THE REGULAR STOCHASTIC BLOCK MODEL ON SEVERAL-COMMUNITY NETWORKS

SAYAR KARMAKAR, MOUMANTI PODDER

Abstract. This paper studies the regular stochastic block model comprising several communities: each of the $k$ non-overlapping communities, for $k \geq 3$, possesses $n$ vertices, each of which has total degree $d$. The values of the intra-cluster degrees (i.e. the number of neighbours of a vertex inside the cluster it belongs to) and the inter-cluster degrees (i.e. the number of neighbours of a vertex inside a cluster different from its own) are allowed to vary across clusters. We discuss three main results. The first of these compares the probability measure induced by our model with the uniform measure on the space of $d$-regular graphs on $kn$ vertices; the second establishes that the clusters, under rather weak assumptions, are unique asymptotically almost surely as $n \to \infty$; the third shows that efficient weak recovery of the clusters is possible under suitable assumptions on the eigenvalues of the $k \times k$ matrix of intra-cluster and inter-cluster degrees.

1. Introduction

This paper concerns itself with the regular stochastic block model, henceforth abbreviated as RSBM, that is used to study clustered networks. These networks exhibit community structure, whereby the individuals participating in the network, typically indicated as nodes or vertices of a graph, are split into overlapping or non-overlapping groups, usually with dense connections internally and sparser connections between different groups. Community structure is common in many complex networks such as computer and information networks ([30]), online social networks and biological networks ([21, 10, 16, 38]) that include protein-protein and gene-gene interactions ([34]), biological neural networks ([20]), metabolic networks ([54]) etc. Detecting communities in clustered networks has been pursued with fervour ([44, 45, 46, 5, 43, 27]), since communities often act as meta-nodes in a network and individuals within the same community tend to exhibit behavioural and functional similarities, simplifying the analysis of the underlying features of the network. The characteristics displayed by each distinct community may also vary greatly from the average properties of the network. The existence of communities may also significantly affect the spreading of rumours, epidemics etc. within the network.

The stochastic block model, henceforth abbreviated as SBM, (introduced in [23], surveyed in [1]), has been the most popular model, so far, in studying clustered networks. In its most simplified form, this model comprises $2n$ vertices that are partitioned into two equi-sized clusters. Edges between all pairs of vertices appear mutually independently, with probability $p$ if both vertices belong to the same cluster, and probability $q$ if they belong to different clusters. Letting the intra-cluster average...
degree be \( a \sim \rho n \) and the inter-cluster average degree be \( b \sim q_n \), \( \text{(37)} \) and \( \text{(3)} \) studied the SBM in the regime where \( a, b = O(\log n) \), whereas \( \text{(11), (13), (14), (39), (42), (40), (41)} \) and \( \text{(33)} \) studied SBM in the regime where \( a, b = O(1) \), as \( n \) grows to \( \infty \). Other variants of this model studied are the Bayesian SBM \( \text{(48)} \), degree-corrected block models \( \text{(51)} \), labeled SBM \( \text{(29, 22)} \), SBM in sparse hypergraphs \( \text{(47)} \) etc. We also refer the reader to \( \text{(15), (26), and (12)} \) for discussions on relations between community detection in SBM and the minimum bisection problem that seeks to partition a graph of \( 2n \) vertices into two equi-sized parts such that the number of edges across the parts is minimized.

Since the essence of our paper is to focus on the case where the given model constitutes several underlying clusters, we emphasize on the following developments in the literature. In \( \text{(13)} \), it was conjectured that if the signal-to-noise ratio of a given SBM is strictly higher than 1, then it is possible to detect communities in polynomial time, or, in other words, the well-known Kesten-Stigum threshold is achieved; moreover, if the number of underlying communities in the model exceeds 4, it is possible to detect the communities information-theoretically for some signal-to-noise ratio strictly lower than 1. It was shown in \( \text{(7)} \) that the Kesten-Stigum threshold is achieved in SBM’s with multiple communities satisfying certain asymmetry assumptions, whereas the full conjecture of \( \text{(13)} \), for several clusters, was established in \( \text{(2)} \). The extension of SBM from two to several communities has proven to be a veritable challenge during the course of development of this field.

The RSBM was introduced in \( \text{(8)} \), although two regular versions of the SBM in the sparse regime was proposed in \( \text{(33)} \). As in \( \text{(8)} \), we assume that each intra-cluster degree and each inter-cluster degree exceeds 3, ensuring that the resulting graph is connected with high probability. The RSBM differs from the SBM with constant average intra-cluster and inter-cluster degrees in that the latter has a positive probability of possessing isolated vertices. In RSBM, the imposition of the constraint that each vertex has a constant number of neighbours in each given cluster gives more structure to the graph, but at the same time, robs the model of the edge-independence that is present in SBM.

We now highlight the novelties as well as describe the organization of our paper. First and foremost, we emphasize that our model is far more general than that of \( \text{(8)} \) in that it takes into account multiple communities as well as intra-cluster and inter-cluster degree values that differ across communities. We answer similar questions as those in \( \text{(8)} \), but the proof techniques, despite bearing similarities in a few places, are more involved and require more careful analysis. To set the stage, in \( \text{(1.1)} \) we describe the notations and terminology used throughout the paper; in \( \text{(1.2)} \) we describe the model and its underlying measure in details; in \( \text{(1.3)} \) we describe the well-known configuration model and the associated exploration process, and their importance in the generation of uniformly random regular or bipartite-regular graphs.

In \( \text{(2)} \) we show that the measure induced by RSBM on \( kn \) vertices, each with degree \( d \), where \( k \) denotes the number of communities and \( n \) the number of members in each community, is distinct from the measure that makes a uniformly random selection out of the collection of all \( d \)-regular graphs on \( kn \) vertices. In \( \text{(3)} \) we show that under rather weak assumptions, the underlying clusters of the model are unique almost surely as \( n \) approaches \( \infty \). We draw attention of the reader to a key difference between our analysis and the analysis in \( \text{(8), (3.2.2)} \): while they had the symmetry, around 1/2, of the binary entropy function \( H(\alpha) = -\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha) \) in their favour due to the presence of only two clusters, we require a somewhat different strategy to handle the higher number of clusters in our model. Even in the homogeneous case, where all the intra-cluster degrees are the same, there is need for a thorough case-by-case analysis that is much more intricate than in \( \text{(8)} \). We emphasize here that in the homogeneous case, our analysis allows for the inter-cluster degrees to exceed, by far, the intra-cluster degrees. This is a significant generalization.
over the much more usual assumption of denser intra-cluster connections and sparser inter-cluster connections found in the literature. In the heterogeneous scenario, we need a more restricted range of intra-cluster and inter-cluster degrees, as described in [3.1] and [3.2], Theorem 3.1. Finally, in §4 we discuss our result pertaining to the recovery of the clusters of RSBM. The much more complicated implementation of the principal ideas of [Lemma 1, §3] in our set-up is not to be missed. Unlike [3], it no longer suffices to consider the second largest eigenvalue alone, but rather requires consideration of the $k$ largest eigenvalues (counting multiplicities), and the corresponding eigenvectors, of a suitable self-avoiding matrix associated with the RSBM graph.

1.1. Notations. Given $n, d \in \mathbb{N}$, we denote by $\mathcal{R}_d^n$ the set of all $d$-regular graphs on $n$ labeled vertices, and by $\mathcal{B}_d^n$ the set of all $d$-bipartite-regular graphs on $2n$ labeled vertices where each cluster comprises $n$ vertices. We shall denote by $\mu_d^n$ the uniform measure on $\mathcal{R}_d^n$.

Given a graph $G$, we denote by $V(G)$ its vertex set and by $E(G)$ its edge set. Given $S \subset V(G)$ and $v \in V(G)$, we let $\deg_S(v)$ denote the number of edges $\{u, v\}$ where $u \in S$. We denote by $G|_S$ the subgraph of $G$ that is induced on $S$. For disjoint subsets $S_1$ and $S_2$ of $V(G)$, we denote by $\deg(S_1, S_2)$ the number of edges $\{u_1, u_2\}$ where $u_1 \in S_1$ and $u_2 \in S_2$. We let $G|_{S_1, S_2}$ denote the subgraph with vertex set $S_1 \cup S_2$ and edge set $\{\{u_1, u_2\} : u_1 \in S_1, u_2 \in S_2\}$. Given $v \in V(G)$ and $r \in \mathbb{N}$, we let $B(v, r) = \{u \in V(G) : \rho(u, v) \leq r\}$ be the neighbourhood of radius $r$ around $v$, where $\rho$ is the usual graph metric. We let $\delta B(v, r) = \{u \in V(G) : \rho(u, v) = r\}$ denote the boundary of $B(v, r)$.

Given an infinite sequence of graphs $\{G_n\}$ and a graph property $A$, we say that $A$ holds asymptotically almost surely (a.a.s.) for this sequence if $\mathbb{P}[G_n \text{ satisfies property } A] \to 1$ as $n \to \infty$.

For any $\alpha \in (0, 1)$, recall that the Shannon entropy for a Bernoulli($\alpha$) distribution is given by $H(\alpha) = -\alpha \log_2 \alpha - (1 - \alpha) \log_2(1 - \alpha)$. This will be used in §3.

The rest of §1 will be relevant in §4. Given a graph $G$ with $V(G) = \{v_1, \ldots, v_n\}$, a path between vertices $v_i$ and $v_j$ is a sequence $(v_{i_0}, v_{i_1}, \ldots, v_{i_s})$ for some positive integer $s$, such that $v_{i_0} = v_i$, $v_{i_s} = v_j$, and $\{v_{i_{t-1}}, v_{i_t}\} \in E(G)$ for each $1 \leq t \leq s$. We call a path self-avoiding if the vertices $v_{i_0}, v_{i_1}, \ldots, v_{i_{s-1}}, v_{i_s}$ are all distinct. For any $s \in \mathbb{N}$, we define the length-$s$ self-avoiding matrix $S^{(s)}$ by setting its $(i, j)$-th entry $S_{i, j}^{(s)}$ to be the number of self-avoiding paths of length $s$ between $v_i$ and $v_j$. Note that $S^{(1)}$ is the adjacency matrix of $G$.

A vector $u = (u_1, \ldots, u_m)$ in $\mathbb{R}^m$ is said to be a unit vector if $\|u\|_2 = \sqrt{\sum_{i=1}^m u_i^2} = 1$. For two vectors $u$ and $v$ in $\mathbb{R}^m$, we let $u \perp v$ indicate that $u$ is orthogonal to $v$, i.e. the scalar product $\langle u, v \rangle = \sum_{i=1}^m u_i v_i = 0$. Consider two sequences $\{u_n\}_n$ and $\{w_n\}_n$ of unit vectors such that both $u_n$ and $w_n$ belong to $\mathbb{R}^m$ for some $m_n \in \mathbb{N}$, for each $n$. We say that they are asymptotically aligned if $\lim_{n \to \infty} \langle u_n, w_n \rangle = 1$. Given a sequence of unit vectors $\{u_n\}_n$ and a sequence of subspaces $\{S_n\}_n$ where $u_n \in \mathbb{R}^{m_n}$ and $S_n$ is a subspace of $\mathbb{R}^{m_n}$ for some $m_n \in \mathbb{N}$, we say that $\{u_n\}_n$ asymptotically belongs to $\{S_n\}_n$ if there exists a sequence of unit vectors $\{w_n\}_n$, with $w_n \in S_n$ for each $n$, such that $\{u_n\}_n$ and $\{w_n\}_n$ are asymptotically aligned.

1.2. Description of the model. Our $k$-cluster RSBM, denoted $\mathcal{G}_A^n$, has the following parameters:

(i) $n$ denotes the number of vertices in each cluster,
(ii) $k$ denotes the number of clusters,
(iii) and $A = (A_{i,j})_{1 \leq i,j \leq k}$ is a $k \times k$ symmetric matrix of strictly positive integers such that, for some $d \in \mathbb{N}$,

$$\sum_{j=1}^k a_{i,j} = d \text{ for all } i = 1, \ldots, k. \quad (1.1)$$
Starting with \( kn \) labeled vertices, we uniformly randomly partition them into \( k \) clusters \( C_1, \ldots, C_k \), each of size \( n \). Independent of each other, we now place on the vertices of \( C_i \) a uniformly random member of \( R_{a_{i,i}}^n \), and across the clusters \( C_i \) and \( C_j \) a uniformly random member of \( B_{a_{i,j}}^n \), for all \( i, j \in \{1, \ldots, k\} \) with \( i \neq j \). The criterion in (1.1) ensures that any realization of our model will be \( d \)-regular. All the parameters except \( n \) remain fixed throughout our analysis. We analyze the asymptotic behaviour of the model as \( n \to \infty \).

1.3. Configuration model and exploration process. The configuration model plays a crucial role as a tool in our arguments in [3] and [4]. Given \( d, n \in \mathbb{N} \) such that \( dn \) is even, this model (see [3]) allows us to generate a \( d \)-regular random graph on \( n \) labeled vertices \( v_1, \ldots, v_n \) (possibly with self-loops and parallel edges) according to the following procedure, also known as the exploration process:

(i) Fix a total order \( v_1 < v_2 < \cdots < v_n \) on the vertex set, and let \( \Xi_i = \{ \xi_{i,j} : 1 \leq j \leq d \} \) denote the set of half-edges emanating from \( v_i \), for all \( i = 1, \ldots, n \). Let \( \Xi = \bigcup_{i=1}^{n} \Xi_i \). We define a total ordering on \( \Xi \) as follows: all half-edges in \( \Xi_i \) come before every half-edge in \( \Xi_{i+1} \) for all \( i = 1, \ldots, n-1 \), and within each \( \Xi_i \), we have \( \xi_{i,j} < \xi_{i,j+1} \) for all \( 1 \leq j \leq d-1 \).

(ii) We first choose \( \xi \) uniformly randomly from the set \( \Xi \setminus \{ \xi_{1,1} \} \) and form the edge \( \{ \xi, \xi_{1,1} \} \).

