

Role of Normalization in Spectral Clustering for Stochastic Blockmodels

Purnamrita Sarkar
psarkar@eecs.berkeley.edu
Department of EECS
U. C. Berkeley

Peter J. Bickel
bickel@stat.berkeley.edu
Department of Statistics
U. C. Berkeley

October 8, 2013

Abstract

Spectral Clustering clusters elements using the top few eigenvectors of their (possibly normalized) similarity matrix. The quality of Spectral Clustering is closely tied to the convergence properties of these principal eigenvectors. This rate of convergence has been shown to be identical for both the normalized and unnormalized variants ([17]). However normalization for Spectral Clustering is the common practice ([16], [19]). Indeed, our experiments also show that normalization improves prediction accuracy. In this paper, for the popular Stochastic Blockmodel, we theoretically show that under spectral embedding, normalization shrinks the variance of points in a class by a constant fraction. As a byproduct of our work, we also obtain sharp deviation bounds of empirical principal eigenvalues of graphs generated from a Stochastic Blockmodel.

1 Introduction

Networks appear in many real-world problems. Any dataset of co-occurrences or relationships between pairs of entities can be represented as a network. For example, the Netflix data can be thought of as a giant bipartite network between customers and movies, where edges are formed via ratings. Facebook is a network of friends, where edges represent who knows whom. Weblogs link to other blogs and give rise to blog networks. Networks can also be implicit; for example, in machine learning they are often built by computing pairwise similarities between entities.

Many problems in machine learning and statistics are centered around community detection. Viral marketing functions by understanding how information propagates through friendship networks, and community detection is key to this. Link farms in the World Wide Web are basically malicious tightly connected clusters of webpages which exploit web-search algorithms to increase their rank. These need to be identified and removed so that search results are authentic and do not mislead users.

Spectral Clustering ([5], [7]) is a widely used network clustering algorithm. The main idea is to represent every entity by its coefficients along the top k eigenvectors of

a graph, and then cluster this spectral representation. Due to its computational ease and competitive performance, emerging application areas of Spectral Clustering range widely from parallel computing [11], CAD (computer aided design) [10], parallel sparse matrix factorization [18] to image segmentation [20], general clustering problems in Machine learning [16], and most recently, to fitting and classification using network block models [19, 21].

A Stochastic Blockmodel is a widely used generative model for networks with labeled nodes ([12], [19]). It assigns nodes to k different classes and forces all nodes in the same class to be stochastically equivalent. For example, in a two class Stochastic Blockmodel, any pair of nodes belonging to different classes link with probability γ_n (a deterministic quantity possibly dependent on the size of the graph, i.e. n).

Recently the consistency properties of Spectral Clustering in the context of Stochastic Blockmodels have attracted a significant amount of attention. Rohe et al. [19] showed that, under general conditions, for a sequence of normalized graphs with growing size generated from a Stochastic Blockmodel, Spectral Clustering yields the correct clustering in the limit. In a subsequent paper Sussman et al. [21] showed that an analogous statement holds for an unnormalized sequence of graphs. For finite k , the above results can also be obtained using direct applications of results from [17].

This prior theoretical work does not distinguish between normalized and unnormalized Spectral Clustering, and hence cannot be used to support the common practice of normalizing matrices for Spectral Clustering. Let us elucidate this issue by an example. We symmetrize the political blogs network [1], which is a directed network of hyperlinks connecting nodes representing weblogs about US politics. The nodes are labeled as “liberal” and “conservative” blogs. Misclassification rates using Spectral Clustering for the political blogs dataset are 4% for Normalized vs. 37% for Unnormalized for this dataset. In this paper, we present both theoretical arguments and empirical results to give a quantitative argument showing that normalization improves the quality of clustering.

While existing work [19, 21] bounds the classification accuracy, we do not take that route, since upper bounds can not be used to compare two methods. Instead, we focus on the variance within a class under the spectral representation using the top k eigenvectors. In this representation, by virtue of stochastic equivalence, points are identically distributed around their respective class centers. Hence the empirical variance can be computed using the average squared distance of points from their class center.

We show that for both the normalized and unnormalized methods, the squared distance between the two cluster centers is an order larger than the within class variance, thus leading to perfect classification in the limit. However, our results also indicate that the variance of points in a class increases as the graph gets sparser; hence methods which reduce the within-class variance are desired. We show that normalization indeed reduces the variance of points in a class by a constant fraction for a large parameter regime. Also, in this setting, the distance between the class centers can be thought of as bias; curiously, this distance approaches the same deterministic quantity with or without normalization. So, normalization does not change the bias, but shrinks the variance asymptotically.

We first prove that, for the simple case of disconnected clusters ($\gamma_n = 0$), the within class variance is asymptotically four times less when the matrix is normalized. While the simple case is on its own uninteresting, the proof idea is heavily used for the more general case. In the general setting, we derive asymptotic expressions of the within class variances; then by evaluating them we show that as γ_n increases, the ratio of within class variances (with normalization to without normalization) increases. For sparse graphs, this ratio is less than one for a large regime of γ_n (in particular, when γ_n is less than four fifths of the within class linkage probabilities).

This parameter regime with more within-class edges than across class edges can seem to be easy to cluster. However in sparse graphs, clustering this case can be relatively difficult. Of course, as n grows, both methods have enough data to distinguish between the clusters and behave similarly. But for small and sparse graphs, it is indeed an important regime. Sussman et al. [21] present a parameter setting where normalization is shown to hurt classification accuracy empirically. We show that this is but a partial picture; and in fact there is a large parameter regime where spectral clustering with normalized matrices yields tighter and hence better clusters.

Using quantifiable link prediction experiments on real world graphs and classifications tasks in labeled simulated graphs, we show that normalization leads to better classification accuracies for the regime dictated by our theory, and yields higher link prediction accuracy on sparse real world graphs.

2 Preliminaries and Import of Previous Work

In this paper we will only work with two class blockmodels. Given a binary $n \times 2$ class membership matrix Z , the edges in A are simply outcomes of $\binom{n}{2}$ independent Bernoulli coin flips. The Stochastic Blockmodel ensures stochastic equivalence of nodes within the same block; i.e. all nodes within the same block have identical probability of linking with other nodes in the graph.

Thus the conditional expectation matrix P can be described by three probabilities, namely $\alpha_n, \beta_n, \gamma_n$; where α_n and β_n denote the probabilities of connecting within the first and second classes (C_1 and C_2) respectively, and γ_n denotes the probability of connecting across two classes. All statements in this paper are conditioned on $\alpha_n, \beta_n, \gamma_n$ and Z .

Definition 2.1 (A Stochastic Blockmodel). *Let $Z \in \{0, 1\}^{n \times 2}$ be a fixed and unknown matrix of class memberships such that every row has exactly one 1, the first and second columns have $n\pi$ and $n(1 - \pi)$ ones respectively. A Stochastic Blockmodel with parameters $(\alpha_n, \beta_n, \gamma_n, Z)$ generates symmetric graphs with adjacency matrix A such that, $P(A(i, i) = 1) = 0, \forall i$. For $i > j$, $A_{ij} = A_{ji}$ are independent with $P(A_{ij} = 1|Z) = P_{ij}$, where P is symmetric with $P_{ij} = \alpha_n$ for $i, j \in C_1$, γ_n for $i \in C_1, j \in C_2$ and β_n for $i, j \in C_2$.*

For ease of exposition we will assume that the rows and columns of A are permuted such that all elements of the same class are grouped together. We have $P := E[A|Z]$. Clearly, P is a blockwise constant matrix with zero diagonal by construction.

We use a similar parametrization as [2] to allow for decaying edge probabilities as n grows. Formally α_n , β_n and γ_n are proportional to a common rate variable ρ_n where $\rho_n \rightarrow 0$ as $n \rightarrow \infty$, forcing all edge probabilities to decay at the same rate. Thus it suffices to replace α_n , β_n or γ_n by ρ_n in orders of magnitude, e.g. the expected degree of nodes in either class is $C_0 n \rho_n$. We use “ C_0 ” to denote a generic positive constant. All expectations are conditioned on Z ; for notational convenience we write $E[\cdot]$ instead of $E[\cdot|Z]$.

Since Spectral Clustering uses eigenvectors of A , the eigen-structure of P is of interest. For two symmetric matrices P_B , P_D , and $P := P_B + P_D$, Weyl’s inequality tells us that $\lambda_i(P_B) + \lambda_n(P_D) \leq \lambda_i(P) \leq \lambda_i(P_B) + \lambda_1(P_D)$, where $\lambda_i(\cdot)$ is the i^{th} largest eigenvalue of matrix (\cdot) . Let P_D be a diagonal matrix with first $n\pi$ elements along the diagonal as $-\alpha_n$ and remaining $n(1 - \pi)$ elements as $-\beta_n$. Hence P_B is a blockwise constant matrix of rank two. Thus, $\lambda_i(P_B)$ is $O(n\rho_n)$ for $i \in \{1, 2\}$, and zero otherwise. Also $\lambda_i(P_D)$ is of the form $-C_0\rho_n \forall i$. Hence we have:

$$\text{When } \alpha_n\beta_n \neq \gamma_n^2, \quad \lambda_i = \begin{cases} O_P(n\rho_n) & \text{For } i \leq 2 \\ O_P(\rho_n) & \text{Otherwise} \end{cases} \quad (1)$$

Let $\mathbf{v}_i(\lambda_i)$ denote the i^{th} eigenvector (eigenvalue) of matrix P . The ordering is in decreasing order of λ_i . We will denote the i^{th} empirical eigenvector (eigenvalue) by $\hat{\mathbf{v}}_i(\hat{\lambda}_i)$.

Now we will define the normalized counterparts of the above quantities. Let $\tilde{A} := D^{-1/2}AD^{-1/2}$, and also let $\tilde{P} := \mathcal{D}^{-1/2}P\mathcal{D}^{-1/2}$, where D and \mathcal{D} are the diagonal matrices of degrees and expected degrees respectively. We denote the first two eigenvectors by \mathbf{u}_1 and \mathbf{u}_2 , and the first two eigenvalues by ν_1 and ν_2 . Similar to P , \mathbf{u}_1 and \mathbf{u}_2 also are piecewise constant vectors, with corresponding coefficients \tilde{x}_1 , \tilde{y}_1 and \tilde{x}_2 , \tilde{y}_2 . The empirical counterparts of the eigenvectors and values are denoted by $\tilde{\mathbf{u}}_i$, $\tilde{\nu}_i$. One interesting fact about $\tilde{\mathbf{u}}_1$ is that the i^{th} entry is proportional to $\sqrt{\tilde{d}_i}$, where \tilde{d}_i is the degree of node i . However, one cannot explicitly obtain the form of $\hat{\mathbf{u}}_1(i)$.

Among the many variants of Spectral Clustering, we consider the algorithm used by [19]. The idea is to compute $n \times k$ matrix \hat{Q} with the top k eigenvectors of \tilde{A} along its columns, and applies the **kmeans** algorithm on the rows of \hat{Q} . The **kmeans** algorithm searches over different clusterings and returns a local optima of an objective function that minimizes the squared Euclidian distance of points from their respective cluster centers.

Probabilistic bounds on misclassification error of Spectral Clustering under the Stochastic Blockmodel has been obtained in previous work [19, 21]. However, upper bounds cannot be used for comparing two algorithms. It should also be noted that, analysis of misclassification error relies on **kmeans** achieving the global optimum, which is not guaranteed. Instead, we show that some simple clustering quality metrics are improved by normalization, and these metrics are computable in terms of an appropriately defined deviation of empirical eigenvectors from their population counterparts.

