, using Lemma 4.1 and Theorem 2.3, we bound the product $D_\ell^2(i) D_\ell^2(j)$ by $O(\log^4(n) \beta^{4\ell})$ on the event that the coupling with independent copies fails, and by $\mathbf{E}(D_\ell^2(i)) \mathbf{E}(D_\ell^2(j))$ on the event that it succeeds. We then bound each of the individual terms in this product as in the control of the expectation done in the first half of the proof, thus obtaining the upper bound for the second moment:

$$o(1) + \frac{2}{n^2} \sum_{i < j} [O(n^{-2\epsilon} \log^4(n)) + (\mathbf{E}(\Delta_\infty^2))^2 + o(1)].$$

It readily follows that

$$\lim_{n \rightarrow \infty} \mathbf{E} \left[\frac{1}{n} \sum_{i=1}^n \beta^{-2\ell} D_\ell^2 - \mathbf{E}(\Delta_\infty^2) \right]^2 = 0.$$

The convergence in probability (31) follows.

We now turn to establishing (32). We shall only consider the case of sign +, the other being handled similarly. Fix some arbitrarily small $\delta > 0$. Because τ is a continuity point of the distribution of Δ_∞ , we can find two bounded Lipschitz-continuous functions f, g such that

$$f(u) \leq \mathbf{1}_{u \geq \tau} \leq g(u), \quad u \in \mathbb{R}$$

and

$$0 \leq \mathbf{E}(g(\Delta_\infty) - f(\Delta_\infty)) \leq \delta.$$

Consider then the empirical sum

$$\frac{1}{n} \sum_{i \in n_+} f(y_i \sqrt{n \mathbf{E}(\Delta_\infty^2)}).$$

Writing

$$y_i = \frac{\beta^{-\ell} D_\ell(i)}{\sqrt{\sum_{j=1}^n \beta^{-2\ell} D_\ell^2(j)}},$$

we have that this empirical sum differs from the simpler one

$$\frac{1}{n} \sum_{i \in n_+} f(\beta^{-\ell} D_\ell(i)) \tag{46}$$

by at most

$$K \frac{1}{n} \sum_{i \in n_+} \beta^{-\ell} |D_\ell(i)| \times \left| 1 - \sqrt{\frac{\mathbf{E}(\Delta_\infty^2)}{A}} \right|,$$

where K is the Lipschitz continuity constant of function f and A is the empirical sum in (31). This correction tends to zero in probability by dominated convergence. Indeed, convergence to zero of $1 - \sqrt{\mathbf{E}(\Delta_\infty^2)/A}$ has just been established. By similar arguments as before based on Theorem 2.3 and Lemmas 4.6 and 4.1, the empirical average of the $|\beta^{-\ell} D_\ell(i)|$ is bounded. Convergence in probability of (46) to $(1/2)\mathbf{E}(f(\Delta_\infty))$ is then established by evaluating the first and second moments of this sum as previously done.

The same argument can be applied to g , eventually leading to the convergence in probability

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i \in n_+} g(y_i \sqrt{n \mathbf{E}(\Delta_\infty^2)}) = \frac{1}{2} \mathbf{E}(g(\Delta_\infty)).$$

It readily follows that

$$\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i \in n_+} \mathbf{1}_{y_i \geq \tau / \sqrt{n \mathbf{E}(\Delta_\infty^2)}} - \frac{1}{2} \mathbf{P}(\Delta_\infty \geq \tau) \right| \leq \delta.$$

As δ is arbitrary, this establishes (32).

Pick again an arbitrary $\delta > 0$, two pairs of Lipschitz-continuous functions f_{\pm} and g_{\pm} such that

$$f_{\pm}(u) \leq \mathbf{1}_{\pm u \geq t} \leq g_{\pm}(u),$$

and

$$0 \leq \mathbf{E}(g_{\pm}(\pm \Delta_{\infty})) - \mathbf{E}(f_{\pm}(\pm \Delta_{\infty})) \leq \delta.$$

The empirical overlap (36) reads

$$\frac{1}{n} \left[\sum_{i \in n_+} - \sum_{i \in n_-} \right] \left(2 \cdot \mathbf{1}_{x_i \sqrt{n \mathbf{E}(\Delta_{\infty}^2)} \geq t} - 1 \right).$$

The difference $(n_+ - n_-)/n$ is of order $1/\sqrt{n}$ and thus vanishes. We upper-bound the remaining terms by

$$\frac{2}{n} \left[\sum_{i \in n_+} g_+(x_i \sqrt{n \mathbf{E}(\Delta_{\infty}^2)}) - \sum_{i \in n_-} f_-(x_i (-\sqrt{n \mathbf{E}(\Delta_{\infty}^2)})) \right]$$