Having constructed the first \( k \) edges, we find the smallest half-edge \( \xi_{i,j} \) yet unmatched with another half-edge, and choosing a \( \xi \) uniformly randomly from the remaining subset of half-edges, we form the edge \( \{ \xi_{i,j}, \xi \} \). Thus we form a perfect matching on \( \Xi \).

We also describe here the exploration process aimed at generating a random \( d \)-regular bipartite graph in which each cluster contains \( n \) vertices. In this case, we label the vertices of one partition as \( u_1 < u_2 < \cdots < u_n \) and the other as \( v_1 < v_2 < \cdots < v_n \). We let \( \Xi_i = \{ \xi_{i,j} : 1 \leq j \leq d \} \) denote the set of half-edges emanating from \( u_i \) and \( \Gamma_i = \{ \gamma_{i,j} : 1 \leq j \leq n \} \) the set of half-edges emanating from \( v_i \), and the total orderings on \( \Xi = \bigcup_{i=1}^{n} \Xi_i \) and on \( \Gamma = \bigcup_{i=1}^{n} \Gamma_i \) are analogous to the one described above. We first choose, uniformly randomly, a \( \gamma \) out of \( \Gamma \), and form the edge \( \{ \gamma, \xi_{1,1} \} \).

After having constructed the \( k \)-th edge, we find the smallest \( \xi_{i,j} \) in \( \Xi \) that is yet to be matched with a half-edge from \( \Gamma \). We choose, uniformly randomly, a half-edge \( \xi \) from \( \Gamma \) that has not yet been matched, and form the edge \( \{ \xi, \xi_{i,j} \} \). This leads to a perfect matching between \( \Xi \) and \( \Gamma \).

It has been shown in [5] that in either of the cases above, the probability that the generated random graph is simple, i.e. devoid of self-loops and parallel edges, stays bounded away from 0 as \( d \) stays bounded and \( n \to \infty \). We can thus condition on the event that the generated random graph is simple, which in turn allows us to prove all results of [3] and [4] using the exploration process.

Henceforth, we call a half-edge emanating from a vertex in \( C_i \) and matched with a half-edge from a vertex in \( C_j \), a half-edge of type \( \{i, j\} \), for all \( i, j \in \{1, \ldots, k\} \).

2. Comparing RSBM with uniform measure on \( d \)-regular graphs on \( kn \) vertices

We state here the first of our three main results. Let \( \mu_A^n \) denote the probability measure of \( G_A^n \) and \( S_A^n \) the support of \( \mu_A^n \). Recall from [1] that \( \mu_{d}^{kn} \) denotes the uniform random measure on \( R_{d}^{kn} \).

**Theorem 2.1.** Under the above set-up, keeping the matrix \( A \) fixed, we have

\[
\lim_{n \to \infty} \left\| \mu_A^n, \mu_d^{kn} \right\|_{TV} = 1, \tag{2.1}
\]

where \( TV \) denotes the total variation distance between two probability measures.
The proof begins with stating two well-known results. For given \( n \) and \( d \) with \( 1 \leq d = o\left(n^{1/2}\right) \), [36, Corollary 5.3] states that
\[
|R_d^n| = C \frac{(nd)!}{(nd/2)!(nd/2)(d!)^n},
\]
where \( C = C(n,d) \) remains bounded as \( n \) grows. Similarly, [35, Theorem 2] states that
\[
|B_d^n| = C' \frac{(dn)!}{(d!)^n},
\]
where \( C' = C'(n,d) \) remains bounded as \( n \) grows. Given \( kn \) labeled vertices, we choose the vertex sets for the clusters \( C_1, \ldots, C_k \) in
\[
\binom{n}{k} \left( \binom{k-1}{n} \right) \left( \frac{2n}{n} \right) \frac{(kn)!}{(n!)^k} \sim \frac{\sqrt{2\pi(kn)}}{\left\{2\pi n \left(\frac{n}{e}\right)^n\right\}^k} \sqrt{\frac{k}{2\pi n}} \frac{k^{kn}}{(n-1)^{kn/2}} = \Theta \left( \frac{k^{kn}}{n^{(k-1)/2}} \right)
\]
many ways. The number of possible \( a_{i,i} \)-regular graphs on \( C_i \), for each \( i = 1, \ldots, k \), equals, by (2.2),
\[
\Theta \left( \frac{(na_{i,i})!}{(na_{i,i}/2)!2^{na_{i,i}/2}(a_{i,i})^{na_{i,i}/2}} \right) = \Theta \left( \frac{\sqrt{2\pi(a_{i,i})}}{\left\{2\pi a_{i,i} \left(\frac{a_{i,i}}{e}\right)^{a_{i,i}/2}\right\}^{2n}} \frac{a_{i,i}^{a_{i,i}/2}}{a_{i,i}^{a_{i,i}/2} \cdot a_{i,i}^{e-a_{i,i}/2} \cdot 2^{a_{i,i}/2} \left(2\pi a_{i,i}\right)^{a_{i,i}/2} \cdot a_{i,i}^{e-a_{i,i}/2}} \right) = \Theta \left( \frac{n^{a_{i,i}/2} e^{a_{i,i}/2}}{a_{i,i}^{a_{i,i}/2} (2\pi)^{a_{i,i}/2}} \right).
\]
Similarly, by (2.3), the number of possible \( a_{i,j} \)-bipartite-regular graphs across clusters \( C_i \) and \( C_j \), for each \( i \neq j \) and \( i, j \in \{1, \ldots, k\} \), is given by
\[
\Theta \left( \frac{(a_{i,j})!}{(a_{i,j})!} \right) = \Theta \left( \frac{\sqrt{2\pi(a_{i,j})}}{\left\{2\pi a_{i,j} \left(\frac{a_{i,j}}{e}\right)^{a_{i,j}/2}\right\}^{2n}} \frac{a_{i,j}^{a_{i,j}/2} n^{a_{i,j}/2} e^{a_{i,j}/2}}{a_{i,j}^{a_{i,j}/2} (2\pi)^{a_{i,j}/2} e^{a_{i,j}/2}} \right) = \Theta \left( \frac{(a_{i,j})!}{(2\pi)^{a_{i,j}/2} e^{a_{i,j}/2}} \right).
\]
Therefore, combining these estimates, the total number of possible realizations of \( G_A^n \) on a given set of \( kn \) labeled vertices becomes
\[
\Theta \left( \frac{k^{kn}}{n^{(k-1)/2}} \prod_{i=1}^{k} \frac{n^{a_{i,i}/2} e^{a_{i,i}/2}}{a_{i,i}^{a_{i,i}/2} (2\pi)^{a_{i,i}/2} e^{a_{i,i}/2}} \prod_{i<j} \left( \frac{n^{a_{i,j}/2} e^{a_{i,j}/2}}{a_{i,j}^{a_{i,j}/2} (2\pi)^{a_{i,j}/2} e^{a_{i,j}/2}} \right) \right) = \Theta \left( \frac{k^{kn} n^{kd/2} e^{kd/2} k^{kn/2} (2(k^2 - 3k + 2)/4) d^{kd/2}}{n^{(k-1)/2} \prod_{i<j} a_{i,j}^{a_{i,j}/2} (2\pi)^{kd/2} e^{kd/2}} \right).
\]
On the other hand, from (2.4), the number of \( d \)-regular graphs on \( kn \) vertices is
\[
\Theta \left( \frac{(kd)!}{(kd/2)!2^{kd/2}(d!)^n} \right) = \Theta \left( \frac{\sqrt{2\pi(kd)}}{\left\{2\pi (kd/2) \left(\frac{kd}{e}\right)^{kd/2}\right\}^{2n}} \frac{d^{kd/2}}{(2\pi)^{d/2} e^{d/2}} \right) = \Theta \left( \frac{(kd)^{kd/2} e^{kd/2}}{(2\pi)^{kd/2} d^{kd/2}} \right).
\]
Theorem 3.1. Suppose the model described in Setting $x_i$ for each $i \in \{1, \ldots, m\}$ with \(\sum_{i=1}^{m} x_i = 1\), we have

\[
\prod_{i=1}^{m} x_i^{\alpha_i} \geq \left(\sum_{i=1}^{m} \alpha_i \right)^{-1}. \tag{2.6}
\]

Setting $x_j = a_{i,j}$ and $\alpha_j = \frac{a_{i,j}}{d}$ for all $j \in \{1, \ldots, k\}$, from (2.4), we get

\[
\prod_{j=1}^{k} a_{i,j}^{a_{i,j}/d} \geq \frac{d}{k} \implies \prod_{j=1}^{k} a_{i,j}^{a_{i,j}/d} \geq \frac{d^d}{k^d}
\]

for each $i = 1, \ldots, k$. This in turn yields

\[
\left\{ \prod_{i,j} a_{i,j}^{a_{i,j}} \right\}^{n/2} \geq \frac{d^{knd/2}}{k^{n/2}}. \tag{2.7}
\]

We also observe that given positive integers $x_1, \ldots, x_k$ for any $k \in \mathbb{N}$, the following inequality holds:

\[
k \prod_{i=1}^{k} x_i \geq \sum_{i=1}^{k} x_i. \tag{2.8}
\]

From (2.1), (2.5), (2.7) and (2.8), we see that

\[
\mu_d^k (S_A^n) \leq \Theta \left( \frac{k^{kn} n^{knd/2} e^{knd/2} n^{(k^2-3k+2)/4} \prod_{i,j \in \{a_{i,j} + 1\}/2} (2\pi)^n k^{n/2}}{(knd)^{knd/2} e^{knd/2}} \right)
\]

\[
\leq \Theta \left( \left\{ \prod_{i,j} a_{i,j}^{a_{i,j}} \right\}^{n/2} \left( \prod_{i,j \in \{a_{i,j} + 1\}/2} (2\pi)^n k^{n/2} \right) \right)
\]

\[
\leq \Theta \left( \left\{ \prod_{i,j} a_{i,j}^{a_{i,j}} \right\}^{n/2} \frac{(k^3)}{(2\pi)^k} \right) = \Theta \left( \left\{ \frac{k^3}{(2\pi)^k} \right\}^{n/2} \right). \tag{2.9}
\]

The ratio \(\frac{k^3}{(2\pi)^k}\) is strictly less than 1 for all $k$ that satisfy \(\frac{k^3}{\log k} > \frac{2}{\log(2\pi)}\). Now, the function \(f(x) = \frac{x-1}{\log x}\) is strictly increasing in $x$ for all $x \geq 2$, and \(f(3) = \frac{2}{\log 3} > \frac{3}{\log(2\pi)}\). This shows that the ratio \(\frac{k^3}{(2\pi)^k}\) is strictly less than 1 for all $k \geq 3$, thus showing that the bound in (2.9) is $o(1)$.

3. Almost sure uniqueness of clusters in RSBM

We now state the second of our three main results.

**Theorem 3.1.** Suppose the model described in Setting satisfies the following conditions:

(i) There exists a positive constant $C$ such that $Ca_{i,i} > B_i$ for each $i \in \{k\}$, where $B_i = \max \left\{ a_{i,\ell} : \ell \in \{1, \ldots, k\} \setminus \{i\} \right\}$.

(ii) When not all intra-cluster degrees are equal, for $i, j \in \{1, \ldots, k\}$ with $a_{i,i} < a_{j,j}$, there exist constants $\delta_{i,j} \in (0, \frac{1}{4})$, independent of all entries of $A$, such that

\[
a_{i,i} \geq \left( \frac{1}{2} + 2\delta_{i,j} \right) a_{j,j}. \tag{3.1}
\]
Moreover, there exist constants $\epsilon_{i,j} \in \left(0, \frac{1}{2}\right)$, independent of the entries of $A$, such that for all $i, j \in \{1, \ldots, k\}$ with $a_{i,i} > a_{j,j}$ and $B_j > a_{i,i} - a_{j,j}$,

$$\left(\frac{1}{2} - 2\epsilon_{i,j}\right)a_{i,i} \geq B_j. \quad (3.2)$$

Then, for all sufficiently large values of the entries of the matrix $A$, the clusters $C_i$, $i = 1, \ldots, k$, are a.a.s. unique as $n \to \infty$ while the matrix $A$ stays fixed.

From the discussion in [1.3] it suffices to establish Theorem 3.1 on the random multigraph in which each of the intra-cluster regular graphs and inter-cluster bipartite-regular graphs is generated via the configuration model. Fix any non-negative $\alpha_1, \ldots, \alpha_k$ such that $\sum_{i=1}^k \alpha_i = 1$, and subsets $C_i$ of $C$ such that $|C_i| = \alpha_i n$ for $i = 1, \ldots, k$. Let $D = \bigcup_{i=1}^k C_i$. To prove Theorem 3.1, it is enough to establish Proposition 3.2. We note here that although Proposition 3.2 is stated for $a_{1,1}$, its proof will be analogous if we replace $a_{1,1}$ by any $a_{i,i}$ for $i \in \{2, \ldots, k\}$.

**Proposition 3.2.** Assume that the hypotheses of Theorem 3.1 hold, and that the entries of $A$ are sufficiently large. Suppose there exist at least two distinct $i, j \in \{1, \ldots, k\}$ such that $\alpha_i$ and $\alpha_j$ are strictly positive. Then a.a.s. the following cannot be true simultaneously:

(i) the subgraph $G|D$ is $a_{1,1}$-regular;

(ii) there exists a partition of $V(G) \setminus D$ into subsets $D_2, \ldots, D_k$, each of size $n$, such that $G|_{D_i}$ is $a_{i,i}$-regular, $G|_{D_i, D_j}$ is $a_{i,j}$-bipartite-regular and $G|_{D_i, D_j}$ is $a_{i,j}$-bipartite-regular for all distinct $i, j \in \{2, \ldots, k\}$.