2.0.1 Quality Metrics

The quality metrics are defined as follows: the algorithm passes the empirical eigenvectors to an oracle who knows the cluster memberships. The oracle computes cluster centers $K_k := \sum_{i \in C_k} \hat{Q}_i / |C_k|$, for us $k \in \{1, 2\}$. Let d_{11}^2 denote the mean squared distance of points in C_1 from K_1 . To be concrete, we can write $\hat{d}_{11}^2 = \sum_{i \in C_1} \|\hat{Q}_i - K_1\|^2 / n\pi$. Similarly define \hat{d}_{12}^2 as the mean square distance of points in C_1 from K_2 , i.e. $\hat{d}_{12}^2 = \sum_{i \in C_1} \|\hat{Q}_i - K_2\|^2 / n\pi$. One can analogously define \hat{d}_{22}^2 and \hat{d}_{21}^2 . When the spectral embedding uses \tilde{A} we will denote the distances by \tilde{d}_{11}^2 and \tilde{d}_{12}^2 .

Although \hat{d}_{11}^2 seems like a simple average of squared distances; it actually has useful information about the quality of clustering. For definiteness, let us take the unnormalized case and examine points in C_1 . By stochastic equivalence, $\forall i \in C_1$, $\{\hat{v}_1(i), \hat{v}_2(i)\}$ are identically distributed (albeit dependent) random variables. Now \hat{d}_{11}^2 essentially is the trace of the 2×2 sample variance matrix, and hence measures the variance of these random variables.

Ideally a good clustering algorithm should satisfy $\hat{d}_{11}^2 / \hat{d}_{12}^2 \xrightarrow{P} 0$. But we will show that, this ratio converges to zero at the same rate, with or without normalization, in consistence with previous work ([17], [19], and [21]). Furthermore, we will show that $\tilde{d}_{12}^2 / \hat{d}_{12}^2 \xrightarrow{P} 1$, i.e. the two methods do not distinguish between \hat{d}_{12}^2 .

Interestingly, our results also imply that \hat{d}_{11}^2 becomes smaller as the graphs become sparser, i.e. ρ_n decreases. Hence, if a method can be shown to reduce the variance of points in a class by a constant fraction it would be preferable for sparse graphs. Indeed we show that $\tilde{d}_{11}^2 / \hat{d}_{11}^2$ converges to a constant which is *less than 1* for a broad range of parameter settings of α_n, β_n and γ_n . In the simple disconnected case with $\gamma_n = 0$, this constant is $1/4$.

Another advantage of \hat{d}_{11}^2 is that, it can be conveniently expressed in terms of an appropriately defined deviation of empirical eigenvectors from their population counterpart. For any population and empirical eigenvector pair $\{\mathbf{v}, \hat{\mathbf{v}}\}$, we consider the following orthogonal decomposition: $\mathbf{v} = c\hat{\mathbf{v}} + \mathbf{r}$, where $c := \mathbf{v}^T \hat{\mathbf{v}}$. The norm of residual \mathbf{r} will measure the deviation of $\hat{\mathbf{v}}$ from \mathbf{v} . The deviation of $\hat{\mathbf{u}}$ from \mathbf{u} can be measured similarly.

Since we are interested in two class blockmodels, we will mostly use $\mathbf{r}_i, i \in \{1, 2\}$ as the residual of the i^{th} empirical eigenvector from its population counterpart, and $c_{jj} := \mathbf{v}_j^T \hat{\mathbf{v}}_j$. We denote by $\mathbf{v}(C_1) := \sum_{i \in C_1} v(i) / n\pi$. A key fact is that $\mathbf{v}_1, \mathbf{v}_2$ (or $\mathbf{u}_1, \mathbf{u}_2$) are both piecewise constant.

$$d_{11}^2 = \frac{1}{c_{11}^2} \left(\sum_{i \in C_1} \frac{r_1(i)^2}{n\pi} - r_1(C_1)^2 \right) + \frac{1}{c_{22}^2} \left(\sum_{i \in C_1} \frac{r_2(i)^2}{n\pi} - r_2(C_1)^2 \right) \quad (2)$$

$$d_{12}^2 = \frac{1}{n\pi} \sum_{i \in C_1} \|\hat{Q}_i - K_2\|^2 = d_{11}^2 + \|K_1 - K_2\|^2 \quad (3)$$

We will denote the distances obtained from A by $\hat{d}_{..}$ and from \tilde{A} by $\tilde{d}_{..}$. For a wide regime of $(\alpha_n, \beta_n, \gamma_n)$, we prove that \tilde{d}_{11}^2 is asymptotically a constant factor smaller

and hence better than \widehat{d}_{11}^2 . First, using results from [8] we will prove that for $\gamma_n = 0$, $\widehat{d}_{11}^2 = 1/4\widehat{d}_{11}^2(1 + o_P(1))$. In this case, the result can be proven using existing results on Erdős-Rényi graphs [8] and a simple application of Taylor's theorem. In order to generalize the result to $\gamma_n \neq 0$, we would need new convergence results for A and \widetilde{A} generated from a Stochastic Blockmodel. All results rely on the following assumption on ρ_n :

Assumption 2.1. *We assume $\log n/n\rho_n \rightarrow 0$, as $n \rightarrow \infty$.*

This assumption ascertains with high probability that the sequence of growing graphs are not too sparse. The expected degree is $np = O(n\rho_n)$, and this is the most commonly used regime where norm convergence of matrices can be shown ([9], [17], [4]). Note that this is also the sharp threshold for connectivity of Erdős-Rényi graphs ([3]). We will now define a semi-sparse Stochastic Blockmodel, which is what we would work with for this paper.

Definition 2.2 (A semi-sparse Stochastic Blockmodel). *Define a Stochastic Blockmodel with parameters $\alpha_n, \beta_n, \gamma_n$ and Z (see Definition 2.1). Let α_n, β_n and γ_n be deterministic quantities of the form $C_0\rho_n$. If ρ_n satisfies Assumption 2.1, we call the Stochastic Blockmodel $(\alpha_n, \beta_n, \gamma_n, Z)$ a semi-sparse Stochastic Blockmodel.*

Here is how the paper is organized: we present the main results in Section 3. The proof of the simple $\gamma_n = 0$ case is in Section 4, whereas the expressions of \widehat{d}_{11}^2 and \widehat{d}_{12}^2 in the general case appear in Section 5. We derive the expressions of \widetilde{d}_{11}^2 and \widetilde{d}_{12}^2 in Section 6. Experiments on simulated and real data appear in Section 7. The proofs of accompanying lemmas presented in Sections 5 and 6 are provided in Appendices A, B respectively. Finally, proofs of all other results used in this paper are presented in Appendix C.

2.1 Import of Previous Work

By virtue of stochastic equivalence of points belonging to the same class, eigenvectors of P map the data to k distinct points. This is why consistency of Spectral Clustering is closely tied to consistency properties of empirical eigenvalues and eigenvectors. We will show that current theoretical work on eigenvector consistency does not distinguish between the use of normalized or unnormalized A .

One of the earlier results on consistency of Spectral Clustering can be found in [22], where weighted graphs generated from a geometric generative model are considered. While this is an important work, this does not apply to our random network model.

For any symmetric adjacency matrix A with independent entries, one can use results on random matrix theory from Oliveira [17] to show that the empirical eigenvectors of a semi-sparse Stochastic Blockmodel converge to their population counterpart at the same rate with or without normalization. If $\mathbf{p} : [n]^2 \rightarrow [0, 1]$ denotes the probability function $P(A_{ij} = 1) = 1 - P(A_{ij} = 0) = \mathbf{p}(i, j)$, and $d_{\mathbf{p}}$ denotes the expected degree, then:

Theorem 2.1. (Theorem 3.1 of [17]) *For any constant $c > 0$ there exists another constant $C = C(c) > 0$, independent of n or \mathbf{p} , such that the following holds. Let*

$d := \min_{i \in [n]} d_{\mathbf{p}}(i)$, $\Delta := \max_{i \in [n]} d_{\mathbf{p}}(i)$. If $\Delta > C \log n$, then for all $n^{-c} \leq \delta \leq 1/2$,

$$P \left(\|A - P\| \leq 4\sqrt{\Delta \log(n/\delta)} \right) \geq 1 - \delta.$$

Moreover, if $d \geq C \log n$, then for the same range of δ :

$$P \left(\|\tilde{A} - \tilde{P}\| \leq 14\sqrt{\log(4n/\delta)/d} \right) \geq 1 - \delta.$$

Let $\Pi_{a,b}(A)$ denote the orthogonal projector onto the space spanned by the eigenvectors of A corresponding to eigenvalues in $[a, b]$. A simple consequence of Theorem 2.1 is that for suitably separated population eigenvalues the operator norm of the difference of the eigenspaces also converges to zero.

Corollary 2.1. (Corollary 3.2 of [17]) Given some $x > 0$, let $N_x(P)$ be the set of all pairs $a < b$ such that $a + x < b - x$ and P has no eigenvalues in $(a - x, a + x) \cup (b - x, b + x)$. Then for $x > 4\sqrt{\Delta \log(n/\delta)}$,

$$\begin{aligned} \|A - P\| &\leq 4\sqrt{\Delta \log(n/\delta)} \\ \Rightarrow \forall (a, b) \in N_x(P), \|\Pi_{a,b}(A) - \Pi_{a,b}(P)\| &\leq \left(\frac{4(b - a + 2x)}{\pi(x^2 - x\sqrt{\Delta \log(n/\delta)})} \right) \sqrt{\Delta \log(n/\delta)} \end{aligned}$$

Similarly define $N_x(\tilde{P})$. Then for $x > 14\sqrt{\log(4n/\delta)/d}$,

$$\begin{aligned} \|\tilde{A} - \tilde{P}\| &\leq 14\sqrt{\frac{\log(4n/\delta)}{d}} \\ \Rightarrow \forall (a, b) \in N_x(\tilde{P}), \|\Pi_{a,b}(\tilde{A}) - \Pi_{a,b}(\tilde{P})\| &\leq \left(\frac{4(b - a + 2x)}{\pi(x^2 - x\sqrt{\log(n/\delta)/d})} \right) \sqrt{\log(n/\delta)/d} \end{aligned}$$

In particular the R.H.S.'s of the above equations hold with probability $\geq 1 - \delta$ for any $n^{-c} < \delta < 1/2$.

A straightforward application of this corollary yields that Spectral Clustering for Stochastic Blockmodel's with A and \tilde{A} lead to $O_P(\sqrt{\log n/n\rho_n})$ convergence of empirical eigenvectors to their population counterparts. Further analysis shows that the fraction of misclassified nodes go to zero at the same rate for A and \tilde{A} . We defer the proof to Appendix C.

Corollary 2.2. Let A be generated from a semi-sparse Stochastic Blockmodel (Definition 2.2) with $\gamma_n > 0$ and $\alpha_n \beta_n \neq \gamma_n^2$. Then, for $i \in \{1, 2\}$, $\|\mathbf{v}_i \mathbf{v}_i^T - \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T\| = O_P(\sqrt{\log n/n\rho_n})$. Furthermore the fraction of misclassified nodes can be bounded by $O_P(\log n/n\rho_n)$ for both methods.