Letting K denote the Lipschitz-continuity constant for both g_+ and f_- , this last display differs from

$$\frac{2}{n} \left[\sum_{i \in n_+} g_+(\beta^{-\ell} D_{\ell}(i)) - \sum_{i \in n_-} f_-(-\beta^{-\ell} D_{\ell}(i)) \right] \quad (47)$$

by at most

$$\frac{2K}{n} \sum_{i=1}^n \left| (x_i - y_i) \sqrt{n \mathbf{E}(\Delta_{\infty}^2)} + \beta^{-\ell} |D_{\ell}(i)| \right| \times \left| 1 - \sqrt{\frac{\mathbf{E}(\Delta_{\infty}^2)}{A}} \right|. \quad (48)$$

Because of the assumed convergence in probability $\lim_{n \rightarrow \infty} \|x - y\| = 0$, the first error term necessarily tends to zero in probability by Cauchy-Schwarz inequality. The second term is dealt with as mentioned in the proof of the previous lemma. Finally, using the coupling lemmas 4.6 and 4.1, by evaluating the first and second moments of (47), we obtain the convergence in probability

$$\lim_{n \rightarrow \infty} \frac{2}{n} \left[\sum_{i \in n_+} g_+(\beta^{-\ell} D_{\ell}(i)) - \sum_{i \in n_-} f_-(-\beta^{-\ell} D_{\ell}(i)) \right] = \mathbf{E}(g_+(\Delta_{\infty}) - f_-(\Delta_{\infty})).$$

The latter term is then an upper bound on the limsup of the empirical overlap. By the same approach, we obtain a lower bound of

$$\mathbf{E}(f_+(\Delta_{\infty}) - g_-(\Delta_{\infty}))$$

on the liminf of the overlap. These upper and lower bounds differ by at most 2δ , and differ from $\mathbf{P}(\Delta_{\infty} \geq t) - \mathbf{P}(\Delta_{\infty} \leq -t)$ by at most δ . Since δ is arbitrary, this establishes the announced convergence in probability of the empirical overlap to quantity x where

$$x = \mathbf{P}(\Delta_{\infty} \geq t) - \mathbf{P}(\Delta_{\infty} \leq -t)$$

is strictly positive by our choice of t . □

K Proof of Lemma 4.4

Proof. The first and second evaluations follow by noting that the vectors whose difference is considered in the left-hand side agree on the set of entries i whose ℓ -neighborhood is cycle-free. Thus

$$\begin{aligned} |B^{(\ell)}e - \{S_\ell(i)\}| &\leq \sqrt{|B|}O(\log(n)\alpha^\ell) \\ &\leq O(\log^3(n)\alpha^{2\ell}), \end{aligned}$$

and the same bound holds for $|B^{(\ell)}\sigma - \{D_\ell(i)\}|$. This upper bound is $o(\sqrt{n}\beta^\ell)$ so that the first two assertions follow, by further noticing that $|\{D_\ell(i)\}| = \Theta(\beta^\ell)$, as follows from combining Theorem 2.3 with Lemmas 4.6 and 4.1 as in the proof of Theorem 4.2.

For the third assertion, consider the scalar product $\langle \{S_\ell(i)\}, \{D_\ell(i)\} \rangle$. Using the same arguments as in the proof of Theorem 4.2 we obtain that this scalar product is indeed $o(|\{S_\ell(i)\}| \times |\{D_\ell(i)\}|)$. \square

L Proof of Lemma 4.5

Proof. To establish the lower bound of (27), note that by Cauchy-Schwarz,

$$\langle e, B^{(\ell)}B^{(\ell)}e \rangle \leq |e| \times |B^{(\ell)}B^{(\ell)}e|.$$

However the left-hand side reads $|B^{(\ell)}e|^2$. Thus

$$|B^{(\ell)}B^{(\ell)}e| \geq \frac{|B^{(\ell)}e|^2}{|e|}.$$

However it holds that $|B^{(\ell)}e| = \Theta(\sqrt{n}\alpha^\ell)$ (this follows from the methods in the proof of Theorem 4.2). Since $|e| = \sqrt{n}$, the lower bound in (27) follows. For the upper bound, we note that by Lemma 4.3 and Theorem 2.3, the max row sum for matrix $B^{(\ell)}$ is of order $O(\log(n)\alpha^\ell)$.