In order to prove Proposition 3.2, we start with the assumption that $G|D$ is $a_{1,1}$-regular. The proof requires consideration of a few different cases depending on the values of the $\alpha_i$'s, and these are addressed in Lemma 3.3, 3.1, 3.2 and 3.3.

**Lemma 3.3.** If $a_{i,i} > a_{1,1}$ and $\alpha_i > \frac{B_1}{a_{i,i} - a_{1,1} + B_1}$, then the conclusion of Proposition 3.2 holds.

**Proof.** Assume that $\alpha_i > 0$ for some $i$ such that $a_{i,i} > a_{1,1}$, and that Proposition 3.2 does not hold. As $G|D$ is $a_{1,1}$-regular, we have $\deg_{C_i \setminus C_j}(v) \geq (a_{i,i} - a_{1,1})$ for every $v \in C_i$, yielding $\deg(C_i, C_j \setminus C_i) \geq (a_{i,i} - a_{1,1})\alpha_i n$. On the other hand, each $u \in C_i \setminus C_j$ belongs to precisely one of the remaining clusters $D_2, \ldots, D_k$. If $u \in D_j$, then $\deg_{C_i}(u) \leq \deg_D(u) = a_{1,j}$. Thus $\deg(C_i, C_j \setminus C_i) \leq B_1(1 - \alpha_i)n$. These two inequalities together yield

$$(a_{i,i} - a_{1,1})\alpha_i \leq B_1(1 - \alpha_i) \implies \alpha_i \leq \frac{B_1}{a_{i,i} - a_{1,1} + B_1},$$

thus completing the proof. \qed

From here onward, we only consider those $\ell \in \{1, \ldots, k\}$ such that $\alpha_\ell > 0$, without mentioning so every time. We shall let $i$ denote that index in $\{1, \ldots, k\}$ (if this is not unique, we choose any such $i$ and fix it) for which $\alpha_i \geq \alpha_\ell$ for all $\ell \in \{1, \ldots, k\} \setminus \{i\}$. Note that this guarantees, by the pigeon hole principle, that $\alpha_i \geq \frac{1}{k}$.

Under the assumption that $G|D$ is $a_{1,1}$-regular, we have $\sum_{\ell=1}^k \deg_{C_\ell}(v) = a_{1,1}$ for every $v \in C_i$. On the other hand, $\deg_{C_i}(v) + \deg_{C_i \setminus C_j}(v) = a_{i,i}$ for each $v \in C_i$. These together imply

$$\sum_{\ell \in \{1, \ldots, k\} \setminus \{i\}} \deg_{C_\ell}(v) = a_{1,1} - a_{i,i} + \deg_{C_i \setminus C_j}(v). \quad (3.3)$$

To prove Proposition 3.2 we first condition on the $\sigma$-field $\mathcal{F}$ comprising the following information:

(i) the vertex sets of $C_\ell$ and $C_\ell$ for all $\ell = 1, \ldots, k$,
(ii) the subgraph $\mathcal{G}|_{C_i}$ induced on $C_i$. Given $\mathcal{F}$, we enumerate the vertices of $C_i$ as $v_1, \ldots, v_n$ such that $C_i = \{v_1, \ldots, v_{\alpha_i n}\}$. From (3.3), we set

$$g_s = a_{1,1} - a_{i,i} + \deg_{\mathcal{G}|_{C_i}}(v_s)$$

for all $s = 1, \ldots, \alpha_i n$. The random variables $g_s$ are measurable with respect to $\mathcal{F}$. The conditional probability of the event that $\mathcal{G}|_D$ is $a_{1,1}$-regular is bounded above by the conditional probability of the event

$$A = \left\{ \sum_{\ell \in \{1, \ldots, k\} \setminus \{i\}} \deg_{\mathcal{G}_\ell}(v_s) = g_s \text{ for all } s = 1, \ldots, \alpha_i n \right\}.$$ 

We show that the probability of the event $A$ is $o(1)$ as $n \to \infty$.

First, we express $A$ as the union of pairwise disjoint events. For $g \in \mathbb{N}$, let us define the following subset of ordered $(k-1)$-tuples of non-negative integers:

$$S_g = \left\{ (m_1, \ldots, m_{i-1}, m_{i+1}, \ldots, m_k) : 0 \leq m_\ell \leq a_{i,\ell} \text{ for all } \ell \in \{1, \ldots, k\} \setminus \{i\}, \sum_{\ell \in \{1, \ldots, k\} \setminus \{i\}} m_\ell = g \right\}.$$ 

Then $A$ can be written as the union of the events

$$A\left(m^{(s)} : s = 1, \ldots, \alpha_i n\right) = \left\{ \deg_{\mathcal{G}_\ell}(v_s) = m^{(s)}_\ell \text{ for all } \ell \in \{1, \ldots, k\} \setminus \{i\} \text{ and } s = 1, \ldots, \alpha_i n \right\}$$

where $m^{(s)} = \left( m^{(s)}_\ell : \ell \in \{1, \ldots, k\} \setminus \{i\} \right)$ belongs to $S_{g_s}$ for all $s = 1, \ldots, \alpha_i n$. Note, from the mutual independence of the subgraphs $\mathcal{G}_{C_i, \mathcal{C}_\ell}$ for $\ell \in \{1, \ldots, k\} \setminus \{i\}$, that

$$P\left[A\left(m^{(s)} : s = 1, \ldots, \alpha_i n\right) \mid \mathcal{F}\right] = \prod_{\ell \in \{1, \ldots, k\} \setminus \{i\}} P\left[\deg_{\mathcal{G}_\ell}(v_s) = m^{(s)}_\ell \text{ for all } s = 1, \ldots, \alpha_i n \right] \mid \mathcal{F}\right] \leq P\left[\deg_{\mathcal{G}_\ell}(v_s) = m^{(s)}_\ell \text{ for all } s = 1, \ldots, \alpha_i n \right] \mid \mathcal{F}\right] \leq n^{-1}$$

for each $\ell \in \{1, \ldots, k\} \setminus \{i\}$. The goal now is to fix any $m^{(s)} \in S_{g_s}$ for each $s$ and establish that the probability of the event $\left\{ \deg_{\mathcal{G}_\ell}(v_s) = m^{(s)}_\ell \text{ for all } s = 1, \ldots, \alpha_i n \right\}$ for at least one $\ell \in \{1, \ldots, k\} \setminus \{i\}$ is $o(n^{-1})$ as $n \to \infty$.

Since $m^{(s)} \in S_{g_s}$ for each $s$, from (3.4), we have

$$\sum_{\ell \in \{1, \ldots, k\} \setminus \{i\}} \sum_{s=1}^{\alpha_i n} m^{(s)}_\ell = \sum_{s=1}^{\alpha_i n} \sum_{\ell \in \{1, \ldots, k\} \setminus \{i\}} m^{(s)}_\ell = \sum_{s=1}^{\alpha_i n} g_s = \deg(C_i, \mathcal{C}_i \setminus C_i) + (a_{1,1} - a_{i,i})\alpha_i n.$$ (3.6)

For $G$ uniformly randomly chosen from $\mathcal{R}_d^n$, [17], Theorem 1.1 showed that $\gamma \geq 1 - \frac{\sqrt{d}}{2}$ a.a.s. as $n \to \infty$, where $\gamma$ is the spectral gap for the adjacency matrix of $G$. Given a $d$-regular graph $G$ on $n$ vertices and a subset $S$ of $V(G)$ with $|S| \leq \frac{n}{2}$, [31], Theorem 13.14] (see also [25], [28], and [9], Theorem 6) established that

$$\frac{\gamma}{2} \leq \frac{\deg(S, V(G) \setminus S)}{d|S|}.$$ 

Combining these, we get

$$\deg(C_i \setminus C_i, C_i) \geq \min\{\alpha_i, 1 - \alpha_i\} \left(\frac{1}{2} - \frac{1}{\sqrt{\alpha_i}}\right) a_{i,i} n.$$ (3.7)
From (3.6) and (3.7), we get
\[ \sum_{t \in \{1, \ldots, k\} \setminus \{i\}} a_{i,n} m^{(s)}_t \geq \min\{\alpha_i, 1 - \alpha_i\} \left( 1 - \frac{1}{\sqrt{a_{i,i}}} \right) a_{i,i} n + (a_{1,1} - a_{i,i}) \alpha_i n. \]

By the pigeon-hole principle, there exists at least one \( j \in \{1, \ldots, k\} \setminus \{i\} \) such that
\[ \sum_{s=1}^{\alpha_i n} m^{(s)}_j \geq \frac{1}{k-1} \left\{ \min\{\alpha_i, 1 - \alpha_i\} \left( 1 - \frac{1}{\sqrt{a_{i,i}}} \right) a_{i,i} n + (a_{1,1} - a_{i,i}) \alpha_i n \right\}. \tag{3.8} \]

For the rest of the proof, we fix such a \( j \), and establish the following lemma:

**Lemma 3.4.** Let \( A_j \) denote the event that there exist \( C_i = \{v_1, \ldots, v_{\alpha_i n}\} \subset C_i \) and \( C_j \subset C_j \) such that \( \deg_{C_j}(v_s) = m^{(s)}_j \) for all \( s = 1, \ldots, \alpha_i n \). Then \( P[A_j] = o(n^{-1}) \).

The proof of this lemma is accomplished through the consideration of three different cases, in 3.1, 3.2 and 3.3.

3.1. **When \( a_{1,1} \geq a_{i,i} \).** We note at the very outset that the analysis of 3.1 is enough for the special and commonly studied situation where all intra-cluster degrees are the same. We set \( G_j = \sum_{s=1}^{\alpha_i n} m^{(s)}_j \), so that from (3.8), we have, for all \( a_i n \) sufficiently large,
\[ \frac{G_j}{n} \geq \frac{\min\{\alpha_i, 1 - \alpha_i\} a_{i,i}}{4(k-1)}. \tag{3.9} \]

We refer the reader to [8, Lemma 2] for the following inequality:
\[ P\left[ \deg_{C_j}(v_s) = m^{(s)}_j \right] \leq P\left[ \deg_{C_j}(v_s) \in \{\eta_j, \eta_j + 1\} \right] \text{ for all } s = 1, \ldots, \alpha_i n, \sum_{s=1}^{\alpha_i n} \deg_{C_j}(v_s) = \sum_{s=1}^{\alpha_i n} m^{(s)}_j \right| F, \tag{3.10} \]

where \( \eta_j = \left\lfloor \frac{G_j}{a_{i,n}} \right\rfloor \). Notice that we have the trivial bound \( G_j = \deg(C_i, C_j) \leq \deg(C_j, C_i) = \alpha_j a_{i,j} n \), so that
\[ \eta_j \leq \frac{\alpha_j a_{i,j}}{\alpha_i}. \tag{3.11} \]

We invoke the configuration model discussed in 3.1.1 and outline, in the next paragraph, some foundational aspects of the argument that resemble [8, Lemma 5]. Let \( \xi_{i,j} \) denote the half-edges of type \( \{i, j\} \) emanating from vertex \( v_s \), for each \( s = 1, \ldots, \alpha_i n \). Let \( B_t \) denote the indicator random variable of the event that \( \xi_t \) is matched with a half-edge of type \( \{i, j\} \) emanating from \( C_j \). Conditioned on \( B_1, \ldots, B_t \), the random variable \( B_{t+1} \) is Bernoulli with probability
\[ \tilde{p}_t = \frac{\alpha_j a_{i,j} n - \sum_{t' \leq t} B_{t'}}{a_{i,j} n - t}. \]

For all \( 1 \leq t \leq \alpha_i n \), we see that \( |\tilde{p}_t - \tilde{p}_{t-1}| \leq O(n^{-1}) \), so that for all \( s = 1, \ldots, \alpha_i n \), there exists \( p_s \in (0, 1) \) such that
\[ \left| \left| \deg_{C_j}(v_s) \right| G_{s-1}, \text{Bin}(a_{i,j}, p_s) \right|_{TV} = O\left( \frac{1}{n} \right), \tag{3.12} \]

where \( G_s \) denotes the \( \sigma \)-field generated by \( \{\xi_t, 1 \leq t \leq a_{i,j}s\} \), for each \( s \). Given that each of \( \deg_{C_j}(v_1), \ldots, \deg_{C_j}(v_{s-1}) \) takes values in \( \{\eta_j, \eta_j + 1\} \), the number of half-edges emanating from \( C_j \)
that have not yet been matched is at least $\alpha_j a_{i,j} n - (\eta_j + 1) (s - 1)$ and at most $\alpha_j a_{i,j} n - \eta_j (s - 1)$. Thus

$$\frac{\alpha_j a_{i,j} n - (\eta_j + 1) (s - 1)}{a_{i,j} (n - s + 1)} \leq p_s \leq \frac{\alpha_j a_{i,j} n - \eta_j (s - 1)}{a_{i,j} (n - s + 1)}.$$  

(3.13)

We shall now consider three different ranges of values of $\alpha_j$, where $j$ is as chosen by \[3.8\]. First, consider

$$\frac{c}{a_{i,j}} < \alpha_j \leq \frac{1}{2},$$

(3.14)

where

$$\log_2 c > \max \left\{ 16 k + 1, \frac{2 (k - 1) C}{4^s - 1} \right\},$$

(3.15)

where $C$ is as in \[3.11\] of Theorem \[3.1\]. We first consider the case of $\eta_j \neq 0$. For each $s = 1, \ldots, \lfloor \frac{\alpha_j n}{4} \rfloor$, from \[3.11\] and \[3.13\], we have

$$p_s \geq \frac{\alpha_j a_{i,j} n - 2 \eta_j \cdot \frac{\alpha_j n}{4}}{a_{i,j} n} \geq \frac{\alpha_j a_{i,j} n - 2 \cdot \frac{\alpha_j a_{i,j} n}{a_{i,j} n} \cdot \frac{\alpha_j n}{4}}{\alpha_j a_{i,j} n} = \frac{\alpha_j}{2} > \frac{c}{2 a_{i,j}}.$$  

(3.16)

Using \[3.12\], \[3.16\], the fact that the mode of the $\text{Bin}(n, p)$ distribution is $\lfloor (n + 1) p \rfloor$, and the same argument as in \[3.8\], Lemma 4, we conclude that, for all $s = 1, \ldots, \lfloor \frac{\alpha_j n}{4} \rfloor$,

$$\mathbb{P} \left[ \deg_{C_j}(v_s) \in \{\eta_j, \eta_j + 1\} \mid G_{s-1} \right] \leq \mathbb{P} \left[ \text{Bin} \left( a_{i,j}, p_s \right) = \lfloor (a_{i,j} + 1) p_s \rfloor \right] + O \left( \frac{1}{n} \right)$$

$$\leq O \left( \sqrt{\frac{1}{a_{i,j} p_s}} \right) \leq O \left( \sqrt{\frac{2}{c}} \right).$$  

(3.17)

We now use union bounds and Stirling’s approximation to bound above $\mathbb{P}[A_j]$ by

$$\left( \frac{n}{\alpha_j n} \right) \left( \frac{n}{\alpha_j n} \right) \left( O \left( \sqrt{\frac{2}{c}} \right) \right)^{\frac{\alpha_j n}{4}} \sim O \left( \frac{2^{H(\alpha_j) n + H(\alpha_j) n}}{\sqrt{\alpha_j \alpha_j (1 - \alpha_j) (1 - \alpha_j) n}} \left( \frac{2}{c} \right)^{\frac{\alpha_j n}{8}} \right)$$

$$\leq O \left( \frac{2^{2n}}{\sqrt{\alpha_j \alpha_j (1 - \alpha_j) (1 - \alpha_j) n}} \left( \frac{2}{c} \right)^{\frac{\alpha_j n}{8}} \right),$$

which is $o(n^{-1})$ by \[3.15\].