Spectral Clustering with \tilde{A} derived from a Stochastic Blockmodel with growing number of blocks has been shown to be asymptotically consistent [19]. Further, the fraction of mis-clustered nodes is shown to converge to zero under general conditions. These results are extended to show that Spectral Clustering on unnormalized A also

enjoys similar asymptotic properties [21]. Sussman et al. [21] also give an example of parameter setting for a Stochastic Blockmodel where Spectral Clustering using unnormalized A outperforms that using \tilde{A} . We however demonstrate using theory and experiments that, this is only a partial picture, and there is a large regime of parameters where normalization indeed improves performance.

For the convenience of the reader, we list the different variables and their orders of magnitude in Table 1. For deterministic quantities x_n and c_n , $x_n \asymp c_n$, simply means that x_n/c_n converges to some constant as $n \rightarrow \infty$. For two random variables X_n and Y_n , we use $X_n \sim Y_n$ to denote $X_n = Y_n(1 + o_P(1))$. For the scope of this paper $\|\cdot\|$ will always mean the L_2 norm, unless otherwise specified.

Table 1: Table of notations.

ρ_n	Edge probability	I	The $n \times n$ identity matrix.
n	Number of nodes in the network	Z	$n \times 2$ binary matrix of class memberships
C_i	The i^{th} group, $i \in \{1, 2\}$	π	$ C_1 /n$
D	Diagonal matrix of degrees	\mathcal{D}	Diagonal matrix of expected degrees, conditioned on Z
A	Adjacency matrix	\tilde{A}	$D^{-1/2} A D^{-1/2}$
P	$E[A Z]$	\tilde{P}	$\mathcal{D}^{-1/2} P \mathcal{D}^{-1/2}$
α_n	$P[A_{ij} = 1 i \in C_1, j \in C_1] \asymp \rho_n$	μ_1	$\mathcal{D}_{ii}/n, i \in C_1; = \pi\alpha_n + (1 - \pi)\gamma_n - \alpha_n/n \asymp \rho_n$
β_n	$P[A_{ij} = 1 i \in C_2, j \in C_2] \asymp \rho_n$	μ_2	$\mathcal{D}_{ii}/n, i \in C_2; = \pi\gamma_n + (1 - \pi)\beta_n - \beta_n/n \asymp \rho_n$
γ_n	$P[A_{ij} = 1 i \in C_1, j \in C_2] \asymp \rho_n$	μ	$\sum_i \mathcal{D}_{ii}/n^2 = \pi\mu_1 + (1 - \pi)\mu_2 \asymp \rho_n$
d_i	$D_{ii}, i \in \{1, \dots, n\} = O_P(n\rho_n)$	\bar{d}_i	$\sum_j [A_{ij} - E[A_{ij} Z]] = O_P(\sqrt{n\rho_n})$
$\bar{d}_i^{(1)}$	$\sum_{j \in C_1} [A_{ij} - E[A_{ij} Z]] = O_P(\sqrt{n\rho_n})$	$\bar{d}_i^{(2)}$	$\sum_{j \in C_2} [A_{ij} - E[A_{ij} Z]] = O_P(\sqrt{n\rho_n})$
E_1	$\sum_{i \in C_1} d_i$	E_2	$\sum_{i \in C_2} d_i$
E	$\sum_i d_i$	$x(C_1)$	The average of \mathbf{x} restricted to C_1 , i.e. $\sum_{i \in C_1} \mathbf{x}(i)$
λ_i	i^{th} largest eigenvalue of P in magnitude $\asymp n\rho_n$, for $i \in \{1, 2\}$	ν_i	i^{th} largest eigenvalue of \tilde{P} in magnitude $\asymp 1$, for $i \in \{1, 2\}$
\mathbf{v}_i	i^{th} eigenvector of P	\mathbf{u}_i	i^{th} eigenvector of \tilde{P}
x_k	$\mathbf{v}_k(i), k \in \{1, 2\}, i \in C_1 \asymp 1/\sqrt{n}$	\tilde{x}_k	$\mathbf{u}_k(i), k \in \{1, 2\}, i \in C_1 \asymp 1/\sqrt{n}$
y_k	$\mathbf{v}_k(i), k \in \{1, 2\}, i \in C_2 \asymp 1/\sqrt{n}$	\tilde{y}_k	$\mathbf{u}_k(i), k \in \{1, 2\}, i \in C_2 \asymp 1/\sqrt{n}$
$\hat{\lambda}_i$	i^{th} largest eigenvalue of A in magnitude;	$\tilde{\nu}_i$	i^{th} largest eigenvalue of \tilde{A} in magnitude
$\hat{\mathbf{v}}_i$	i^{th} eigenvector of A	$\tilde{\mathbf{u}}_i$	i^{th} eigenvector of \tilde{A}
K_1	$\sum_{j \in C_1} \hat{Q}_j/n\pi$	K_2	$\sum_{j \in C_2} \hat{Q}_j/n\pi$
\hat{Q}	$n \times 2$ matrix of top two empirical eigenvectors (of A) along the columns	Q	The population variant of \hat{Q}
$\hat{d}_{k\ell}^2$	$\sum_{i \in C_k} \ \hat{Q}_i - K_\ell\ ^2/n\pi$	$\tilde{d}_{k\ell}^2$	Variant of $\hat{d}_{k\ell}^2$ using eigenvectors of \tilde{A}
C	$\hat{Q}^T Q$	c_{ij}	$C_{ij} := \mathbf{v}_i^T \hat{\mathbf{v}}_j$
$\hat{\mathbf{r}}_i$	$\mathbf{v}_i - (\mathbf{v}_i^T \hat{\mathbf{v}}_i) \hat{\mathbf{v}}_i, i \in \{1, 2\}$	$\tilde{\mathbf{r}}_i$	$\mathbf{u}_i - (\mathbf{u}_i^T \tilde{\mathbf{u}}_i) \tilde{\mathbf{u}}_i, i \in \{1, 2\}$

3 Main Results

For the general case we derive the following asymptotic expressions of d_{11}^2 and d_{12}^2 . We recall that d_{11}^2 measures the variance of points in class one under the spectral representation, whereas d_{12}^2 basically measures the distance between the class centers, which can also be thought of as bias. We will show that normalizing A asymptotically reduces the variance without affecting the bias. The proofs can be found in Sections 5 and 6. In Figure 1(A) we plot the ratio of $\tilde{d}_{11}^2/\hat{d}_{11}^2$ for a case where $\alpha_n = \beta_n$, $\pi = 1/2$. The figure shows that for a large regime of α_n, γ_n , in particular when $\gamma_n < \alpha_n$ this ratio is less than one.

Theorem 3.1. *Let A be the adjacency matrix generated from a semi-sparse Stochastic Blockmodel $(\alpha_n, \beta_n, \gamma_n, Z)$ where $\gamma_n > 0$ and $\alpha_n\beta_n \neq \gamma_n^2$. We define λ_i, x_i, y_i , for $i \in \{1, 2\}$ and π as in Table 1.*

$$\tilde{d}_{11}^2 \sim \left[\left(\frac{x_1^2}{\lambda_1^2} + \frac{x_2^2}{\lambda_2^2} \right) n\pi\alpha_n(1 - \alpha_n) + \left(\frac{y_1^2}{\lambda_1^2} + \frac{y_2^2}{\lambda_2^2} \right) n(1 - \pi)\gamma_n(1 - \gamma_n) \right] \quad (4)$$

$$\tilde{d}_{12}^2 \sim 1/n\pi(1 - \pi) \quad (5)$$

Theorem 3.2. *Let A be the adjacency matrix generated from a semi-sparse Stochastic Blockmodel $(\alpha_n, \beta_n, \gamma_n, Z)$ where $\gamma_n > 0$ and $\alpha_n\beta_n \neq \gamma_n^2$. We define μ_1, μ_2, ν_2 , and π as in Table 1.*

$$\tilde{d}_{11}^2 \sim \left[\frac{n\pi\alpha_n(1 - \alpha_n)}{n^3\pi\mu_1^2} \left(\frac{1}{4} + \frac{(1 - \pi)\gamma_n}{\mu_1\nu_2^2} \right) \right. \quad (6)$$

$$\left. + \frac{n(1 - \pi)\gamma_n(1 - \gamma_n)}{n^3\mu_1^2} \left(\frac{1}{4\pi} + \frac{\pi\alpha_n}{(1 - \pi)\mu_2\nu_2^2} \right) \right]$$

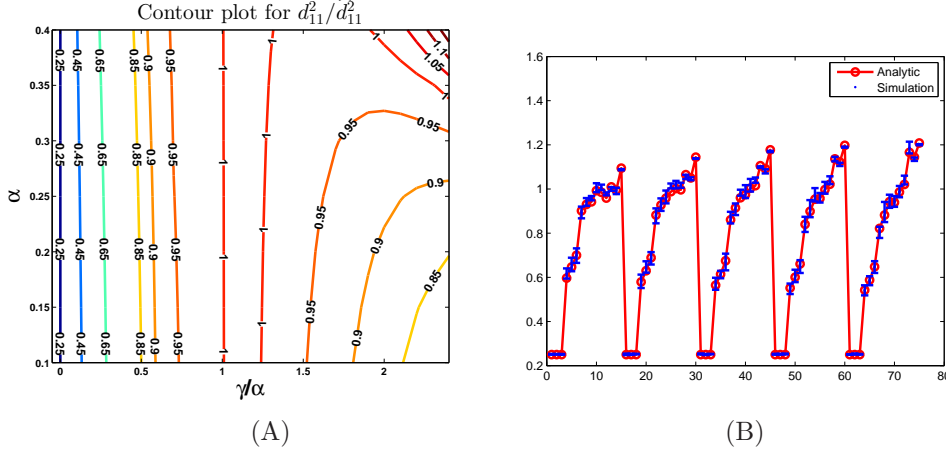
$$\tilde{d}_{12}^2 \sim \frac{1}{n\pi(1 - \pi)} \quad (7)$$

While both d_{11}^2 and d_{12}^2 both are approaching zero in probability, d_{11}^2/d_{12}^2 is $C_0/n\rho_n$ for both the normalized and unnormalized cases. In our regime of ρ_n this translates to perfect classification as $n \rightarrow \infty$. This is not unexpected because existing literature has established that Spectral Clustering with both A and \tilde{A} are consistent in our regime of ρ_n . Also, both \tilde{d}_{12}^2 and \hat{d}_{12}^2 approach the same constant; however for a large parameter regime, $\tilde{d}_{11}^2/\hat{d}_{11}^2$ approaches a constant fraction.

While it may seem that a constant fraction is not significant, one should recall that the d_{11}^2 measures the variance of the points in C_1 , whereas d_{12}^2 measures the distance between the two cluster centers. Later we will show that the variance of the points in a cluster increases as the edge probability ρ_n decreases. Hence, although in the limit both methods can differentiate between the clusters, for small sparse graphs shrinking the variance by a constant factor is indeed desirable.

Finally, in Figure 1(B) we plot the analytic ratio vs. ratios computed using simulations. For $n = 1000$, we vary $\alpha_n \in [0.4, 0.6]$, $\beta_n \in [0.5, 0.9]$ and γ_n/α_n between zero and two such that $\forall \alpha_n, \gamma_n \leq 1$, and $\gamma_n^2 \neq \alpha_n\beta_n$. We vectorize the three dimensional array of α_n, β_n and γ_n , and plot the ratio on the Y axis; every point on the X axis stands for a different setting.