The lower bound in (28) is established similarly, from the inequality

$$\langle \sigma, B^{(\ell)}B^{(\ell)}\sigma \rangle \leq |\sigma| \times |B^{(\ell)}B^{(\ell)}\sigma|$$

The upper bound requires additional arguments. Assuming the 2ℓ -neighborhood of i is cycle-free, the i -th entry of vector $B^{(\ell)}B^{(\ell)}\sigma$ can be written as

$$\sum_{d=0}^{\ell} \sum_{j: d_G(i,j)=2d} \sigma_j |\{k : d_G(i,k) = d_G(j,k) = \ell\}|.$$

We control the magnitude of this quantity in the tree model; using coupling we will then transpose the corresponding estimates to the original scenario.

Let then \mathcal{T} denote a branching process with offspring $\text{Poi}(\alpha)$. The process of spins is then constructed by sampling uniformly the root's spin, and then propagating spins in a Markovian fashion with transition matrix $(a/(a+b)b(a+b), b(a+b), a(a+b))$ that is $\alpha^{-1}M$. Its eigenvalues are thus $(1, \beta/\alpha)$.

The variable of interest is written

$$X = \sum_{d=0}^{\ell} \sum_{j: d(j,i)=2d} \sigma_j |\{k : d(j,k) = d(i,k) = \ell\}|.$$

We evaluate its second moment conditionally on \mathcal{T} by writing X^2 as

$$X^2 = \sum_{d=0}^{\ell} \sum_{d'=0}^{\ell} \sum_{j': d(j', i)=2d'} \sum_{j: d(j, i)=2d} \sigma_j \sigma_{j'} \times \\ \times |\{k : d(j, k) = d(i, k) = \ell\}| \cdot |\{k' : d(j', k') = d(i, k') = \ell\}|.$$

Now it holds that

$$\mathbf{E}(\sigma_j \sigma_{j'} | \mathcal{T}) = O\left(\left(\frac{\beta}{\alpha}\right)^{d(j, j')}\right).$$

We will use this formula, and further distinguish nodes j' according to their distance $2(d + d' - \tau)$ for $\tau = 0, \dots, 2(d \wedge d')$. This yields

$$\mathbf{E}(X^2 | \mathcal{T}) = \sum_{d, d'=0}^{\ell} \sum_{\tau=0}^{2(d \wedge d')} \sum_{j': d(j', i)=2d'} \sum_{j: d(j, i)=2d} O\left(\left(\frac{\beta}{\alpha}\right)^{2(d+d'-\tau)}\right) \times \\ \times |\{k : d(j, k) = d(i, k) = \ell\}| \cdot |\{k' : d(j', k') = d(i, k') = \ell\}|.$$

Note now that with high probability, we have the following evaluations

$$\begin{aligned} |\{k : d(j, k) = d(i, k) = \ell\}| &= \tilde{O}(\alpha^{\ell-d}), \\ |\{k' : d(j', k') = d(i, k') = \ell\}| &= \tilde{O}(\alpha^{\ell-d'}), \\ |\{j : d(j, i) = 2d\}| &= \tilde{O}(\alpha^{2d}), \\ |\{j' : d(j', i) = 2d' \& d(j, j') = 2(d + d' - \tau)\}| &= \tilde{O}(\alpha^{2d'-\tau}). \end{aligned}$$

Plugging these in, we have

$$\begin{aligned} \mathbf{E}(X^2 | \mathcal{T}) &= \sum_{d, d'=0}^{\ell} \sum_{\tau=0}^{2(d \wedge d')} \tilde{O}\left(\left(\frac{\beta}{\alpha}\right)^{2(d+d'-\tau)}\right) \alpha^{2\ell-d-d'+2(d+d')-\tau} \\ &= \sum_{d, d'=0}^{\ell} \sum_{\tau=0}^{2(d \wedge d')} \tilde{O}\left(\alpha^{2\ell} \left(\frac{\beta^2}{\alpha}\right)^{d+d'-\tau}\right) \\ &= \tilde{O}(\alpha^{2\ell} (\beta^2/\alpha)^{2\ell}) \\ &= \tilde{O}(\beta^{4\ell}). \end{aligned}$$

By coupling (techniques of Theorem 4.2 involving Tchebitchev inequality, based on the bounds of Theorem 2.3 and Lemmas 4.6 and 4.1) we thus have that with high probability,

$$|B^{(\ell)} B^{(\ell)} \sigma| = \tilde{O}(\sqrt{n\beta^{4\ell}}) = \tilde{O}(\beta^{\ell} |B^{(\ell)} \sigma|)$$

as announced. □