Now we consider the case of $\eta_j = 0$. We enumerate the vertices of $C_j$ so that $\deg_{C_j}(v_s) = 1$ for $s = 1, \ldots, G_j$ and $\deg_{C_j}(v_s) = 0$ for all $G_j + 1 \leq s \leq \alpha_j n$. Inspired by the lower bound in \[3.13\], set

$$f(s - 1) = \frac{\alpha_j a_{i,j} n - s + 1}{a_{i,j} (n - s + 1)}, \text{ for all } s \in [G_j].$$

The function $f$ being strictly increasing for $\alpha_j$ as in \[3.14\], a uniform lower bound on $p_s$ for all $s = 1, \ldots, G_j$ is $f(0) = \alpha_j > \frac{1}{a_{i,j}}$. By similar computations as used in deriving in \[3.17\], we get

$$\mathbb{P} \left[ \deg_{C_j}(v_s) = 1 \mid G_{s-1} \right] \leq O \left( \sqrt{\frac{1}{c}} \right)$$

(3.18)
for all $s = 1, \ldots, G_j$. Note that $\alpha_j \leq \alpha_i$ by our choice of $i$, and $\alpha_i + \alpha_j \leq 1$ implies that $\alpha_j \leq 1 - \alpha_i$. Hence, from (3.9), (3.14) and (3.15) we conclude that
\[
\frac{G_j}{n} \geq \frac{\alpha_j a_{i,i}}{4(k-1)} > \frac{ca_{i,i}}{4(k-1)a_{i,j}} > \frac{2 \cdot 4^{8k-1}}{(k-1)c}.
\] (3.19)

In this case, the upper bound on $P[A_j]$ is given by
\[
\left( \frac{n}{\alpha_i n} \right) \left( \frac{n}{\alpha_j n} \right) \left\{ O \left( \frac{1}{\sqrt{c}} \right) \right\}^{G_j} \leq O \left( \frac{2^n}{\sqrt{\alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j)n}} \left( \frac{1}{c} \right)^{4(k-1)c} \right)
\]
which is $o(n^{-1})$ due to (3.15).

Next, we consider the following range of values of $\alpha_j$:
\[
\frac{1}{a_{i,i}^2} < \alpha_j < \frac{c}{a_{i,j}},
\] (3.20)
where $c$ is as in (3.15). At the very outset of this case, we note that, if $\alpha_i \leq \frac{1}{2}$, we must have
\[
c \geq \alpha_j a_{i,j} \geq \frac{G_j}{n} > \frac{\alpha_i a_{i,i}}{4(k-1)},
\]
implying that $\alpha_i \leq \frac{4(k-1)c}{a_{i,i}}$. For all $a_{i,i}$ sufficiently large, this upper bound is smaller than $\frac{1}{k}$, giving us a contradiction.

Remark 3.5. The above reasoning shows that for the range $\alpha_j \leq \frac{c}{a_{i,j}}$, we need not consider $\alpha_i \leq \frac{1}{2}$.

When $\alpha_i > \frac{1}{2}$, we have, by similar reasoning as above,
\[
c \geq \frac{1 - \alpha_i a_{i,i}}{4(k-1)} \implies \alpha_i \geq 1 - \frac{4c(k-1)}{a_{i,i}}.
\] (3.21)

For all $a_{i,i}$ sufficiently large, this yields:
\[
\alpha_j < 1 - \alpha_i \leq \frac{4c(k-1)}{a_{i,i}} < \frac{1}{2} < 1 - \frac{4c(k-1)}{a_{i,i}} \leq \alpha_i < 1 - \alpha_j,
\] (3.22)
and by the concave nature of the entropy function and its symmetry around $\frac{1}{2}$, we conclude that
\[
H(\alpha_i) + H(\alpha_j) \leq 2H \left( \frac{4c(k-1)}{a_{i,i}} \right).
\] (3.23)

We first address the case of $\eta_j \neq 0$. From (3.13) and (3.20), for all $s = 1, \ldots, \left\lfloor \frac{\alpha_i n}{2} \right\rfloor$, we get
\[
p_s \leq \frac{\alpha_j a_{i,j} n}{a_{i,j} \left( n - \frac{\alpha_i n}{2} \right)} < \frac{2c}{a_{i,j}},
\] (3.24)
so that $\deg_{G_j}(v_s)$, conditioned on $G_{s-1}$, is stochastically dominated by $\Bin \left( a_{i,j}, \frac{2c}{a_{i,j}} \right)$, which in turn can be approximated by the Poisson($2c$) distribution. Thus
\[
P \left[ \deg_{G_j}(v_s) \in \{ \eta_j, \eta_{j+1} \} \middle| G_{s-1} \right] \leq P \left[ \text{Poisson($2c$)} \geq 1 \right] = \gamma
\] (3.25)
where $\gamma$ is a constant that depends only on $c$. Using (3.23) and (3.25), and $\frac{\alpha_i n}{2} > \frac{n}{4}$, we get the following upper bound on $\frac{1}{n} \log_2 P[A_j]$:
\[
- \frac{\log_2 \left( \alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) \right)}{2n} - \frac{\log_2 n}{n} + 2H \left( \frac{4c(k-1)}{a_{i,i}} \right) + \frac{\log_2 \gamma}{4}.
\] (3.26)
The first two terms approach 0 as \( n \to \infty \). The last term is a strictly negative constant, and as \( a_{i,j} \) grows, the third term goes to 0. Hence (3.20) is strictly negative for all sufficiently large \( a_{i,i} \), as \( n \to \infty \).

Remark 3.6. Observe that, in the above argument, nowhere has the lower bound on \( \alpha_j \) from (3.20) been used. This shows that as far as the case of \( \eta_j \neq 0 \) is concerned, our proof of Lemma 3.4 for the regime of (3.1) ends here. In the rest of (3.1) we only consider \( \eta_j = 0 \).

Now, we consider \( \eta_j = 0 \) and \( \alpha_j \) in the range given by (3.20). If \( G_j \geq \frac{a_{j,n}}{2} \), then the same argument as above will be enough.

Remark 3.7. This shows that for all \( \alpha_j \leq \frac{1}{a_{i,j}} \) and \( \eta_j = 0 \), as long as \( G_j \geq \frac{a_{j,n}}{2} \), our proof of Lemma 3.4 is already complete. Henceforth, we only consider \( \eta_j = 0 \) and \( G_j < \frac{a_{j,n}}{2} \).

If \( G_j < \frac{a_{j,n}}{2} \), then for each \( s = 1, \ldots, G_j \), the bound in (3.21) holds, and hence so does (3.25). Together with (3.9) and (3.22), this yields the following upper bound on \( \frac{1}{n} \log_2 P[A_j] \):

\[
- \frac{\log_2(a_{i,j})(1 - \alpha_i)2n}{2} + 2H(\alpha_i) + \frac{(1 - \alpha_i)a_{i,i} \log_2 \gamma}{4(k-1)}.
\]

Again, the first two terms approach 0 as \( n \to \infty \). We focus on the last two terms. Using the lower bound on \( \alpha_j \) from (3.20) and the fact that \( x \log x > (1-x)\log(1-x) \) for all \( x \in (0,1) \), we get:

\[
2H(\alpha_i) + \frac{(1 - \alpha_i)}{4(k-1)}a_{i,i} \log_2 \gamma \leq -4(1 - \alpha_i) \log_2(1 - \alpha_i) + \frac{(1 - \alpha_i)a_{i,i} \log_2 \gamma}{4(k-1)}
\]

\[
\leq 8(1 - \alpha_i) \log_2 a_{i,j} + \frac{(1 - \alpha_i)a_{i,i} \log_2 \gamma}{4(k-1)}
\]

\[
\leq 8(1 - \alpha_i) \{ \log_2 C + \log_2 a_{i,i} \} + \frac{(1 - \alpha_i)a_{i,i} \log_2 \gamma}{4(k-1)}
\]

\[
= (1 - \alpha_i) \left\{ 8 \log_2 C + 8 \log_2 a_{i,i} + \frac{a_{i,i} \log_2 \gamma}{4(k-1)} \right\}.
\]

As \( a_{i,i} \) grows to \( \infty \) much faster than \( \log_2 a_{i,i} \), and the coefficient of \( a_{i,i} \) is a strictly negative constant whereas that of \( \log_2 a_{i,i} \) is a positive one, hence this is strictly negative for all \( a_{i,i} \) sufficiently large.

Finally, we consider

\[
\alpha_j \leq \frac{1}{a_{i,j}}, \tag{3.27}
\]

and by Remarks 3.5 and 3.6, we need only consider \( \alpha_i \geq \frac{1}{2} \) and \( \eta_j = 0 \). From (3.13), for all \( s = 1, \ldots, \left\lfloor \frac{a_{j,n}}{2} \right\rfloor \), we have

\[
p_s \leq \frac{\alpha_j a_{i,j} n}{a_{i,j} (n - \frac{a_{j,n}}{2})} < 2\alpha_j \leq 2(1 - \alpha_i),
\]

so that

\[
P \left[ \deg_{C_j}(v_s) \in \{\eta_j, \eta_j + 1\} \mid F_{s-1} \right] \leq P \left[ \text{Bin}(a_{i,j}, p_s) \geq 1 \right] + O(n^{-1}) \leq a_{i,j}p_s < 2(1 - \alpha_i)a_{i,j}. \tag{3.28}
\]

By Remark 3.7, we need only consider \( G_j < \frac{a_{j,n}}{2} \), so that (3.28) holds for all \( s = 1, \ldots, G_j \). By (3.9) and (3.27), we get:

\[
\frac{1}{a_{i,j}} \geq \alpha_j a_{i,j} \geq \frac{G_j}{n} \geq \frac{(1 - \alpha_i)a_{i,i}}{4(k-1)} \implies 1 - \alpha_i \leq \frac{4(k-1)}{a_{i,j}a_{i,i}}. \tag{3.29}
\]
For any fixed positive integer \( r > 2 \), for all \( a_{i,i} \) sufficiently large, by (3.29) of Theorem 3.1, we have
\[
a_{i,i} > C \frac{1}{2^{r-2}} (4(k-1))^r \quad \Rightarrow \quad a_{i,i}^{r-1} \geq \left\{ 4(k-1) \right\}^{r-1} 2^r C a_{i,i} \geq \left\{ 4(k-1) \right\}^{r-1} 2^r a_{i,j},
\]
so that by (3.29) we get
\[
\log_2 \left\{ 2(1 - \alpha_i) a_{i,j} \right\} \leq \frac{r - \log_2 (1 - \alpha_i)}{n}.
\] (3.30)

By (3.9) and (3.28), we get the following upper bound on \( \frac{1}{n} \log_2 P[A_j] \):
\[
- \frac{\log_2 (\alpha_i a_{j} (1 - \alpha_i)(1 - \alpha_j))}{2n} - \frac{\log_2 n}{n} + 2H(\alpha_i) + \frac{(1 - \alpha_i) a_{i,j}}{4(k-1)} \log_2 \left\{ 2(1 - \alpha_i) a_{i,j} \right\}.
\]

Again, it suffices to focus on the last two terms, and by (3.30), we get the following bound:
\[
2H(\alpha_i) + \frac{(1 - \alpha_i) a_{i,j}}{4(k-1)} \log_2 \left\{ 2(1 - \alpha_i) a_{i,j} \right\} \leq 4(1 - \alpha_i) \log_2 (1 - \alpha_i) + \frac{a_{i,j}}{4(k-1)} \frac{\log_2 (1 - \alpha_i)}{r}
\]
\[
= (1 - \alpha_i) \log_2 (1 - \alpha_i) \left\{ -4 + \frac{a_{i,j}}{4(k-1)r} \right\},
\]
which is strictly negative for all \( a_{i,j} \) sufficiently large. This brings us to the end of §3.1

3.2. When \( a_{1,1} < a_{i,i} \) and \( \alpha_i > \frac{1}{2} \). Note that, by Lemma 3.3, this situation arises only when \( B_1 > a_{i,i} - a_{1,1} \), and we need only consider \( \alpha_i \leq \frac{B_1}{a_{i,i} - a_{1,1} + B_1} \). From (3.8) and the hypothesis of Theorem 3.1 for all sufficiently large \( a_{i,i} \), we get:
\[
\frac{G_j}{n} \geq \frac{1}{k-1} \left( \frac{1}{2} - \frac{1}{\sqrt{a_{i,i}}} \right) a_{i,i} - \frac{\alpha_i (a_{i,i} - a_{1,1})}{k-1}
\]
\[
\geq \frac{1}{k-1} \left( \frac{1}{2} - \frac{1}{\sqrt{a_{i,i}}} \right) a_{i,i} - \frac{(1 - \alpha_i)(a_{i,i} - a_{1,1} + B_1)\alpha_i}{k-1}
\]
\[
= \frac{1 - \alpha_i}{k-1} \left\{ \frac{a_{i,i}}{2} - \sqrt{a_{i,i}} - B_1 \right\}
\]
\[
\geq \frac{1 - \alpha_i}{k-1} \left\{ \left( \frac{1}{2} - \epsilon_{i,1} \right) a_{i,i} - B_1 \right\} > \frac{(1 - \alpha_i)\epsilon_{i,1} a_{i,i}}{k-1},
\] (3.31)

where the last inequality follows from (3.2).