Figure 1: (A) Contour plot for $\tilde{d}_{11}^2/\hat{d}_{11}^2$, for a blockmodel with $\pi = 1/2, \alpha_n = \beta_n$. Y axis has varying α_n , X axis has varying γ_n/α_n . (B) Analytic (red circle) vs. simulated ratios (blue dots with errorbars) $\tilde{d}_{11}^2/\hat{d}_{11}^2$. X axis has different settings of $(\alpha_n, \beta_n, \gamma_n)$ triples.



For each setting we compare the average $\tilde{d}_{11}^2/\hat{d}_{11}^2$ ratio (R in blue) by simulating ten Stochastic Blockmodel graphs with the analytic formula ($R_{analytic}$, in red). The mean, median and maximum of $|R(i, j, k) - R_{analytic}(i, j, k)|/R(i, j, k)$ are given by 0.01, 0.008 and 0.03. Each repetition of the sawtooth like pattern corresponds to increasing (α_n, γ_n) values for a fixed β_n value.

We note that the ratio increases for large (α_n, γ_n) pairs. The simulated experiment used by [21] where unnormalized performed better than normalized was with $\alpha_n = 0.42, \beta_n = 0.50, \gamma_n = 0.42$ (and $\pi = 0.60$), where the analytic ratio also is larger than one, and the graph is very close to an Erdős-Rényi graph. The mean, median and maximum absolute relative error for \tilde{d}_{12}^2 (\hat{d}_{12}^2) from $1/n\pi(1 - \pi)$ are 0.01, 0.003 and 0.15 (0.01, 0.004 and 0.17) respectively. In both cases the maximum happens for the $\{\alpha_n, \beta_n, \gamma_n\}$ combination where $|\alpha_n\beta_n - \gamma_n^2|$ is the smallest, leading to most instability. Since all our $o_P(1)$ terms are $O_P(1/\sqrt{n\rho_n})$, for this experiment these errors are indeed justifiable.

A SPECIAL CASE. When $\gamma_n = 0$, we have $\lambda_1 = n\mu_1$, and $x_1 = 1/\sqrt{n\pi}$, which immediately shows that normalization shrinks d_{11}^2 by a factor of four. We call this simple case the zero communication case, which can be thought of as two disconnected Erdős-Rényi graphs. Under Assumption 2.1 each of the smaller graphs will be connected with probability tending to one. Luxburg [15] already established that Spectral Clustering achieves perfect classification in this scenario. We merely present this simple setting because the ideas and proof techniques used here will be carried over to the following sections where $\gamma_n \neq 0$.

Corollary 3.1. *Let A be the adjacency matrix generated from a semi-sparse Stochastic Blockmodel $(\alpha_n, \beta_n, \gamma_n, Z)$ (see Definition 2.2) where $\gamma_n = 0$ and $\alpha_n\beta_n \neq \gamma_n^2$. We*

have:

$$\frac{\widehat{d}_{11}^2}{\widetilde{d}_{11}^2} \sim 4 \quad (8) \quad \frac{\widehat{d}_{11}^2}{\widetilde{d}_{12}^2} = O_P\left(\frac{1}{n\rho_n}\right) \quad (10)$$

$$\frac{\widehat{d}_{12}^2}{\widetilde{d}_{12}^2} \sim 1 \quad (9) \quad \frac{\widetilde{d}_{11}^2}{\widetilde{d}_{12}^2} = O_P\left(\frac{1}{n\rho_n}\right) \quad (11)$$

The same holds for normalized and unnormalized versions of d_{22}^2 and d_{21}^2 .

4 The Zero Communication Case

In this section we will present our result for two class block models (see Definition 2.2) where $\gamma_n = 0$. We will heavily use the following orthogonal decomposition of the population eigenvectors:

$$\mathbf{v}_k := c_{kk}\widehat{\mathbf{v}}_k + \mathbf{r}_k \text{ for } k \in \{1, 2\}, \text{ where } c_{kk} = \mathbf{v}_k^T \widehat{\mathbf{v}}_k.$$

Since $\gamma_n = 0$, A can be thought of as two disconnected Erdős-Rényi graphs of size $n\pi$ and $n(1 - \pi)$ (let the two adjacency matrices be denoted by A_1 and A_2 respectively). W.L.O.G we assume $\pi\alpha_n > (1 - \pi)\beta_n$ so that $\lambda_1 = n\pi\alpha_n + O(\rho_n)$ and $\lambda_2 = n(1 - \pi)\beta_n + O(\rho_n)$. We also assume that rows and columns of A are permuted so that the first $n\pi$ entries are from C_1 (we will not use this in our proofs, it only helps the exposition).

Füredi and Komlós [8] show that for $i \in \{1, 2\}$, $\widehat{\lambda}_i = \lambda_i + O_P(1)$, and $\max_{i>2} |\lambda_i| = O_P(\sqrt{n\rho_n})$. Hence for large n , the second largest eigenvalue will come from A_2 , and will have zeros along the first class, similar to the second population eigenvector. Thus, $\widehat{r}_1(i) = 0$ for $i \in C_2$, and vice-versa.

Further, little algebra reveals that $K_1 = \{\widehat{c}_{11}/\sqrt{n\pi}, 0\}$ and $K_2 = \{0, \widehat{c}_{22}/\sqrt{n(1 - \pi)}\}$. Computing \widehat{d}_{11}^2 or \widetilde{d}_{11}^2 requires one to compute the norm and average of $\widehat{\mathbf{r}}_k$ and $\widetilde{\mathbf{r}}_k$ restricted to C_1 (see Equation 2). For $\gamma_n = 0$, this reduces to examining $\widehat{\mathbf{r}}$ and $\widetilde{\mathbf{r}}$ for two Erdős-Rényi graphs.

Let us consider an Erdős-Rényi graph $G_{n,p}$. Since self loops are prohibited, the conditional expectation matrix P is simply $p(\mathbf{1}\mathbf{1}^T - I)$, which has n eigenvalues, the largest of which is $\lambda := (n - 1)p$, and the rest are all $-p$. We denote by d_i the degree of node i , and $\bar{d}_i := d_i - (n - 1)p$.

Let λ, \mathbf{v} be respectively the principal eigenvalue and eigenvector pair whose empirical counterparts are given by $\widehat{\lambda}, \widehat{\mathbf{v}}$. Also let $\widetilde{\lambda}, \widetilde{\mathbf{v}}$ be the corresponding quantities for \widetilde{A} . We require that all of $\widehat{\mathbf{v}}, \mathbf{v}$ and $\widetilde{\mathbf{v}}$ are unit-length. We denote by $\langle x_i \rangle$ the a n length vector with the i^{th} entry equaling x_i . We note that $\mathbf{v} = \langle 1/\sqrt{n} \rangle$, and $\widetilde{\mathbf{v}} = \left\langle \sqrt{d_i / \sum_j d_j} \right\rangle$. Let $\widehat{c} := \widehat{\mathbf{v}}^T \mathbf{v}$ and $\widetilde{c} := \widetilde{\mathbf{v}}^T \mathbf{v}$. We will prove that $\|\widehat{\mathbf{r}}\|^2 \sim 4\|\widetilde{\mathbf{r}}\|^2$, which will help us prove Corollary 3.1.

Before proceeding with the result, for ease of exposition we recall the orders of magnitudes of some random variables used in the proof. Let E denote $\sum_i d_i$. We

have $\sum_i \bar{d}_i = O_P(n\sqrt{\rho_n})$ (this is simply twice the sum of $\binom{n}{2}$ centered Bernoulli(p) variables) and $\sum_i \bar{d}_i^2 = n^2 p(1 + o_P(1))$. The later result can be obtained by showing that the expectation is $n(n-1)p$ and the standard deviation is of a smaller order. A detailed proof can be found in [8].

Lemma 4.1. *Write the first population eigenvector \mathbf{v} of an Erdős-Rényi (n, p) graph adjacency matrix A as $\mathbf{v} = \widehat{c}\widehat{\mathbf{v}} + \widehat{\mathbf{r}}$. If $p = O(\rho_n)$ satisfies Assumption 2.1, we have:*

$$\|\widehat{\mathbf{r}}\|^2 \sim \frac{1}{((n-1)p)^2} \frac{\sum_i \bar{d}_i^2}{n}$$

Proof. Before delving into the proof, we state the main result from [8]. For an Erdős-Rényi graph, $\widehat{\lambda}_1 = \frac{\mathbf{1}^T A \mathbf{1}}{n} + (1-p) + O_P(1/\sqrt{n})$. Since $\frac{\mathbf{1}^T A \mathbf{1}}{n} - (n-1)p = O_P(\sqrt{p(1-p)})$, we have: $\widehat{\lambda}_1 - (n-1)p = O_P(1)$. As the explicit form of $\widehat{\mathbf{v}}$ is not known, the following step is used to compute the norm of $\widehat{\mathbf{r}}$.

$$(A - \widehat{\lambda}_1 I)\widehat{\mathbf{r}} = \left\langle (d_i - \widehat{\lambda}_1) / \sqrt{n} \right\rangle \quad (12)$$

The proof is straight-forward. First we see that,

$$A\mathbf{v} - \widehat{\lambda}_1 \mathbf{v} = A(\widehat{c}\widehat{\mathbf{v}} + \widehat{\mathbf{r}}) - \widehat{\lambda}_1 \mathbf{v} = (A - \widehat{\lambda}_1 I)\widehat{\mathbf{r}}.$$

Using $\mathbf{v} = \langle \mathbf{1} / \sqrt{n} \rangle$, $A\mathbf{v} - \widehat{\lambda}_1 \mathbf{v} = \left\langle (d_i - \widehat{\lambda}_1) / \sqrt{n} \right\rangle$, thus proving Equation 12. Now Equation 12 and standard norm-inequalities yields $\|A - \widehat{\lambda}_1 I\| \leq (\widehat{\lambda}_1 + \max(\widehat{\lambda}_2, |\widehat{\lambda}_n|))$, where $\widehat{\lambda}_i$ is the i^{th} largest eigenvalue of A .

Now, using results from [6] we have $\max(\widehat{\lambda}_2, |\widehat{\lambda}_n|) = O_P(\sqrt{np})$, and hence $\|A - \widehat{\lambda}_1 I\| \sim np$. Interestingly, note that $\widehat{\mathbf{r}} \perp \widehat{\mathbf{v}}$, and hence $\|A\widehat{\mathbf{r}}\|/\|\widehat{\mathbf{r}}\| = O_P(\sqrt{np})$. Hence $\|(A - \widehat{\lambda}_1 I)\widehat{\mathbf{r}}\| \geq \widehat{\lambda}_1(1 + o_P(1))\|\widehat{\mathbf{r}}\|$. Combining this with the former upper bound, we have

$$\|(A - \widehat{\lambda}_1 I)\widehat{\mathbf{r}}\| \sim \widehat{\lambda}_1 \|\widehat{\mathbf{r}}\|.$$

Since, $\bar{d}_i/\sqrt{n} = O_P(\sqrt{p(1-p)})$, and $E[d_i] = (n-1)p$, we have:

$$\sum_i \frac{(d_i - \widehat{\lambda}_1)^2}{n} = \sum_i \frac{\bar{d}_i^2}{n} + (\widehat{\lambda}_1 - (n-1)p)^2 - 2(\widehat{\lambda}_1 - (n-1)p) \frac{\sum_i \bar{d}_i}{n} \sim \sum_i \frac{\bar{d}_i^2}{n}$$

The last step is true because $\sum_i \frac{\bar{d}_i^2}{n} = O_P(np(1-p))$, whereas both $\widehat{\lambda}_1 - (n-1)p$ and $\sum_i \bar{d}_i/n$ are $O_P(1)$. Simple application of the Cauchy Schwartz inequality shows that the cross term is also $O_P(1)$. Now we have:

$$\widehat{\mathbf{r}}^T \widehat{\mathbf{r}} = \frac{1}{\widehat{\lambda}_1^2} \sum_i \frac{(d_i - \widehat{\lambda}_1)^2}{n} \sim \frac{1}{((n-1)p)^2} \frac{\sum_i \bar{d}_i^2}{n} \quad (13)$$

□

Since the form of $\tilde{\mathbf{u}}$ is known, $\tilde{\mathbf{r}}^T \tilde{\mathbf{r}}$ can be obtained by using element-wise Taylor expansion. The only complication arises because we often approximate the norm of a length n vector by the norm of its first or second order Taylor expansion, where n is growing to infinity. Hence we present the following helping lemma, where we formalize sufficient conditions for neglecting lower order terms in such an expansion.

Lemma 4.2. *Consider length n vector $\mathbf{x}_n := c_n + \mathbf{x}_n^1 + \mathbf{R}_n$ where c_n is a vector of constants c . If both $\|\mathbf{R}_n\| = o_P(\|\mathbf{x}_n^1\|)$ and $|\sum_i x_n^1(i)/n| = o_P(\|\mathbf{x}_n^1\|/\sqrt{n})$, as $n \rightarrow \infty$, $\sum_i (x_n(i) - \sum_i x_n(i)/n)^2 \sim \|\mathbf{x}_n^1\|^2$.*

The following lemma has the asymptotic form of $\|\tilde{\mathbf{r}}\|^2$.

Lemma 4.3. *Write the first population eigenvector \mathbf{v} of an Erdős-Rényi (n, p) graph normalized adjacency matrix \tilde{A} as $\mathbf{u} = \tilde{c}\tilde{\mathbf{u}} + \tilde{\mathbf{r}}$. If $p = O(\rho_n)$ satisfies Assumption 2.1,*

$$\|\tilde{\mathbf{r}}\|^2 \sim \frac{1}{4n(n-1)p^2} \sum_i \frac{\bar{d}_i^2}{n}$$

Proof Sketch. We will use the fact that $\|\tilde{\mathbf{r}}\|^2 = 1 - \tilde{c}^2 = \sum_i (\tilde{u}_i - \sum_i \tilde{u}_i/n)^2$. Since one can explicitly obtain the expression of u_i , the basic idea is to use term by term Taylor approximation to obtain the norm. However, the issue is that we are summing over n elements where n is going to infinity, and extra care is required for the remainder terms; in particular, we will bound them uniformly over n .

It is easy to check that the vector $\langle \sqrt{d_i/E} \rangle$ is an eigenvector of \tilde{A} with eigenvalue one. By virtue of Assumption 2.1 we know that A is connected with high probability, and so the principal eigenvalue has multiplicity one. Thus $\tilde{\mathbf{u}}(i) = \sqrt{d_i/E}$. Now termwise Taylor approximation gives:

$$\tilde{u}_i = \frac{1}{\sqrt{n}} + \frac{\bar{d}_i}{2\sqrt{n(n-1)p^2}} + R, \quad (14)$$

where R is a length n vector of remainder terms. We will now invoke Lemma 4.2. Let c_n be the vector of constants $1/\sqrt{n}$, and $x_n^1 := \frac{\bar{d}_i}{2\sqrt{d_0 E_0}}$. Hence $\|x_n^1\| \sim C_0/\sqrt{n\rho_n}$, and the mean of x_n^1 is $O_P(1/\sqrt{n^3\rho_n}) = o_P(\|x_n^1\|/\sqrt{n})$. Using standard probabilistic arguments and the form of R we show that $\|R\| = o_P(\|1/\sqrt{n\rho_n}\|)$ (details in Appendix C). Hence we have:

$$\|\tilde{\mathbf{r}}\|^2 = \sum_i (\tilde{u}_i - \sum_i \tilde{u}_i/n)^2 \sim \frac{1}{4n(n-1)p^2} \sum_i \frac{\bar{d}_i^2}{n}.$$

□

PROOF OF COROLLARY 3.1. In order to compute \tilde{d}_{11}^2 and \hat{d}_{11}^2 , we need to compute the norms and averages of $\hat{\mathbf{r}}_k$ and $\tilde{\mathbf{r}}_k, k \in \{1, 2\}$ restricted to class C_1 . First note that $\hat{\mathbf{r}}_1(C_1) = \hat{\mathbf{r}}_1^T \mathbf{v}_1/\sqrt{n\pi} = \|\hat{\mathbf{r}}_1\|^2/\sqrt{n\pi}$ by construction, and $\hat{r}_2(i) = 0$, for $i \in C_1$. Hence $\sum_{i \in C_1} \hat{r}_1(i)^2 = \|\hat{\mathbf{r}}_1\|^2$.

Using Equation 2, $\hat{d}_{11}^2 = (\|\hat{\mathbf{r}}_1\|^2/n\pi - \|\hat{\mathbf{r}}_1\|^4/n\pi)/\hat{c}_{11}^2 = \|\hat{\mathbf{r}}_1\|^2/n\pi$. But $\|\hat{\mathbf{r}}_1\|^2$ is the norm-square of the residual of the principal eigenvector from A_1 which is a Erdős-Rényi $(n\pi, \alpha_n)$ graph (see Lemma 4.1).

Now we consider the corresponding quantities from \tilde{A} . The only issue is that A has two disconnected components (each of which are connected w.h.p. via Assumption 2.1), and hence \tilde{A} will have two eigenvalues equal to one; hence the first two eigenvectors can be any two orthogonal vectors spanning this eigenspace. Since Euclidean distances (e.g. $\hat{d}_{11}^2, \hat{d}_{12}^2$ etc.) are preserved under rotation, any such pair of vectors yield the same answer.

We will construct \mathbf{u}_1 and $\tilde{\mathbf{u}}_1$ as follows. \mathbf{u}_2 and $\tilde{\mathbf{u}}_2$ are defined analogously.

$$\mathbf{u}_1(i) = \begin{cases} 1/\sqrt{n\pi} & \text{For } i \in C_1 \\ 0 & \text{Otherwise} \end{cases} \quad \tilde{\mathbf{u}}_1(i) = \begin{cases} \sqrt{d_i/E} & \text{For } i \in C_1 \\ 0 & \text{Otherwise} \end{cases}$$

Since \mathbf{u} and \mathbf{v} are identical in the zero communication case, we have $\hat{d}_{11}^2 = \|\hat{\mathbf{r}}_1\|^2/n\pi$. However, $\tilde{\mathbf{r}}_1$ is simply the residual of the principal eigenvector from \tilde{A}_1 . Now an application of Lemmas 4.1 and 4.3 proves Equation 8.

As for \hat{d}_{12}^2 , note that $\hat{\mathbf{v}}(C_1) = \mathbf{v}_1^T \hat{\mathbf{v}}/\sqrt{n\pi} = \hat{c}_{11}/\sqrt{n\pi}$. Hence, $K_1 = \{\hat{c}_{11}/\sqrt{n\pi}, 0\}$ and $K_2 = \{\hat{c}_{22}/\sqrt{n(1-\pi)}, 0\}$. Thus $\|K_1 - K_2\|^2 = (1 - o_P(1))/n\pi(1-\pi)$, since both $\hat{c}_{11}^2 = 1 - \hat{\mathbf{r}}^T \hat{\mathbf{r}}$ and $\hat{c}_{22}^2 = 1 - \tilde{\mathbf{r}}^T \tilde{\mathbf{r}}$ are $1 - o_P(1)$ (Lemma 4.1). Since $\hat{d}_{11}^2 = O_P(1/n^2 \rho_n)$ is of smaller order than $\|K_1 - K_2\|^2$, using Equation 3 we see that $\hat{d}_{12}^2 \sim 1/n\pi(1-\pi)$. An identical argument shows that $\tilde{d}_{12}^2 \sim 1/n\pi(1-\pi)$ yielding Equation 9. With or without normalization, we have $d_{11}^2 = O_P(1/n^2 \rho_n)$, whereas $d_{12}^2 \sim 1/n\pi(1-\pi)$; this yields Equations 10 and 11. Finally, an identical argument proves the result for the normalized and unnormalized versions of d_{22}^2 and d_{21}^2 . \square

5 The General Case: Unnormalized A

In this section we obtain expressions for d_{11}^2 and d_{12}^2 when $\gamma_n \neq 0$ for A . First we give a simple lemma describing the eigen-structure of the conditional probability matrix P . The proof is simple and is deferred to the appendix.

Lemma 5.1. *Define a stochastic blockmodel (see Definition 2.2) with parameters $(\alpha_n, \beta_n, \gamma_n, Z)$, where $\gamma_n > 0$ and $\alpha_n \beta_n \neq \gamma_n^2$. The two population eigenvectors of P are piecewise constant with first $n\pi$ coefficients x_1 and x_2 respectively, and the second $n(1-\pi)$ coefficients y_1 and y_2 respectively. These coefficients are of the form C_0/\sqrt{n} , and they satisfy the following:*

$$x_1^2 + x_2^2 = 1/n\pi; \quad y_1^2 + y_2^2 = 1/n(1-\pi); \quad x_1 y_1 + x_2 y_2 = 0 \quad (15)$$

The two population eigenvalues λ_1 and λ_2 are of the form $C' n \rho_n$ and $C'' n \rho_n$, where C' and C'' are deterministic constants asymptotically independent of n ; also, $|\lambda_1 - \lambda_2|$ is of the form $C''' n \rho_n$ for some arbitrary constant C''' , when $\gamma_n > 0$.

We will now lay the groundwork for our result on \hat{d}_{11}^2 and \hat{d}_{12}^2 . In order to extend the simple zero-communication case to the general case, we will need some key results.

In this section, we first list those results, and then state the formal lemmas which show that such a result can actually be proved. The proofs are deferred to Appendix A. Recall the following decomposition of the population eigenvector:

$$\mathbf{v}_k = c_{kk} \widehat{\mathbf{v}}_k + \widehat{\mathbf{r}}_k \quad (16)$$

We would need the following three key components in order to use the same technique as in Lemma 4.1:

1. Sharp deviation of empirical eigenvalues. For $\gamma_n = 0$, we have $\widehat{\lambda}_k = \lambda_k + O_P(1)$.
2. Upper bound on $\|A\widehat{\mathbf{r}}_k\|$. For $\gamma_n = 0$, we have $\|A\widehat{\mathbf{r}}_k\| = O_P(1)$.
3. Bound on the average of $\widehat{\mathbf{r}}_k$ restricted to C_1 . For $\gamma_n = 0$ we have $\widehat{\mathbf{r}}_k(C_1) = O_P(1/n^{3/2}\rho_n)$.

In Section A we will provide detailed proofs of the following theorems, which show that the above results are also true when $\gamma_n \neq 0$.