We again split the analysis into three parts depending on the ranges of values of \( \alpha_j \) as given in (3.14), (3.20) and (3.27), with \( c \) satisfying the following condition:
\[
\log_2 c > \max \left\{ 16k + 1, \frac{2C(k-1)}{\epsilon_{i,1}} \right\}.
\] (3.32)

We first consider the case of \( \eta_j \neq 0 \) and then the case of \( \eta_j = 0 \) in each of these ranges.

When we are in the regime of (3.14) and \( \eta_j \neq 0 \), we note that the bounds in (3.15) and (3.16) hold, and therefore the same analysis as before goes through. When \( \alpha_j \leq \frac{c}{\alpha_j} \) and \( \eta_j \neq 0 \), the bounds in (3.21) and (3.22) hold for all \( s = 1, \ldots, \left\lfloor \frac{\alpha_i}{2} \right\rfloor \). Now, from (3.31), we have:
\[
c \geq \alpha_j a_{i,j} \geq \frac{G_j}{n} \geq \frac{(1 - \alpha_i)\epsilon_{i,1} a_{i,i}}{k-1} \quad \Rightarrow \quad \alpha_i \geq 1 - \frac{c(k-1)}{\epsilon_{i,1} a_{i,i}}.
\] (3.33)

Then \( \frac{1}{n} \log_2 P[A_j] \) can be bounded above by
\[
- \frac{\log_2 \left\{ \alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) \right\}}{2n} - \frac{\log_2 n}{n} + H(\alpha_i) + H(\alpha_j) + \frac{\alpha_i}{2} \log_2 \gamma
\]
of which the first two terms approach 0 as \( n \to \infty \). The remaining terms can be bounded above by

\[
2H \left( \frac{c(k-1)}{\epsilon_{i,1}a_{i,i}} \right) + \frac{\log_2 \gamma}{4},
\]

of which the first term can be made arbitrarily small by choosing \( a_{i,i} \) sufficiently large, and the last term is a strictly negative constant. Hence the above is strictly negative for all \( r > a \).

For any fixed positive integer \( s \), the bound of (3.18) holds for all \( a \). By (3.31) and since \( \alpha_j \leq 1 - a_i \), we get

\[
\frac{G_j}{n} \geq \frac{\alpha_j \epsilon_{i,1}a_{i,i}}{k-1} > \frac{c\epsilon_{i,1}a_{i,i}}{(k-1)a_{i,j}} > \frac{2 \cdot 4^k \epsilon_{i,1}^n}{C(k-1)}
\]

by (3.31), (3.32) and (i). Thus an upper bound on \( P[A_j] \) is given by

\[
O \left\{ \frac{2H(\alpha_i)n + H(\alpha_i)n}{\sqrt{\alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) n}} \left( \frac{1}{c} \right) \frac{G_j/2}{n} \right\} \leq O \left\{ \frac{22n}{\sqrt{\alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) n}} \left( \frac{1}{c} \right) \frac{4^k \epsilon_{i,1}^n}{C(k-1)} \right\}
\]

which is \( o(n^{-1}) \) for \( c \) as in (3.32).

Next, we consider \( \alpha_j \) as in (3.20). Again, the bounds in (3.24) and (3.25) hold for all \( a \). From (3.31), we get the following upper bound on \( n \log_2 P[A_j] \):

\[
- \frac{\log_2 \{ \alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) \}}{2n} - \frac{\log_2 n}{\sqrt{2} n} + 2H(\alpha_i) + \frac{(1 - \alpha_i) \epsilon_{i,1}a_{i,i}}{k-1} \log_2 \gamma.
\]

The sum of the last two terms can be bounded above by

\[
-4(1 - \alpha_i) \log_2(1 - \alpha_i) + \frac{(1 - \alpha_i) \epsilon_{i,1}a_{i,i}}{k-1} \log_2 \gamma < 8(1 - \alpha_i) \log_2 a_{i,j} + \frac{(1 - \alpha_i) \epsilon_{i,1}a_{i,i}}{k-1} \log_2 \gamma
\]

\[
\leq (1 - \alpha_i) \left\{ 8 \log_2 C + 8 \log_2 a_{i,i} + \frac{\epsilon_{i,1}a_{i,i}}{k-1} \log_2 \gamma \right\},
\]

which is strictly negative for all \( a_{i,i} \) sufficiently large.

Finally, we consider \( \alpha_j \) in the range given in (3.27). The bound in (3.28) holds for all \( s = 1, \ldots, G_j \). From (3.31), we have

\[
\frac{1}{a_{i,j}} \geq \alpha_j a_{i,j} \geq \frac{G_j}{n} \geq \frac{(1 - \alpha_i) \epsilon_{i,1}a_{i,i}}{k-1} \implies \alpha_i \geq 1 - \frac{k-1}{\epsilon_{i,1}a_{i,i}a_{i,j}}.
\]

For any fixed positive integer \( r > 2 \) and all \( a_{i,i} \) sufficiently large,

\[
a_{i,i} > C^{\frac{1}{r-1}} \left( \frac{k-1}{\epsilon_{i,1}} \right)^{1\frac{r-1}{2}} \frac{2^r}{2},
\]

and by the same reasoning as in (3.30), using (3.35) we conclude that

\[
\log_2 \{ 2(1 - \alpha_i) a_{i,j} \} \leq \frac{\log_2(1 - \alpha_i)}{r}.
\]

An upper bound on \( \frac{1}{n} \log_2 P[A_j] \) is given by

\[
- \frac{\log_2 \{ \alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) \}}{2n} - \frac{\log_2 n}{\sqrt{2} n} + 2H(\alpha_i) + \frac{(1 - \alpha_i) \epsilon_{i,1}a_{i,i}}{k-1} \log_2 \{ 2(1 - \alpha_i) a_{i,j} \}.
\]
of which the first two terms approach 0 as \( n \to \infty \), and the sum of the last two terms can be bounded by

\[
-4(1 - \alpha_i) \log_2(1 - \alpha_i) + \frac{(1 - \alpha_i) \epsilon_{i,1} a_{i,i} \log_2(1 - \alpha_i)}{k - 1} = (1 - \alpha_i) \log_2(1 - \alpha_i) \left\{ -4 + \frac{\epsilon_{i,1} a_{i,i}}{r(k - 1)} \right\}
\]

which is strictly negative for all \( a_{i,i} \) sufficiently large. This brings us to the end of 3.2.

### 3.3. When \( a_{i,i} > a_{1,1} \) and \( \alpha_i \leq \frac{1}{2} \)

By (3.8), we have

\[
\frac{G_j}{n} \geq \frac{\alpha_i}{k - 1} \left( \frac{1}{2} - \frac{1}{\sqrt{a_{i,i}}} \right) a_{i,i} - \frac{\alpha_i(a_{i,i} - a_{1,1})}{k - 1} \\
\geq \frac{\alpha_i}{k - 1} \left( a_{1,1} - \frac{a_{i,i}}{2} \right) - \frac{\alpha_i a_{i,i}}{k - 1},
\]

by (3.11). Note that, if \( \alpha_j \leq \frac{c}{a_{i,j}} \) for any constant \( c > 1 \), then

\[
c \geq \alpha_j a_{i,j} \geq \frac{G_j}{n} \geq \frac{\alpha_i \delta_{1,i} a_{i,i}}{k - 1} \implies \alpha_i \leq \frac{c(k - 1)}{\delta_{1,i} a_{i,i}},
\]

which is strictly less than \( \frac{1}{k} \) for all \( a_{i,i} \) sufficiently large, contradicting our choice of \( i \). Hence we need only consider the range of (3.14) for values of \( \alpha_j \).

When \( \eta_j \neq 0 \), the argument is the same as the corresponding case in (3.11). When \( \eta_j = 0 \), the bound in (3.18) holds, and from (3.36) and \( \alpha_i \geq \frac{1}{k} \), we get the following upper bound on \( \frac{1}{n} \log P[A_j] \):

\[
- \frac{\log_2 \left\{ \alpha_i \alpha_j (1 - \alpha_i)(1 - \alpha_j) \right\}}{2n} - \frac{\log_2 n}{n} + H(\alpha_i) + H(\alpha_j) - \frac{\delta_{1,i} a_{i,i}}{2k(k - 1)} \log_2 c,
\]

and as \( H(\alpha_i) + H(\alpha_j) \leq 2 \), the above expression is strictly negative for all \( a_{i,i} \) sufficiently large.

### 4. Recovery of clusters

This section is dedicated to the identification and recovery of the underlying clusters \( C_i \) of \( G \). Let \( \mathcal{M} \) be the set of all algorithms that take as input a \( d \)-regular graph on \( kn \) vertices, where \( d \) is as in (1.1), and output a partition of the vertex set \( V(G) \) into \( k \) clusters of \( n \) vertices each. An algorithm in \( \mathcal{M} \) is said to allow weak recovery if, with probability approaching 1 as \( n \to \infty \), it outputs a partition \( (C_1', \ldots, C_k') \) such that \( \sum_{i=1}^k |C_i' \Delta C_i| = o(n) \), where \( \Delta \) indicates the symmetric difference between two sets. An algorithm in \( \mathcal{M} \) is said to allow strong recovery if, with probability going to 1 as \( n \to \infty \), it outputs the partition \( (C_1, \ldots, C_k) \). An algorithm in \( \mathcal{M} \) is called efficient if its run time is polynomial in \( n \).

We mention at the very outset that in \( \log \) refers to the natural logarithm. For any \( m \in \mathbb{N} \), we denote by \( e^{(m)} \) the vector in \( \mathbb{R}^m \) in which each coordinate equals 1. Recall the matrix \( A \) from (1.1) and \( d \) from (1.1). By the well-known Perron-Frobenius theorem (see [49, 19]), the largest eigenvalue of \( A \) is \( \lambda_1 = d \), with algebraic multiplicity 1 and \( e^{(k)} \) an eigenvector. Let \( d > \lambda_2 > \cdots > \lambda_k \), denote the distinct eigenvalues of \( A \), with algebraic multiplicities \( r_2, \ldots, r_p \) respectively, where \( r_2, \ldots, r_p \) are positive integers with \( \sum_{j=2}^p r_j = k - 1 \). We set \( r_1 = 1 \).

We define \( R_i = \sum_{j=1}^i r_j \), for each \( i = 1, \ldots, p \). Let \( \{x^{(R_{i-1}+1)}, \ldots, x^{(R_i)}\} \) form an orthonormal basis for the eigenspace of \( A \) corresponding to the eigenvalue \( \lambda_i \), for all \( i = 2, \ldots, p \). Let \( x^{(s)} = \)}
\((x_1^{(s)}, \ldots, x_k^{(s)})\) for each \(s = 2, \ldots, k\). Then, for each \(i = 2, \ldots, p\), we get:

\[
\sum_{t=1}^{k} a_{jt} x_t^{(s)} = \lambda_i x_j^{(s)}, \text{ for each } j \in \{1, \ldots, k\}, \text{ for each } s \in \{R_{i-1} + 1, \ldots, R_i\}. \tag{4.1}
\]

**Remark 4.1.** We note here that \(\{x^{(2)}, \ldots, x^{(k)}\}\) forms an orthonormal basis for the subspace of \(\mathbb{R}^k\) orthogonal to \(e^{(k)}\). This tells us that, given any distinct \(i, j \in \{1, \ldots, k\}\), there must exist some \(s \in \{2, \ldots, k\}\) such that \(x_i^{(s)} \neq x_j^{(s)}\).

We now state the last of our three main results.

**Theorem 4.2.** Assume that there exists some constant \(\eta\), independent of all entries of the matrix \(A\), such that \(\lambda_i^2 > 4(d - 1) + \eta^2\) for each \(i = 2, \ldots, p\). Then, for all \(d\) sufficiently large, there exists an efficient algorithm for weak recovery of the clusters in the model.

**Remark 4.3.** When we have a common value \(d_1\) for all intra-cluster degrees and a common value \(d_2\) for all inter-cluster degrees, we have \(p = 2\) and \(\lambda_2 = d_1 - d_2\). In this case, the hypothesis of Theorem 4.2 boils down to \((d_1 - d_2)^2 > 4(d_1 + (k - 1)d_2 - 1)\).

The rest of \(\square\) is dedicated to the proof of Theorem 4.2. For the rest of the paper, we let \(G \sim G_A\).

For each \(s = 2, \ldots, k\), we define a (random) vector \(\sigma^{(s)} = (\sigma_v^{(s)} : v \in V(G))\) with

\[
\sigma_v^{(s)} = \frac{x_j^{(s)}}{\sqrt{n}} \text{ if and only if } v \in C_j, \text{ for all } j \in \{1, \ldots, k\}, \tag{4.2}
\]

for each \(v \in V(G)\). Note that, as \(\frac{e^{(nk)}}{\sqrt{n}}, x^{(2)}, \ldots, x^{(k)}\) form an orthonormal set, we have

(i) \(\sigma^{(s)} \perp e^{(nk)}\) for each \(s = 2, \ldots, k\);  
(ii) \(\sigma^{(s)} \perp \sigma^{(t)}\) for distinct \(s, t \in \{2, \ldots, k\}\);  
(iii) and \(\|\sigma^{(s)}\|_{L_2} = 1\) for each \(s = 2, \ldots, k\).