In the following lemma we establish a sharp eigenvalue deviation result for block-models similar to the one for Erdős-Rényi graphs presented in [8]. Füredi and Komlós [8] use the Von Mises iteration (also popularly known as power iteration), which intuitively returns a good approximation of the principal eigenvalue in a few iterations if the second eigenvalue is much smaller than the first. In [8] the second largest eigenvalue of the adjacency matrix is shown to be an order smaller than the first; hence two steps of power iteration can be shown to give a $O_P(1)$ close approximation of $\widehat{\lambda}_1$. On the other hand, this approximation can also be shown to be $O_P(1)$ close to the population eigenvalue λ_1 , thus giving the sharp deviation bound.

In a Stochastic Blockmodel the second largest eigenvalue is of the same order as the first, which is problematic. However the third largest eigenvalue can be shown to be $O_P(\sqrt{n\rho_n \log n})$. Therefore we design a two-dimensional Von-Mises style iteration argument, so that at any step, the residual vector is orthogonal to the first two empirical eigenvectors and thus a $O_P(1)$ deviation of the empirical eigenvalues from their population counterparts can be proved. While we prove this result only for the two class blockmodels, the proof can be extended easily to k -class blockmodels.

Lemma 5.2. *Consider a n node network generated from a semi-sparse Stochastic Blockmodel $(\alpha_n, \beta_n, \gamma_n, Z)$ with $\gamma_n > 0$. We have,*

$$\text{For } i \in \{1, 2\}, \quad \widehat{\lambda}_i = \lambda_i + O_P(1)$$

Next we need to show that $\|A\widehat{\mathbf{r}}_k\| = O_P(1)$, $k \in \{1, 2\}$, even when $\gamma_n \neq 0$. For definiteness let $k = 1$. We want to emphasize that proving $\|\widehat{\mathbf{r}}_1\| = O_P(1/\sqrt{n\rho_n})$ is not enough to get the above. By construction $\widehat{\mathbf{r}}_1$ is orthogonal to $\widehat{\mathbf{v}}_1$, and hence $\|A\widehat{\mathbf{r}}_1\|$ can be upper bounded by $|\widehat{\lambda}_2| \|\widehat{\mathbf{r}}_1\|$. For an Erdős-Rényi graph, $\widehat{\lambda}_2 = O_P(\sqrt{n\rho_n})$ leading to an $O_P(1)$ bound, whereas for a Stochastic Blockmodel, $\widehat{\lambda}_2 = O_P(n\rho_n)$ leading to a $O_P(\sqrt{n\rho_n})$ bound. We show the required result by proving that $\widehat{\mathbf{v}}_2^T \widehat{\mathbf{r}}_1 = O_P(1/n\rho_n)$. Since $\widehat{\mathbf{v}}_1$ is orthogonal to $\widehat{\mathbf{v}}_2$, $\widehat{\mathbf{v}}_2^T \widehat{\mathbf{r}}_1 = \widehat{\mathbf{v}}_2^T \mathbf{v}_1$, which we prove to be $O_P(1/n\rho_n)$ in the following lemma.

Lemma 5.3. For a Stochastic Blockmodel $(\alpha_n, \beta_n, \gamma_n, Z)$ with $\gamma_n > 0$, define $\mathbf{v}_i = c_{ii}\hat{\mathbf{v}}_i + \hat{\mathbf{r}}_i, i \in \{1, 2\}$. Also define $c_{12} := \mathbf{v}_1^T \hat{\mathbf{v}}_2$ and $c_{21} := \mathbf{v}_2^T \hat{\mathbf{v}}_1$. We have

$$\begin{aligned} \|\hat{\mathbf{r}}_1\|^2 &= 1 - c_{11}^2 = O_P(1/n\rho_n) & \|\hat{\mathbf{r}}_2\|^2 &= 1 - c_{22}^2 = O_P(1/n\rho_n) \\ c_{12} &:= \mathbf{v}_1^T \hat{\mathbf{v}}_2 = O_P(1/n\rho_n) & c_{21} &:= \mathbf{v}_2^T \hat{\mathbf{v}}_1 = O_P(1/n\rho_n) \end{aligned}$$

The final task is to show that $\hat{\mathbf{r}}_k(C_1)$ and $\hat{\mathbf{r}}_k(C_2)$ are small. The Cauchy-Schwarz inequality gives $|\hat{\mathbf{r}}_k(C_1)| \leq \|\hat{\mathbf{r}}_k\|/\sqrt{n} = O_P(1/n\sqrt{\rho_n})$. However, for a Stochastic Blockmodel, by virtue of stochastic equivalence, \mathbf{v}_k for $k \in \{1, 2\}$ is piecewise constant, i.e. all entries in C_1 have value x_k , whereas all in C_2 have value y_k . Now, entries of $\hat{\mathbf{v}}_k$ in C_1 (C_2) constitute a noisy estimate of x_k (y_k). However, one should be able to get an even better estimate by considering $\hat{v}_k(C_1)$ and $\hat{v}_k(C_2)$. Since $\hat{\mathbf{r}}_k(C_1)$ reflects the error of $\hat{v}_k(C_1)$ around x_k , it is plausible that $\hat{\mathbf{r}}_k(C_1)$ is an order smaller than $\|\hat{\mathbf{r}}_k\|$, which is what we prove in the following lemma.

Lemma 5.4. Write $\mathbf{v}_i := c_{ii}\hat{\mathbf{v}}_i + \hat{\mathbf{r}}_i$ for $i \in \{1, 2\}$. Now we have,

$$\text{For } i, j \in \{1, 2\}, \quad \hat{\mathbf{r}}_i(C_j) = O_P(1/n^{3/2}\rho_n).$$

Before proceeding to prove our main result, we present the following simple concentration results, which are derived in the appendix.

Lemma 5.5. Denote $\bar{d}_i^{(1)}$ and $\bar{d}_i^{(2)}$ as the centered degree of node i restricted to blocks C_1 and C_2 respectively. We have:

$$\sum_{i \in C_1} (\bar{d}_i^{(1)})^2 \sim (n\pi)^2 \alpha_n (1 - \alpha_n); \quad \sum_{i \in C_1} (\bar{d}_i^{(2)})^2 \sim n^2 \pi (1 - \pi) \gamma_n (1 - \gamma_n) \quad (17)$$

$$\sum_{i \in C_1} (x_1 \bar{d}_i^{(1)} + y_1 \bar{d}_i^{(2)})^2 \sim \left(x_1^2 \sum_{i \in C_1} (\bar{d}_i^{(1)})^2 + y_1^2 \sum_{i \in C_1} (\bar{d}_i^{(2)})^2 \right). \quad (18)$$

Now we prove Theorem 3.1. Surprisingly, \hat{d}_{12}^2 can be shown to be $(1+o_P(1))/n\pi(1-\pi)$, which does not depend on the parameters α_n, β_n or γ_n .

5.1 Proof of Theorem 3.1

Proof. We will first prove Equation 4 and then Equation 5.

PROOF OF EQUATION 4. Define $\hat{\mathbf{r}}_i$ as in Equation 16. First note that $\|\hat{\mathbf{r}}_i\|^2 = O_P(1/n\rho_n)$ by Lemma 5.3. An argument similar to Lemma 4.1 gives:

$$\text{For } i \in \{1, 2\} \quad (A - \hat{\lambda}_i I) \hat{\mathbf{r}}_i = (A - P) \mathbf{v}_i + (\lambda_i - \hat{\lambda}_i) \mathbf{v}_i \quad (19)$$

As discussed earlier, we have $\hat{\mathbf{r}}_1^T \hat{\mathbf{v}}_2 = \mathbf{v}_1^T \hat{\mathbf{v}}_2$ since $\hat{\mathbf{v}}_1 \perp \hat{\mathbf{v}}_2$. But from Lemma 5.3, we know that $c_{12} = O_P(1/n\rho_n)$, and hence the projection of $\hat{\mathbf{r}}_1$ on the second eigen-space $\hat{\mathbf{v}}_2 \hat{\mathbf{v}}_2^T$ only contributes $\|\hat{\lambda}_2 c_{12} \hat{\mathbf{v}}_2\| = O_P(1)$. As $\hat{\lambda}_3 = O_P(\sqrt{n\rho_n \log n})$, $\|A \hat{\mathbf{r}}_1\| = O_P(1)$.

We compute \hat{d}_{11}^2 by deriving asymptotic expressions of $1/n\pi \sum_{i \in C_1} \hat{\mathbf{r}}_k(i)^2 - \hat{\mathbf{r}}_k(C_1)^2$, $k \in \{1, 2\}$. First we show that the second term is of lower order than the first. This is because $\sum_{i \in C_1} \hat{\mathbf{r}}_1(i)^2/n\pi \leq \|\hat{\mathbf{r}}_1\|^2/n\pi = O_P(1/n^2\rho_n)$, but $\hat{\mathbf{r}}_1(C_1)^2 = O_P(1/n^3\rho_n^2)$

using Lemma 5.4. We will now focus on the elements of $\widehat{\mathbf{r}}_1$ belonging to C_1 . We also denote by $\widehat{\mathbf{r}}_1(1)$ the subset of $\widehat{\mathbf{r}}_1$ indexed by nodes in C_1 , and thus by $[A\widehat{\mathbf{r}}_1](1)$ the subset of vector $A\widehat{\mathbf{r}}_1$ indexed by C_1 . Also note that, $\|[A\widehat{\mathbf{r}}_1](1)\|^2 \leq \|A\widehat{\mathbf{r}}_1\|^2 = O_P(1)$.

$$[A\widehat{\mathbf{r}}_1](1) - \widehat{\lambda}_1 \widehat{\mathbf{r}}_1(1) = [(A - P)v_1](1) + (\lambda_1 - \widehat{\lambda}_1)v_1(1)$$

$$\sum_{i \in C_1} \widehat{r}_1(i)^2 \sim \frac{\sum_{i \in C_1} (x_1 \bar{d}_i^{(1)} + y_1 \bar{d}_i^{(2)})^2}{\lambda_1^2}$$

The last step is valid because, $\|(A - P)v_1\|$ can be shown to be $O_P(\sqrt{n\rho_n})$ (see Appendix A) whereas $\|[A\widehat{\mathbf{r}}_1](1)\| = O_P(1)$ and $\|(\lambda_1 - \widehat{\lambda}_1)v_1(1)\| = O_P(1)$ using Lemma 5.2. Similarly, $\sum_{i \in C_1} \widehat{r}_2(i)^2 \sim \sum_{i \in C_1} (x_2 \bar{d}_i^{(1)} + y_2 \bar{d}_i^{(2)})^2 / \lambda_2^2$. Hence using Lemma 5.5, Equation 18 we have:

$$\widehat{d}_{11}^2 \sim \frac{1}{n\pi} \left[\left(\frac{x_1^2}{\lambda_1^2} + \frac{x_2^2}{\lambda_2^2} \right) \sum_{i \in C_1} (\bar{d}_i^{(1)})^2 + \left(\frac{y_1^2}{\lambda_1^2} + \frac{y_2^2}{\lambda_2^2} \right) \sum_{i \in C_1} (\bar{d}_i^{(2)})^2 \right].$$

Now Lemma 5.5, Equation 17 yields Equation 4.

PROOF OF EQUATION 5.