The proof of Theorem 4.2 happens via Proposition 4.4. One may draw the parallel between Proposition 4.4 and the combination of Proposition 2, Lemma 7 and Lemma 8 of \([8]\).

Recall from \([11]\) that we denote by \(S^{(s)}\) the length-\(s\) self-avoiding matrix for the graph \(G\), for any \(s \in \mathbb{N}\), as well as the definition of asymptotically aligned sequences of vectors. Let \(c\) be a constant and \(\ell\) an even positive integer such that

\[
c \log d < \frac{1}{6} \text{ and } \ell = c \log n. \tag{4.3}
\]

**Proposition 4.4.** Assume the same hypothesis as in Theorem 4.2 and fix any \(\epsilon \in (0, 1)\). Then the following events happen with high probability, under the measure induced by \(G\), as \(n \to \infty\):

(i) We have

\[
S^{(\ell)} e^{(nk)} = d(d - 1)^{\ell-1} e^{(nk)} + \bar{c}, \tag{4.4}
\]

where \(\|\bar{c}\|_{L_2} = o(n)\).

(ii) For each \(i = 2, \ldots, p\), there exists a positive constant \(A_i\) such that

\[
S^{(\ell)} \sigma^{(s)} = A_i \beta_i^2 (1 + o(1)) \sigma^{(s)} + \tilde{\sigma}^{(s)} \text{ for each } s \in \{R_{i-1} + 1, \ldots, R_i\}, \tag{4.5}
\]

where

\[
\beta_i = \begin{cases} 
\frac{1}{2} \left( \lambda_i + \sqrt{\lambda_i^2 - 4(d-1)} \right) & \text{if } \lambda_i > 0, \\
\frac{1}{2} \left( \lambda_i - \sqrt{\lambda_i^2 - 4(d-1)} \right) & \text{if } \lambda_i < 0,
\end{cases} \tag{4.6}
\]
We state here [8], Lemma 9 which goes through verbatim, with a small correction in ii.

(iii) If $\sigma$ is a unit vector orthogonal to $e^{(nk)}$ and to $\sigma^{(s)}$ for each $s = 2, \ldots, k$, then for each $1 \leq m \leq \ell$, we have

$$
\left\lVert S^{(m)}\sigma \right\rVert_{L_2} \leq (\ell + 1)n^d m^{\ell/2}(1 + o(1)).
$$

(iv) The largest eigenvalue of $S^{(\ell)}$ is $\alpha_1 = d(d - 1)^{\ell-1} + o(1)$, and any unit eigenvector of $S^{(\ell)}$ corresponding to $\alpha_1$ is asymptotically aligned with $e^{(nk)}$. The matrix $S^{(\ell)}$ also has the eigenvalues $\alpha_i = A_i \beta_i^2 (1 + o(1))$, with algebraic multiplicity $r_i$, for each $i = 2, \ldots, p$. Any unit eigenvector of $S^{(\ell)}$ corresponding to $\alpha_i$ asymptotically belongs to the subspace spanned by $\{\sigma^{(s)} : s = R_{i-1} + 1, \ldots, R_i\}$, for $i = 2, \ldots, p$.

Finally, if we denote by $\alpha_{p+1}, \ldots, \alpha_q$ the remaining, distinct eigenvalues of $S^{(\ell)}$, for some $q \leq nk$, then $|\alpha_i| \leq n^d \ell^d/(1 + o(1))$ for all $i = p + 1, \ldots, q$.

Remark 4.5. Note that, since we choose $\ell$ to be even and $A_i$ is strictly positive from ii, $\alpha_i$ is strictly positive for each $i = 2, \ldots, p$.

Proof of Theorem 4.2 We first establish Theorem 4.2 using Proposition 4.4. For each $i = 2, \ldots, p$, from [4.6] and the hypothesis of Theorem 4.2, we have

$$
|\beta_i| > \frac{2\sqrt{d-1} + \eta}{2} > \sqrt{d} \text{ for all } d \text{ such that } \sqrt{d} + \sqrt{d-1} > \frac{2}{\eta}.
$$

Recall the eigenvalues $\alpha_i$, for all $i = 1, \ldots, q$, of the matrix $S^{(\ell)}$, as given in iv of Proposition 4.4.

Using [4.3], we have, for each $i \in \{2, \ldots, p\}$ and each $j \in \{p + 1, \ldots, q\}$,

$$
\log |\alpha_j| - \log |\alpha_i| < \epsilon \log n + \left(\frac{c\log n}{2}\right) \log d - \log A_i - \left(\frac{c\log n}{2}\right) \log |\beta_i| + \log(1 + o(1)),
$$

so that, by setting

$$
\epsilon < \frac{c}{2} (\log |\beta_i| - \log d),
$$

the ratio $|\alpha_j| / |\alpha_i|$ goes to 0 as $n \to \infty$. On the other hand, since $\lambda_i < d$ for each $i = 2, \ldots, p$, we have

$$
|\beta_i| < \frac{d + \sqrt{d^2 - 4(d - 1)}}{2} = \frac{d + d - 2}{2} = d - 1,
$$

which shows that the ratio $\alpha_j / \alpha_i$ goes to 0 as $n \to \infty$. These show that the eigenvalues $\alpha_2, \ldots, \alpha_p$ are well-separated, in absolute value, from both the largest eigenvalue $\alpha_1$ and the bulk $\{\alpha_i : i = \ell = p + 1, \ldots, q\}$. We note here that even if some, or all, of the eigenvalues $\alpha_2, \ldots, \alpha_p$ are equal to one another, it will only take polynomial time to obtain an orthonormal set of eigenvectors of $S^{(\ell)}$ corresponding to these eigenvalues, because of their separation from the remaining eigenvalues of $S^{(\ell)}$. Given any $\varepsilon > 0$, from iv of Proposition 4.4 and Remark 4.1 we can see how to construct, using these eigenvectors, a labeling that identifies accurately at least $(1 - \varepsilon)n$ many of the vertices that belong to $C_j$ for each $j \in \{1, \ldots, k\}$.

The rest of [3] is dedicated to establishing Proposition 4.4. However, a substantial portion of the proof follows mutatis mutandis from the corresponding parts of the argument in [86, 8], and these are clearly pointed out in the sequel. To this end, given any $s \in \mathbb{N}$ and any graph $G$, we call $G$ $s$-tangle-free if, for every $v \in V(G)$, the neighbourhood $B(v, s)$ contains at most one cycle ([3, 52]). We state here [8, Lemma 9] which goes through verbatim, with a small correction in [3].
Lemma 4.6. Let $c$ and $\ell$ be as in (4.3), and let $\delta = 4c\log d$. Let $0 < \varepsilon < 1 - 4\delta$ be a small constant. Then

(i) $G$ is $\ell$-tangle-free with probability $1 - O(n^{-\varepsilon})$;

(ii) letting $X^{(\ell)}$ denote the number of $v \in V(G)$ such that $B(v, \ell)$ contains a cycle,

$$
P \left[ X^{(\ell)} > n^\delta \mid G \text{ is } \ell \text{-tangle-free} \right] \leq (1 + o(1))dn^{-\delta/2}. \tag{4.8}$$

The proof of (i) follows exactly as in [Lemma 2.1, (4.2)]. Most of the proof of (ii) follows the same argument as in [Lemma 9, part (ii), (8)], and we only describe here the part which requires rectification. The probabilities mentioned in this paragraph are all conditioned on the event that $G$ is $\ell$-tangle-free. Defining, for each $v \in V(G)$, the event $T_\ell(v) = \{B(v, \ell) \text{ is a tree}\}$ and following the same argument as in (8), we arrive at

$$
\log \mathbb{P} [T_\ell(v)] \geq -(1 + o(1)) \frac{d^{2\ell+1}}{n} = -(1 + o(1))n^{2c\log d - 1},
$$

which yields

$$
\mathbb{P} [T_\ell(v)] \geq 1 - d(1 + o(1))n^{2c\log d - 1} + O \left( n^{2(2c\log d - 1)} \right) = 1 - d(1 + o(1))n^{2c\log d - 1} + o(n^{-1}).
$$

Hence

$$
\mathbb{P} [B(v, \ell) \text{ contains a cycle}] \leq d(1 + o(1))n^{2c\log d - 1} + o(n^{-1})
$$

for each $v \in V(G)$. Summing over all $v \in V(G)$ and using Markov’s inequality, we get (4.8).

The proof of (ii) of Proposition 4.4, using Lemma 4.6, is the same as [Lemma 7, (i), Page 23, (8)].

4.1. Proof of (ii) of Proposition 4.4 Fix an $i \in \{2, \ldots, p\}$ and any $s \in \{R_i-1 + 1, \ldots, R_i\}$. We shall now prove that (4.5) holds. For each $v \in V(G)$ and each $1 \leq t \leq \ell$, we define

$$
m_t^{(j)}(v) = \left\{ u \in \delta B(v, t) : \sigma_u^{(s)} = \frac{x_j}{\sqrt{n}} \right\}, \text{ for each } j' \in \{1, \ldots, k\}.
$$

Note that if $v$ and $v'$ are two distinct vertices with $\sigma_v^{(s)} = \sigma_{v'}^{(s)}$, we have $m_t^{(j)}(v) = m_t^{(j)}(v')$ for each $j' \in \{1, \ldots, k\}$ on the event $T_\ell(v) \cap T_\ell(v')$. Consequently, it makes sense to define, for each $j, j' \in \{1, \ldots, k\}$, the quantity

$$
m_t^{j,j'} = \left\{ u \in \delta B(v, t) : \sigma_u^{(s)} = \frac{x_j}{\sqrt{n}} \right\} \text{ for any } v \in \mathcal{T} \text{ with } \sigma_v^{(s)} = \frac{x_j}{\sqrt{n}}, \tag{4.9}
$$

where $\mathcal{T} = \{v \in V(G) : T_\ell(v) \text{ holds}\}$. For such a $v$, the $v$-th coordinate of $S^{(t)}\sigma^{(s)}$ is given by

$$
M_t^{(j)} := \left( S^{(t)}\sigma^{(s)} \right)_v = \sum_{w \in \delta B(v, t)} \sigma_w^{(s)} = \frac{1}{\sqrt{n}} \sum_{j' = 1}^k m_t^{j,j'} x_j^{(s)}. \tag{4.10}
$$

Fix any $j'' \in \{1, \ldots, k\}$. Each vertex in $C_{j''} \cap \delta B(v, t-1)$ has precisely $a_{j',j''}$ many neighbours in $C_{j''}$, so that the total number of edges between $\delta B(v, t-1)$ and $C_{j''}$ is $\sum_{j' = 1}^k a_{j',j''} m_t^{j,j'}$. Each vertex in $\delta B(v, t-2)$ has $d - 1$ neighbours in $\delta B(v, t-1)$, hence the number of edges between
\( \mathcal{C}_{j''} \cap \delta B(v, t - 2) \) and \( \delta B(v, t - 1) \) is \((d - 1)m_{t-2}^{j''j''}\). Consequently the number of edges between \( \delta B(v, t - 1) \) and \( \mathcal{C}_{j''} \cap \delta B(v, t) \) is

\[
\sum_{j' = 1}^{k} a_{j', j''}m_{t-1}^{j'j''} - (d - 1)m_{t-2}^{j''j''}.
\]

On the other hand, each vertex in \( \mathcal{C}_{j''} \cap \delta B(v, t) \) has precisely one neighbour in \( \delta B(v, t - 1) \), so that the number of edges between \( \delta B(v, t - 1) \) and \( \mathcal{C}_{j''} \cap \delta B(v, t) \) is \( m_{t}^{j''j''} \). Thus we get the recursion

\[
m_{t}^{j''j''} = \sum_{j' = 1}^{k} a_{j', j''}m_{t-1}^{j'j''} - (d - 1)m_{t-2}^{j''j''}. \tag{4.11}
\]

From (4.10), (4.11) and (4.11), we get:

\[
M_{t}^{(j)} = \frac{1}{\sqrt{n}} \sum_{j'' = 1}^{d} \left\{ \sum_{j' = 1}^{k} a_{j', j''}m_{t-1}^{j'j''} \right\} x_{j''}^{(s)} - \frac{(d - 1)}{\sqrt{n}} \sum_{j'' = 1}^{k} m_{t-2}^{j''j''}x_{j''}^{(s)} = \lambda_{i}M_{t-1}^{(j)} - (d - 1)M_{t-2}^{(j)}. \tag{4.12}
\]

It is immediate that \( M_{0}^{(j)} = \frac{x_{j}^{(s)}}{\sqrt{n}} \) as \( n \to \infty \). Note that \( v \) has precisely \( a_{j', j''} \) many neighbours in \( \mathcal{C}_{j} \), and consequently \( m_{1}^{j''j''} = a_{j', j''} \) for each \( j'' \in \{1, \ldots, k\} \). From (4.11), we have \( M_{1}^{(j)} = \lambda_{i}x_{j}^{(s)} \sqrt{n} \).

Defining the generating function \( G(\zeta) = \sum_{t = 0}^{\infty} M_{t}^{(j)} \zeta^{t} \) for \( \zeta \) within the radius of convergence of the series, we use (4.12) to get

\[
G(\zeta) = \frac{x_{j}^{(s)}}{\sqrt{n}(1 - \lambda_{i}z + (d - 1)\zeta z^{2})}.
\]

Let \( \beta_{i} \) and \( \beta_{i}' \) denote the two roots of the quadratic polynomial \( \zeta^{2} - \lambda_{i}z + (d - 1) \) such that \( |\beta_{i}| > |\beta_{i}'| \).