We recall that Equation 3 gives: $\widehat{d}_{12}^2 = \widehat{d}_{11}^2 + \|K_1 - K_2\|^2$, where $K_k = \{\widehat{v}_1(C_k), \widehat{v}_2(C_k)\}$, $k \in \{1, 2\}$. From Equation 16, we see that $\widehat{v}_i(C_1) = (v_i(C_1) - r_i(C_1))/c_{ii}$, and hence we have:

$$\widehat{v}_i(C_1) - \widehat{v}_i(C_2) = \left(\frac{x_i - y_i}{c_{ii}} \right) - \left(\frac{r_i(C_1) - r_i(C_2)}{c_{ii}} \right) \quad i \in \{1, 2\}$$

We will simply show that $\|K_1 - K_2\|^2 = ((x_1 - y_1)^2 + (x_2 - y_2)^2)(1 + o_P(1))$, which is $= 1 + o_P(1)/n\pi(1 - \pi)$ using Lemma 5.1 (Equation 15). Since x_1, y_1, x_2, y_2 are of the form C_0/\sqrt{n} and $c_{ii}^2 = 1 - r_{ii}^2$, we can show that

$$\sum_{i=1}^2 \left(\frac{x_i - y_i}{c_{ii}} \right)^2 = \frac{1}{n\pi(1 - \pi)} + O_P(1/n^2\rho_n).$$

Also, Lemma 5.4 shows that for $i \in \{1, 2\}$, $r_i(C_1) = O_P(1/n^{3/2}\rho_n)$, and hence we have Equation 5. \square

6 The General Case: Normalized A

As discussed in Section 4, both ν_1 and $\widetilde{\nu}_1$ (see Table 1) equal one, and $\widetilde{\mathbf{u}}_1(i) = \sqrt{d_i/E}$. In our analysis what naturally appears is the following notion of density, defined by expected degree over n . All expectations are conditioned on Z . Let μ_1 and μ_2 the $E[d_i|Z]/n$ for i in C_1 and C_2 respectively. Also let $\mu = \sum_{ij} P_{ij}/n^2$. Hence $\mu_1 := \pi\alpha_n + (1-\pi)\gamma_n - \alpha_n/n$, and $\mu_2 = (1-\pi)\beta_n + \pi\gamma_n - \beta_n/n$, and $\mu = \pi\mu_1 + (1-\pi)\mu_2$.

Also, we recall that $\bar{d}_i^{(1)}$ is the centered $d_i^{(1)}$, i.e. $d_i^{(1)} - n\pi\alpha_n$ when $i \in C_1$ and $d_i^{(1)} - n\pi\gamma_n$ when $i \in C_2$.

The eigenvalues of \tilde{P} can be obtained easily by applying Weyl's inequality to its rank two version, which does not have the restriction of zero diagonals. The argument is analogous to that for eigenvalues of P , and follows in a manner similar to that appearing in the paragraph preceding Equation 1. The expressions for the two principal eigenvalues and eigenvectors of \tilde{P} are stated below. Its proof is deferred to Appendix B.

Lemma 6.1. *Define a semi-sparse Stochastic Blockmodel (see Definition 2.2) with parameters $(\alpha_n, \beta_n, \gamma_n, Z)$, where $\alpha_n\beta_n \neq \gamma_n^2$. The principal eigenvalues ν_1 and ν_2 , and the blockwise entries $\tilde{x}_1, \tilde{y}_1, \tilde{x}_2$ and \tilde{y}_2 of the principal eigenvectors of \tilde{P} are given by:*

$$\begin{aligned} \nu_1 &= 1 & \tilde{x}_1 &= \sqrt{\frac{\mu_1}{n\mu}} & \tilde{y}_1 &= \sqrt{\frac{\mu_2}{n\mu}} \\ \nu_2 &= 1 - \frac{\gamma_n\mu}{\mu_1\mu_2} & \tilde{x}_2 &= \sqrt{\frac{(1-\pi)\mu_2}{n\pi\mu}} & \tilde{y}_2 &= -\sqrt{\frac{\pi\mu_1}{n(1-\pi)\mu}} \end{aligned}$$

For all other eigenvalues of \tilde{P} we have $|\nu_i| = O(1/n)$.

In order to obtain \tilde{d}_{11}^2 (Equation 2), we need $\sum_{i \in C_1} (\tilde{u}_1(i) - \tilde{u}_1(C_1))^2$. Using $\tilde{u}(C_1) = \sum_{i \in C_1} \tilde{u}(i)/n\pi$ and arguing as in Lemma 4.3, we see that:

$$\sum_{i \in C_1} (\tilde{u}_1(i) - \tilde{u}_1(C_1))^2 \sim \frac{1}{4n^3\mu\mu_1} \sum_{i \in C_1} \bar{d}_i^2. \quad (20)$$

Computing $\sum_{i \in C_1} (\tilde{u}_2(i) - \tilde{u}_2(C_1))^2$ requires more in-depth analysis, since $\tilde{\mathbf{u}}_2$ cannot be expressed in closed form as $\tilde{\mathbf{u}}_1$. Instead we look at a “good” approximation of $\tilde{\mathbf{u}}_2$, such that the approximation error cannot mask its $O_P(1/\sqrt{n\rho_n})$ deviation from the population counterpart \mathbf{u}_2 . The very first guess is to construct a vector orthogonal to $\tilde{\mathbf{u}}_1$. In that effect we present \mathbf{u}_g^0 as in Equation 21. Define $E_1 := \sum_{i \in C_1} d_i$, and $E_2 := \sum_{i \in C_2} d_i$.

$$u_g^0(i) = \begin{cases} \frac{\sqrt{d_i}}{E_1} & \text{For } i \in C_1 \\ -\frac{\sqrt{d_i}}{E_2} & \text{For } i \in C_2 \end{cases} \quad (21)$$

In spite of being a fair guess, u_g^0 masks the $O_P(1/\sqrt{n\rho_n})$ error. So we take a Von-Mises iteration step starting with \mathbf{u}_g^0 , and get a finer approximation, namely \mathbf{u}_g . We now present element-wise Taylor expansions of \mathbf{u}_g similar to Section 4.

Lemma 6.2. Define \mathbf{u}_g^0 as in Equation 21. We have:

$$[\tilde{A}\mathbf{u}_g^0]_i = \begin{cases} \frac{\nu_2}{n\pi\sqrt{n\mu_1}} \left(1 - \frac{\bar{d}_i}{2n\mu_1} + \frac{\bar{d}_i^{(1)}}{n\mu_1\nu_2} - \frac{\bar{d}_i^{(2)}}{n\mu_2\nu_2} \frac{\pi}{1-\pi} + M_i \right) & i \in C_1 \\ -\frac{\nu_2}{n(1-\pi)\sqrt{n\mu_2}} \left(1 - \frac{\bar{d}_i}{2n\mu_2} - \frac{\bar{d}_i^{(1)}}{n\mu_1\nu_2} \frac{1-\pi}{\pi} + \frac{\bar{d}_i^{(2)}}{n\mu_2\nu_2} + M'_i \right) & i \in C_2 \end{cases}$$

The remainder vectors M and M' are of norm $o_P(C_0/\sqrt{\rho_n})$

$$\|\tilde{A}\mathbf{u}_g^0\| \sim \nu_2 \sqrt{\frac{\mu}{n^2\pi(1-\pi)\mu_1\mu_2}}$$

The next lemma shows that \mathbf{u}_g has an approximation error of $O_P(\sqrt{\log n/n^2\rho_n^2})$. The proof again is deferred to Appendix B.

Lemma 6.3. Define $\mathbf{u}_g := \tilde{A}\mathbf{u}_g^0/\|\tilde{A}\mathbf{u}_g^0\|$. Let $c_g := (\tilde{\mathbf{u}}_2)^T \mathbf{u}_g$, i.e. the projection of \mathbf{u}_g on $\tilde{\mathbf{u}}_2$ and $\mathbf{r}_g := \mathbf{u}_g - c_g \tilde{\mathbf{u}}_2$. We have,

$$\|\mathbf{r}_g\| = O_P\left(\sqrt{\frac{\log n}{n^2\rho_n^2}}\right); \quad c_g = 1 - o_P(1)$$

Now we are ready to derive the expressions of \tilde{d}_{11}^2 and \tilde{d}_{12}^2 (Theorem 3.2).

6.1 Proof of Theorem 3.2

Proof. We will first prove Equation 6 and then Equation 7.

PROOF OF EQUATION 6. Computing \tilde{d}_{11}^2 only involves the entries of $\tilde{\mathbf{u}}_2$ indexed by nodes in C_1 ; hence we will apply Lemma 4.2 on $\tilde{u}_2(i), i \in C_1$. Using our construction:

$$\tilde{\mathbf{u}}_2 = (\mathbf{u}_g - \mathbf{r}_g)/c_g \quad \text{where } c_g = 1 - o_P(1). \quad (22)$$

Using Lemma 6.2, for $i \in C_1$, we can write each term of \mathbf{u}_g as:

$$u_g(i) = \chi_n(1 + x_n^1(i) + M_i),$$

where \mathbf{x}_n^1 and M are the first and remainder terms in the Taylor expansion of $u_g(i)/\chi_n$. We have:

$$\begin{aligned} \chi_n &:= \frac{\nu_2}{n\pi\sqrt{n\mu_1}\|\tilde{A}\mathbf{u}_g^0\|} \sim \sqrt{\frac{(1-\pi)\mu_2}{n\pi\mu}} \\ x_n^1(i) &:= \frac{\bar{d}_i}{2n\mu_1} + \frac{\bar{d}_i^{(1)}}{n\mu_1\nu_2} - \frac{\bar{d}_i^{(2)}}{n\mu_2\nu_2} \frac{\pi}{1-\pi}, \quad i \in C_1 \end{aligned}$$

We have:

$$\sum_{i \in C_1} \left(\frac{u_g(i) - u_g(C_1)}{\chi_n} \right)^2 = \sum_{i \in C_1} \left((x_n^1(i) - x_n^1(C_1)) + (M_i - M(C_1)) \right)^2$$

While $\|\mathbf{x}_n^1\| = C_0/\sqrt{\rho_n}$ (Lemma 5.5, Equation 18), $x_n^1(C_1) = O_P(1/\sqrt{n^2\rho_n})$, since it involves averages of $O(n^2)$ independent Bernoulli's. Also $\|M\| = o_P(1/\sqrt{\rho_n})$, and hence using a simple application of Cauchy-Schwarz inequality, one has:

$$\sum_{i \in C_1} (u_g(i) - u_g(C_1))^2 \sim \chi_n^2 \sum_{i \in C_1} x_n^1(i)^2. \quad (23)$$

Finally, since $\|\mathbf{r}_g\|^2 = O_P(\log n/(n\rho_n)^2)$ and $\sum_{i \in C_1} (u_g(i) - u_g(C_1))^2 = C_0/n\rho_n$, from Equations 22 and 23 we have:

$$\frac{1}{n\pi} \sum_{i \in C_1} (\tilde{u}_2(i) - \tilde{u}_2(C_1))^2 \sim \frac{\chi_n^2}{n\pi} \sum_{i \in C_1} x_n^1(i)^2 \quad (24)$$

With a little algebra Equations 20 and 24 give:

$$\begin{aligned} \tilde{d}_{11}^2 &\sim \frac{1}{n\pi} \sum_{i \in C_1} \left(\frac{\mu_1}{n\mu} \frac{\bar{d}_i^2}{4n^2\mu_1^2} + \frac{(1-\pi)\mu_2}{n\pi\mu} \left(-\frac{\bar{d}_i}{2n\mu_1} + \frac{\bar{d}_i^{(1)}}{n\mu_1\nu_2} - \frac{\bar{d}_i^{(2)}}{n\mu_2\nu_2} \frac{\pi}{1-\pi} \right)^2 \right) \\ &\sim \frac{1}{n\pi} \sum_{i \in C_1} \left[\frac{(\bar{d}_i^{(1)})^2}{n^3\pi\mu_1^2} \left(\frac{1}{4} + \frac{(1-\pi)\gamma_n}{\mu_1\nu_2^2} \right) + \frac{(\bar{d}_i^{(2)})^2}{n^3\mu_1^2} \left(\frac{1}{4\pi} + \frac{\pi\alpha_n - \alpha_n/n}{(1-\pi)\mu_2\nu_2^2} \right) \right] \end{aligned}$$

The last step uses Lemma 5.5 (Equation 18).