This shows that \( \beta_{i} \) satisfies (4.6). We then have

\[
G(\zeta) = \frac{x_{j}^{(s)}}{\sqrt{n}} \beta_{i}' \left\{ \sum_{t = 0}^{\infty} \beta_{i}'^{t} \zeta^{t} \right\} \left\{ \sum_{t = 0}^{\infty} \beta_{i}'^{t} \zeta^{t} \right\},
\]

so that the coefficient of \( \zeta^{t} \) is given by

\[
M_{t}^{(j)} = \left( \frac{x_{j}^{(s)}}{\sqrt{n}} \right) (d - 1) \beta_{i}'^{t} \sum_{t' = 0}^{t} \left( \frac{\beta_{i}'}{\beta_{i}} \right)^{t'} \beta_{i}'^{t} \left( \frac{x_{j}^{(s)}}{\sqrt{n}} \right) \beta_{i}^{-t}. \tag{4.13}
\]

Set \( A_{i} = (d - 1) \sum_{t' = 0}^{\infty} \left( \frac{\beta_{i}'}{\beta_{i}} \right)^{t'} \), where the sum converges since \( |\beta_{i}| > |\beta_{i}'| \). Therefore

\[
M_{t}^{(j)} = \left\{ A_{i} - (d - 1) \sum_{t' = t + 1}^{\infty} \left( \frac{\beta_{i}'}{\beta_{i}} \right)^{t'} \right\} \beta_{i}'^{t} \left( \frac{x_{j}^{(s)}}{\sqrt{n}} \right), \tag{4.14}
\]

with the error \( (d - 1) \sum_{t' = t + 1}^{\infty} \left( \frac{\beta_{i}'}{\beta_{i}} \right)^{t'} \zeta^{t} = O(1) \) as \( t \to \infty \). In particular, \( (d - 1) \sum_{t' = t + 1}^{\infty} \left( \frac{\beta_{i}'}{\beta_{i}} \right)^{t'} \zeta^{t} = O(1) \) as \( n \to \infty \) for \( \ell \) as in (4.13). From (4.10) and (4.14), we have

\[
\left( S(\ell) \sigma^{(s)} \right)_{v} = A_{i} \beta_{i}'^{t}(1 + O(1))\sigma_{v}^{(s)} \text{ for each } v \in T. \tag{4.15}
\]

We set

\[
\bar{\sigma}^{(s)} = S(\ell) \sigma^{(s)} - A_{i} \beta_{i}'^{t}(1 + O(1))\sigma^{(s)}. \tag{4.16}
\]
By part ii of Lemma 4.6, for any \( v \notin T \) and \( w \in \delta B(v, \ell) \), we have \( S_{v,w}^{(\ell)} \leq 2 \) with high probability. This, along with the fact that \( |\delta B(v, \ell)| \leq d^\ell \), part iii of Lemma 4.6 and our choice of \( \delta \) with \( \epsilon \) as in (4.3), gives us \( \|\tilde{\sigma}^{(s)}\|_{L_2} = o(1) \). This concludes the proof of ii of Proposition 4.4.

4.2. Proof of iii of Proposition 4.4. The proof follows the same lines of argument as that of [Lemma 8, [8]]. Let \( A \) denote the adjacency matrix and \( \overline{A} \) the expected adjacency matrix of \( G \). For any \( s \in \mathbb{N} \), let \( \Delta^{(s)} \) be the matrix where, for two vertices \( u \) and \( v \) in \( G \),

\[
\Delta^{(s)}_{u,v} = \sum_{t=1}^{s} \prod_{\ell=1}^{t} (A_{u_{\ell-1},u_{\ell}} - \overline{A}_{u_{\ell-1},u_{\ell}}),
\]

where the sum is taken over all length-\( s \) self-avoiding paths \( (u = u_0, u_1, \ldots, u_s = v) \) between \( u \) and \( v \). For each \( 1 \leq m \leq \ell \), with \( \ell \) as in (4.3), we define the matrix \( \Gamma^{(\ell,m)} \) where, for \( u, v \in V(G) \),

\[
\Gamma^{(\ell,m)}_{u,v} = \sum_{t=1}^{\ell-m} \prod_{\ell=1}^{t} (A_{u_{\ell-1},u_{\ell}} - \overline{A}_{u_{\ell-1},u_{\ell}}) A_{u_{\ell-m},u_{\ell-m+1}} \prod_{t=\ell-m+2}^{\ell} A_{u_{t-1},u_{t}},
\]

where the sum is taken over all paths of length \( \ell \) obtained by concatenating two self-avoiding paths \( (u = u_0, u_1, \ldots, u_{\ell-m}) \) and \( (u_{\ell-m+1}, \ldots, u_{\ell-m+1}, u_{\ell} = v) \) whose intersection is non-empty. The proof of [8, Theorem 2.2] goes through verbatim, since the proof does not involve the specific structure of the expected adjacency matrix. Thus we have

\[
S^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} \Delta^{(\ell-m)} \overline{A} S^{(m-1)} - \sum_{m=1}^{\ell} \Gamma^{(\ell,m)},
\]

with \( \ell \) as in (4.3).

The principal idea now is the same as that in [8] where, through Lemma 10, Lemma 11, Lemma 12, Proposition 3 and Lemma 13, it was shown that the first and the third terms have small spectral norm, so that it is enough to analyse the spectrum of the second term in (4.17). We first note that [8, Lemma 10, Lemma 11 and Lemma 12] go through verbatim for our set-up. The only change required in the statement of [8, Proposition 3] is as follows: for any edge \( e = \{u,v\} \in E(G) \), we define \( d_e = a_{i,j} \) if \( u \in C_i \) and \( v \in C_j \), for all \( i,j \in \{1, \ldots, k\} \); the proof, however, requires no modification for our set-up. The analogue of Lemma 13 of [8] is as follows.

**Lemma 4.7.** Let \( T \), with root \( \phi \), be a tree contained inside \( G \) and comprising \( |E(T)| = O(\log n) \) edges. Then

\[
\mathbb{E} \left[ \prod_{e \in E(T)} \frac{d_e}{n} \right] \leq \left( \frac{d}{kn} \right)^{|T|} \left\{ 1 + O\left( \frac{(\log n)^2}{n} \right) \right\}.
\]

**Proof.** We fix a leaf \( w \) of \( T \), and let \( \mathcal{F}_w \) denote the \( \sigma \)-field that tells us which cluster \( C_i \) each vertex \( v \in V(T) \setminus \{w\} \) belongs. Let \( s_j \) denote the number of vertices in \( V(T) \setminus \{w\} \) that are in \( C_j \), for each \( j \in \{1, \ldots, k\} \), and let \( s = \sum_{j=1}^{k} s_j \). Let \( u \) be the unique vertex in \( V(T) \) which is the parent to \( w \), and let \( e = \{u,w\} \). For any \( i \in \{1, \ldots, k\} \), if \( u \in C_i \), then \( d_e \in \{a_{i,j} : j = 1, \ldots, k\} \).

On the event \( \{d_e = a_{i,j}\} \), we have \( w \in C_j \). There are a total of \( kn - s - 1 \) many vertices in \( V(G) \setminus V(T) \), of which \( n - s_j - 1 \) many are to be assigned to the cluster \( C_j \), and \( n - s_j' \) many are to be assigned to the cluster \( C_j' \) for each \( j' \in \{1, \ldots, k\} \setminus \{j\} \). The number of such assignments is

\[
\frac{(kn-s-1)!}{(n-s_j-1)! \prod_{j' \in \{1,\ldots,k\} \setminus \{j\}} (n-s_{j'})}.
\]
On the other hand, given the information in $\mathcal{F}_w$, there are $kn - s$ many vertices in $V(\mathcal{G}) \setminus (V(T) \setminus \{w\})$, of which $n - s_j$ many are to be assigned to cluster $C_j$ for each $j' \in \{1, \ldots, k\}$. The number of such assignments is

$$\frac{(kn - s)!}{\prod_{j' \in \{1, \ldots, k\}} (n - s_{j'})!}.$$  

This shows that, since $s = |E(T)| = O(\log n)$,

$$P \left[ d_e = a_{i,j} | \mathcal{F}_w \right] = \frac{n - s_j}{kn - s} \leq \frac{1}{k} + O \left( \frac{s}{k(kn - s)} \right) = \frac{1}{k} + O \left( \frac{\log n}{n} \right).$$

The rest of the proof follows exactly as the proof of Lemma 13 of [8].

After this, the analysis of the spectral norms of both $\Delta^{(\ell)}$ and $\Gamma^{(\ell,m)}$, for each $1 \leq m \leq \ell$, follows mutatis mutandis from the corresponding analysis in [8] (see the derivation of equations (6.19), (6.20), (6.23) and (6.24) of [8]), and we deduce that

$$\max \left\{ ||\Delta^{(\ell)}||_{\text{spec}} , ||\Gamma^{(\ell,m)}||_{\text{spec}} \right\} \leq n^d \delta^{\ell/2}$$

for each $1 \leq m \leq \ell$.

where $|| \cdot ||_{\text{spec}}$ denotes the spectral norm. We now focus on the second term of (4.17). We emulate the argument for the derivation of [8, Equations (6.27) and (6.28)] to deduce, for each $1 \leq m \leq \ell$ and for any $\sigma$ satisfying the conditions in (4.16)

$$\max \left\{ ||\sigma^T S^{(m-1)} \left( \frac{e(nk)}{\sqrt{n^k}} \right) || , ||\sigma^T S^{(m-1)}(s) || , s = 2, \ldots, k \right\} \leq 3\delta^2 - 2^{-1/2} d^m$$

with high probability as $n \to \infty$, where $\delta$ is as in Lemma 4.6.

We now take a closer look at $\mathbb{A}$, which can be written, as $n \to \infty$, as $B - C$, where

$$B = \begin{bmatrix} \frac{a_{11}}{n} e(k) e(k)^T \\ \frac{a_{12}}{n} e(k) e(k)^T \\ \frac{a_{1k}}{n} e(k) e(k)^T \\ \vdots \\ \frac{a_{k1}}{n} e(k) e(k)^T \\ \frac{a_{kk}}{n} e(k) e(k)^T \end{bmatrix}$$

and

$$C = \begin{bmatrix} \frac{a_{11}}{n} I_k \\ 0_k \\ \frac{a_{22}}{n} I_k \\ \vdots \\ 0_k \\ \frac{a_{kk}}{n} I_k \end{bmatrix}$$

where $I_k$ and $0_k$ denote the $k \times k$ identity matrix and the $k \times k$ zero matrix respectively. By (4.14), it is clear that $d, \lambda_2, \ldots, \lambda_p$ are eigenvalues of the matrix $B$, with $e(nk)$ an eigenvector corresponding to $d$ and $\{ \sigma^{(s)} : s = R_{i-1} + 1, \ldots, R_i \}$ forming an orthonormal basis for the eigenspace corresponding to $\lambda_i$ for each $i = 2, \ldots, p$. From the structure of $B$, we conclude that its rank is at most $k$. Consequently, all other eigenvalues of $B$ are 0. The spectral decomposition of $B$ yields

$$B = d \left( \frac{e(nk)}{\sqrt{n^k}} \right) \left( \frac{e(nk)}{\sqrt{n^k}} \right)^T + \sum_{i=2}^p \sum_{s=R_{i-1}+1}^{R_i} \lambda_i \sigma^{(s)} \sigma^{(s)^T}.$$  

This representation, along with (4.18), allows us to write

$$\left| \left| BS^{(m-1)} \sigma \right| \right|_{L_2}^2 = d^2 \left| \sigma^T S^{(m-1)} \left( \frac{e(nk)}{\sqrt{n^k}} \right) \right|^2 + \sum_{i=2}^p \lambda_i^2 \sum_{s=R_{i-1}+1}^{R_i} \left| \sigma^T S^{(m-1)} \sigma^{(s)} \right|^2 \leq O \left( n^{\delta-1} d^{2m} \right).$$

which is $o(1)$ by our choice of $c$ and $\ell$ in (4.13), for all $1 \leq m \leq \ell$. On the other hand, the matrix $C$ has eigenvalues $\frac{a_{11}}{n}, \ldots, \frac{a_{kk}}{n}$, each with algebraic multiplicity $n$, so that its spectral norm is $\|C\|_{\text{spec}} = \frac{1}{n} \max \{ a_{i,j} : i = 1, \ldots, k \}$. The spectral norm of $S^{(m-1)}$ is bounded above by $O(d^{m-1})$. 

RSBM ON SEVERAL COMMUNITIES 21
This is enough to conclude that $|CS^{(m-1)}\sigma|^2_{L_2} = o(1)$. The final conclusion, i.e. (4.7), now follows as in the very last line of [3], Page 34.

4.3. Proof of [13] of Proposition 4.4. That $S^{(\ell)}$, with high probability, has the eigenvalues as specified in [13] and that any unit eigenvector corresponding to $\alpha_1$ is asymptotically aligned with $\nu \sqrt{nk}$, follows exactly as argued in the proof of Proposition 2 of [3] (using [3] and 13 as well as the Courant-Fischer Theorem, [24]). By the Perron-Frobenius Theorem ([19, 14]), we conclude that the algebraic multiplicity of $\alpha_1$ with respect to $S^{(\ell)}$ is 1.

We now focus on the remaining claims made in [13]. First, we assume, without any loss of generality, that $\alpha_2, \ldots, \alpha_p$ are distinct (if two or more of them coincide, we can consider that common value and sum the corresponding $R_i$ values) and re-label the $\alpha_i$’s so that we now have $\alpha_2 > \alpha_3 > \cdots > \alpha_p$. We note here that, because of this assumption, the $\alpha_i$ values specified in [13] and $\ell$ as in (4.3), we have $\alpha_i = o(\alpha_{i+1})$ for all $i = 1, \ldots, p - 1$.

We now show that $\alpha_i$ has algebraic multiplicity $r_i$ for each $i = 2, \ldots, p$, via induction on $i$. We omit the proof of the base case since its argument is exactly as that of the inductive step. We let $E_j$ denote the eigenspace of $S^{(\ell)}$ corresponding to $\alpha_j$ for each $j = 1, \ldots, p$. Suppose we have proved this claim for all $j \leq i - 1$, for some $i \in \{2, \ldots, p\}$ as $n \to \infty$.

If possible, let the algebraic multiplicity of $\alpha_i$ be strictly less than $r_i$, so that $\dim(E_i) \leq r_i - 1$. The subspace $W$ spanned by $E_i$ for all $j = 1, \ldots, i$ has dimension $\dim(W) \leq r_i - 1$. Consequently, $\dim(W^\perp) \geq nk - R_i + 1$, where $W^\perp$ denotes the orthogonal complement subspace of $W$. On the other hand, the subspace $U$ spanned by the mutually orthogonal vectors $e^{(nk)}, \sigma^{(2)}, \ldots, \sigma^{(R_i)}$ has dimension $\dim(U) = R_i$. This implies that there exists some unit vector $v \in U \cap W^\perp$. Let $v = \gamma_i e^{(nk)} + \sum_{s=2}^{R_i} \gamma_s \sigma^{(s)}$. This, along with [13] and [13] of Proposition 4.4 implies that

$$
|v^T S^{(\ell)} v| = \gamma_i^2 \alpha_1 + \sum_{j=2}^{i} \alpha_j \sum_{s=R_{j-1}+1}^{R_i} \gamma_s^2 + v^T \tilde{v} \geq \alpha_i - o(1),
$$

with $\tilde{v}$ denoting the error vector with $||\tilde{v}||_{L_2} = o(1)$. But $v \in W^\perp$, so that by the Courant-Fischer Theorem (23), we have $|v^T S^{(\ell)} v| \leq \alpha_{i+1}$. Since $\alpha_{i+1} = o(\alpha_i)$, this brings us to a contradiction.

Now assume that the algebraic multiplicity of $\alpha_i$ is strictly greater than $r_i$. The argument is very similar: now, $\dim(W) \geq R_i + 1$, whereas $\dim(U^\perp) = nk - R_i$, so that once again we can find a unit vector $v$ in $W \cap U^\perp$. By [13] and [13] of Proposition 4.4 we have $|v^T S^{(\ell)} v| \leq \alpha_{i+1} + o(1)$, whereas $v \in W$ implies that $|v^T S^{(\ell)} v| \geq \alpha_i$, thus leading to the same contradiction as above. This completes the proof of the claim that the algebraic multiplicity of $\alpha_i$ is $r_i$ for each $i = 2, \ldots, p$.

Let $\{w_s : s = 1, \ldots, nk\}$ be any orthonormal set of eigenvectors of $S^{(\ell)}$ such that $w_1$ corresponds to $\alpha_1$ and $\{w_s : s = R_{i-1} + 1, \ldots, R_i\}$ correspond to $\alpha_i$ for each $i = 2, \ldots, p$. We now show that, for each $i \in \{2, \ldots, p\}$ and each $s \in \{R_{i-1} + 1, \ldots, R_i\}$, the vector $\sigma^{(s)}$ asymptotically belongs to $E_i$. This, along with the above-established fact that, with high probability, $\dim(E_i) = r_i$, shows that the subspace $U_i$ spanned by $\{\sigma^{(s)} : s = R_{i-1} + 1, \ldots, R_i\}$ is asymptotically the same as $E_i$. This is enough for us to conclude that any unit eigenvector of $S^{(\ell)}$ corresponding to $\alpha_i$ asymptotically belongs to $U_i$.

To this end, fix $i$ and $s \in \{R_{i-1} + 1, \ldots, R_i\}$. Writing $\sigma^{(s)} = \sum_{t=1}^{nk} \gamma_t w_t$, we claim that

$$
\alpha_1 \gamma_1 \to 0 \quad \text{and} \quad \alpha_j \gamma_t \to 0 \quad \text{for each} \quad t \in \{R_{j-1} + 1, \ldots, R_j\}, \quad \text{for all} \quad j \leq i - 1
$$

as $n \to \infty$. The first part of (4.19) follows exactly as in the proof of Proposition 2, Page 22 of [3]. We now prove the rest of (4.19) by induction on $j$. Suppose we have established this for all
2 \leq j' \leq j - 1$ for some $j < i$. Then, for each $t \in \{R_{j-1} + 1, \ldots, R_j\}$, using ii of Proposition 4.4

$$
\alpha_j \gamma_t = \alpha_j \left< w_t, \sigma^{(s)} \right> = \left< S^{(\ell)} w_t, \sigma^{(s)} \right> = \left< w_t, S^{(\ell)} \alpha^{(s)} + \sigma^{(s)} \right> = \alpha_i \left< w_t, \sigma^{(s)} \right> + \left< w_t, \tilde{\sigma}^{(s)} \right> = \alpha_i \gamma_t + o(1),
$$

which, along with the fact that $\frac{\alpha_i}{\alpha_j} \to 0$ as $n \to \infty$, implies that $\alpha_j \gamma_t \to 0$ as $n \to \infty$. Finally, by ii and (4.19),

$$
\alpha_i^2 = \left\| S^{(\ell)} \sigma^{(s)} \right\|_{L_2}^2 + o(1) = \left\| S^{(\ell)} \sum_{t=1}^{nk} \gamma_t w_t \right\|_{L_2}^2 + o(1)
$$

\begin{align*}
&= (\alpha_1 \gamma_1)^2 + \sum_{j=2}^{i-1} \sum_{t=R_{j-1}+1}^{R_j} (\alpha_j \gamma_t)^2 + \sum_{s=R_{i-1}+1}^{R_i} (\alpha_i \gamma_s)^2 + \sum_{j=i+1}^{p} \sum_{t=R_{j-1}+1}^{R_j} (\alpha_j \gamma_t)^2 + \left\| S^{(\ell)} \sum_{t=k+1}^{nk} \gamma_t w_t \right\|_{L_2}^2 \\
&= o(1) + \sum_{s=R_{i-1}+1}^{R_i} (\alpha_i \gamma_s)^2 + O\left(\alpha_{i+1}^2\right) + O\left(\alpha_{p+1}^2\right) = \alpha_i^2 \sum_{s=R_{i-1}+1}^{R_i} \gamma_s^2 + o(1),
\end{align*}

since $\alpha_{i+1} = o(\alpha_i)$ and $\alpha_{p+1} = o(\alpha_i)$, as discussed earlier. This shows that we must have $\sum_{s=R_{i-1}+1}^{R_i} \gamma_s^2 \to 1$, thus establishing that $\sigma^{(s)}$ asymptotically belongs to $\mathcal{E}_i$.

Remark 4.8. To conclude this paper, we remark that one could include a result pertaining to strong recovery of the clusters $C_i$, $i = 1, \ldots, k$, in the same essence as [Theorem 2, [8]]. However, this would, in tandem with Theorem 4.2, require the implementation of the analogous version of the majority algorithm as described in [§1.0.1, [8]]. But simply emulating the argument of [§5, [8]] leads to the rather strong requirement of $a_{i,i} > \left(\frac{k-1}{2}\right) d$ for each $i \in \{1, \ldots, k\}$. It remains an open question to investigate in what ways said argument may be improved. We express our gratitude to Christopher Hoffman for sharing his novel ideas on improving the performance of the majority algorithm, and hope to pursue this avenue of thought in the near future.

References

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. The Journal of Machine Learning Research, 18(1):6446–6531, 2017.

[2] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. arXiv preprint arXiv:1512.09080, 2015.

[3] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. IEEE Transactions on Information Theory, 62(1):471–487, 2015.

[4] Edward A Bender. The asymptotic number of non-negative integer matrices with given row and column sums. Discrete Mathematics, 10(2):217–223, 1974.

[5] GUILLAUMEJL BLONDELVD et al. Fast unfolding of community hierarchies in large networks, 2008.

[6] Béla Bollobás and Béla Bollobás. Random graphs. Number 73. Cambridge university press, 2001.

[7] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, pages 1347–1357. IEEE, 2015.
Gerandy Brito, Ioana Dumitriu, Shirshendu Ganguly, Christopher Hoffman, and Linh V Tran. Recovery and rigidity in a regular stochastic block model. In Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms, pages 1589–1601. Society for Industrial and Applied Mathematics, 2016.

Gerandy Brito, Ioana Dumitriu, and Kameron Decker Harris. Spectral gap in random bipartite biregular graphs and applications. arXiv preprint arXiv:1804.07808, 2018.

Deepjyoti Choudhury and Arnab Paul. Community detection in social networks: an overview. International Journal of Research in Engineering and Technology, 2(2):6–13, 2013.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. Combinatorics, Probability and Computing, 19(2):227–284, 2010.

Anne Condon and Richard M Karp. Algorithms for graph partitioning on the planted partition model. Random Structures & Algorithms, 18(2):116–140, 2001.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Physical Review E, 84(6):066106, 2011.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. Physical Review Letters, 107(6):065701, 2011.

Martin E. Dyer and Alan M. Frieze. The solution of some random np-hard problems in polynomial expected time. Journal of Algorithms, 10(4):451–489, 1989.

Hossein Fani and Ebrahim Bagheri. Community detection in social networks. Encyclopedia with Semantic Computing and Robotic Intelligence, 1(01):1630001, 2017.

Joel Friedman. A proof of Alon’s second eigenvalue conjecture and related problems. American Mathematical Soc., 2008.

Joel Friedman. A proof of Alon’s second eigenvalue conjecture and related problems. American Mathematical Soc., 2008.

Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, Ferdinand Georg Frobenius, and Germany Mathematician. Über matrizen aus nicht negativen elementen. 1912.

Javier O Garcia, Arian Ashourvan, Sarah Muldoon, Jean M Vettel, and Danielle S Bassett. Applications of community detection techniques to brain graphs: Algorithmic considerations and implications for neural function. Proceedings of the IEEE, 106(5):846–867, 2018.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12):7821–7826, 2002.

Simon Heimlicher, Marc Lelarge, and Laurent Massoulié. Community detection in the labelled stochastic block model. arXiv preprint arXiv:1209.2910, 2012.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social networks, 5(2):109–137, 1983.

Roger A Horn and Charles R Johnson. Matrix analysis. Cambridge university press, 2012.

Mark Jerrum and Alistair Sinclair. Approximating the permanent. SIAM journal on computing, 18(6):1149–1178, 1989.

Mark Jerrum and Gregory B Sorkin. The metropolis algorithm for graph bisection. Discrete Applied Mathematics, 82(1-3):155–175, 1998.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. Physical review E, 83(1):016107, 2011.

Gregory F Lawler and Alan D Sokal. Bounds on the l2 spectrum for markov chains and markov processes: a generalization of cheegers inequality. Transactions of the American mathematical society, 309(2):557–580, 1988.
[29] Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labelled stochastic block model. IEEE Transactions on Network Science and Engineering, 2(4):152–163, 2015.

[30] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Statistical properties of community structure in large social and information networks. In Proceedings of the 17th international conference on World Wide Web, pages 695–704, 2008.

[31] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. Markov chains and mixing times, volume 107. American Mathematical Soc., 2017.

[32] Eyal Lubetzky, Allan Sly, et al. Cutoff phenomena for random walks on random regular graphs. Duke Mathematical Journal, 153(3):475–510, 2010.

[33] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In Proceedings of the forty-sixth annual ACM symposium on Theory of computing, pages 694–703. ACM, 2014.

[34] Marwa Ben Mbarek, Amel Borgi, Walid Bedhiäfi, and Sana Be Hmida. Genetic algorithm for community detection in biological networks. Procedia Computer Science, 126:195–204, 2018.

[35] Brendan D McKay and Xiaoji Wang. Asymptotic enumeration of 0–1 matrices with equal row sums and equal column sums. Linear algebra and its applications, 373:273–287, 2003.

[36] Brendan D McKay and Nicholas C Wormald. Asymptotic enumeration by degree sequence of graphs with degreeo (n 1/2). Combinatorica, 11(4):369–382, 1991.

[37] Frank McSherry. Spectral partitioning of random graphs. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science, pages 529–537. IEEE, 2001.

[38] Seyed Ahmad Moosavi, Mehrdad Jalali, Negin Misaghian, Shahaboddin Shamshirband, and Mohammad Hossein Anisi. Community detection in social networks using user frequent pattern mining. Knowledge and Information Systems, 51(1):159–186, 2017.

[39] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. arXiv preprint arXiv:1202.1499, 2012.

[40] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In Conference on Learning Theory, pages 356–370, 2014.

[41] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing, pages 69–75, 2015.

[42] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. Combinatorica, 38(3):665–708, 2018.

[43] Raj Rao Nadakuditi and Mark EJ Newman. Graph spectra and the detectability of community structure in networks. Physical review letters, 108(18):188701, 2012.

[44] Mark EJ Newman. Detecting community structure in networks. The European Physical Journal B, 38(2):321–330, 2004.

[45] Mark EJ Newman. Fast algorithm for detecting community structure in networks. Physical review E, 69(6):066133, 2004.

[46] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. Physical review E, 74(3):036104, 2006.

[47] Soumik Pal and Yizhe Zhu. Community detection in the sparse hypergraph stochastic block model. arXiv preprint arXiv:1904.05981, 2019.

[48] Tiago P Peixoto. Bayesian stochastic blockmodeling. Advances in network clustering and blockmodeling, pages 289–332, 2019.

[49] Oskar Perron. Zur theorie der matrices. Mathematische Annalen, 64(2):248–263, 1907.

[50] Sara Rahiminejad, Mano R Maurya, and Shankar Subramaniam. Topological and functional comparison of community detection algorithms in biological networks. BMC bioinformatics,
[51] Xiaoran Yan, Cosma Shalizi, Jacob E Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborová, Pan Zhang, and Yaojia Zhu. Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(5):P05007, 2014.