PROOF OF EQUATION 7. Equation 3 gives, $\tilde{d}_{12}^2 = \tilde{d}_{11}^2 + \|K_1 - K_2\|$. $K_i := \{\tilde{u}_1(C_i), \tilde{u}_2(C_i)\}$ for $i \in \{1, 2\}$. The Taylor expansion used in Lemma 4.3 shows that the second order terms are $o_P(1/n)$ whereas the first is of the form C_0/n . For $\mu_1 \neq \mu_2$, neglecting second order terms gives:

$$(\tilde{u}_1(C_1) - \tilde{u}_1(C_2))^2 \sim \frac{(\sqrt{\mu_1} - \sqrt{\mu_2})^2}{n\mu} \quad (25)$$

For the second part, Equation 22 and an argument shown earlier gives:

$$\begin{aligned} (\tilde{u}_2(C_1) - \tilde{u}_2(C_2))^2 &\sim (u_g(C_1) - u_g(C_2))^2 \\ &\sim \left(\sqrt{\frac{(1-\pi)\mu_2}{n\pi\mu}} + \sqrt{\frac{\pi\mu_1}{n(1-\pi)\mu}} \right)^2 \end{aligned} \quad (26)$$

Putting Equations 25 and 26 together yields Equation 7. When $\mu_1 = \mu_2$, the whole contribution comes from the second eigenvector, and \tilde{u}_1 only contributes $o_P(1/n)$ terms. \square

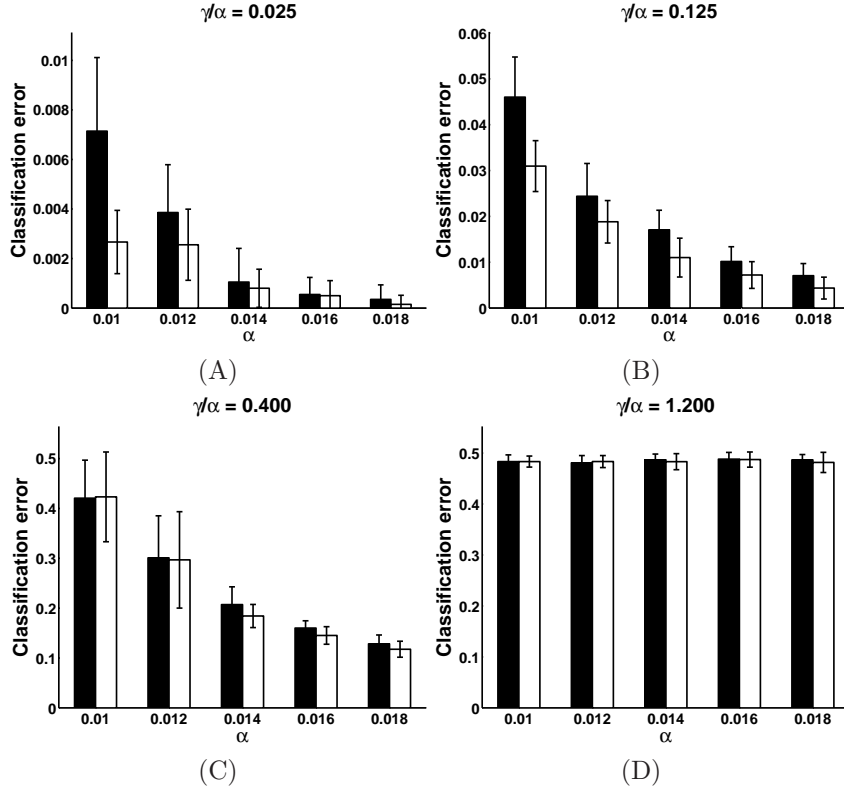
7 Experiments

First, using simulations we demonstrate that normalization improves classification accuracy in the regime surmised by our theoretical results. Next, link prediction

experiments on real world co-authorship networks are presented to show the advantage of normalization. We carry out two simulations; we investigate the behavior of misclassification error with a fixed parameter setting and increasing n and changing parameter settings for a fixed n . For all simulations, a pair of training and test graphs are generated from a Stochastic Blockmodel with a given parameter setting. The model is fitted using Spectral Clustering (with or without normalization) using the training graph whereas misclassification error is computed using the test graph.

7.1 Simulated Networks

Figure 2: For a fixed γ_n/α_n ratio miss-classification error is plotted on the Y axis with increasing α_n on the X axis. (A) $\gamma_n/\alpha_n = 0.025$, (B) $\gamma_n/\alpha_n = 0.125$, (C) $\gamma_n/\alpha_n = 0.4$ and (D) $\gamma_n/\alpha_n = 1.2$.



For a Stochastic Blockmodel with $n = 1000$, $\beta_n = \alpha_n$ and $\pi = 1/2$, we focus on the semi-sparse regime, where expected degree is varied from 10-20. We vary $\alpha_n \in [0.01, 0.018]$ (y axis) and $\gamma_n/\alpha_n \in [0.005, 1.2] \setminus \{1\}$ (x axis). The $\gamma_n/\alpha_n = 1$ case causes instability because it reduces the Stochastic Blockmodel to an Erdős-Rényi graph, and hence is excluded. Since `kmeans` can return a local optimum, we run

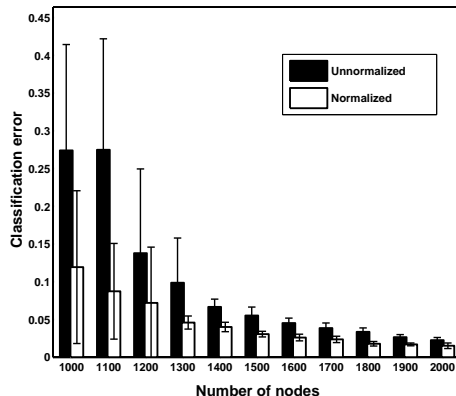
`kmeans` five times and pick the most balanced clustering, in particular the one whose smallest cluster size is largest among the five runs.

For each of the parameter settings average results from twenty random runs are reported with errorbars. In order to ensure that our parameter settings reflect the regime of sparseness required for our theory to hold, we find the connected components of the graph, and only work with those settings where the size of largest connected component is at least 95% of the size of the whole graph. All computations are carried out on the largest connected component. Therefore we never consider the simple case of disconnected clusters. We also assume that $k = 2$ is known.

In each subfigure of Figure 2 we hold γ_n/α_n fixed and plot the classification errors of the two algorithms along the Y axis against increasing α_n values on the X axis. Across the subfigures γ_n/α_n is increased. Our goal is to turn two knobs to adjust the hardness of the classification problem. If one increases α_n for a fixed value of γ_n/α_n , then the problem becomes easier as the expected degree increases with increasing α_n . On the other hand, increasing γ_n/α_n makes it hard to distinguish between clusters.

According to our theoretical results, for small γ_n/α_n ratios, normalization performs better clustering under sparsity. In Figures 2 (A) and (B), we see that normalization always has a smaller average error, although the difference is more striking for small α_n (average degree about 10). As α_n is increased, both methods start to perform equally well. In Figures 2(C) and (D), γ_n/α_n is larger and thus the error rates are also larger. In Figure 2(C) both methods behave similarly, and show improvement with increasing α_n . Finally in Figure 2(D) both misclassify about half of the nodes, since the networks become close to Erdős-Rényi graphs; possibly with more data both methods would perform better.

Figure 3: Miss-classification error with γ_n/α_n on the x axis and increasing n on the y axis.



For the second simulation we fix $\alpha_n = \beta_n = 0.01$, $\gamma_n = 0.002$, $\pi = 0.40$. Now in Figure 3 we plot the errorbars on classification error from twenty random runs along the Y axis and n is varied from 1000 to 2000 in the X axis. One can see that the

normalized method consistently outperforms the unnormalized method; the margin of improvement being more for small n (smaller average degree and hence sparser graphs).

7.2 Real World Networks

Table 2: Table of AUC scores for Real data

Dataset	n	Avg. Degree	AUC scores with Training Links Included			AUC scores with Training Links Excluded		
			Unnorm.	Norm.	Katz	Unnorm.	Norm.	Katz
HepTH	4795	4.6	.67	.82	.87	.59	.79	.79
NIPS	986	4.4	.75	.89	.75	.71	.90	.69
Citeseer	3857	5.6	.79	.96	.97	.65	.93	.90

For real world datasets we use co-authorship networks over T timesteps. The nodes represent authors and edges arise if two authors have co-authored a paper together. Since these networks are unlabeled, we cannot use classification accuracy to measure the quality of Spectral Clustering. Instead, we choose the task of link prediction to quantitatively assess the goodness of clustering. Since the number of clusters is unknown, we learn k via cross validation. We obtain the training graph (A_1) by merging the first $T-2$ datasets, use the $T-1^{th}$ step (A_2) for cross-validating k and use the last timestep (A_3) as the test graph.

We use a subset of the High Energy Physics (HepTH) co-authorship dataset ($T = 6$), the NIPS data ($T = 9$) and the Citeseer data ($T = 11$). Each timestep considers 1-2 years of papers (so that the median degree of the test graph is at least 1). In order to match the degree regime of our theory, we remove all nodes with only one neighbor from the training graph, and work with the largest connected component of the resulting network. Cross validation and testing are done on the corresponding subgraphs of $T - 1$ and T^{th} timesteps respectively. The number of nodes and average degrees are reported in Table 2.

In Section 1 we present the misclassification error on the the political blogs network. This is possible because the entities are labeled as democratic and republican. We preprocess the network as discussed above, and used $k = 2$.

7.2.1 Link Prediction Task

First, we learn the $k \times k$ matrix \hat{P} of within and across class probabilities by counting edges between (or across) two clusters. For testing we pick a hundred nodes at random from nodes with at least one neighbor in the test graph. For node i we construct a prediction vector of length n , whose j^{th} entry is the linkage probability \hat{P}_{ab} learnt using Spectral Clustering; here node i belongs to the a^{th} cluster and node j belongs to the b^{th} cluster. For ground truth we compute the zero one vector representing

presence or absence of an edge between nodes i and j from A_3 . These vectors are concatenated to give one prediction vector and the corresponding ground truth.

Now the AUC score of the prediction vector is computed using the ground truth. This is simply the area under the ROC curve obtained by plotting the false positive rate along the x axis and the true positive rate along the y axis. In order to learn k , we vary k from ten to a hundred. For each value of k we estimate \hat{Z} using Spectral Clustering with k eigenvectors of A_1 (or its normalized counterpart) and then estimate the $k \times k$ conditional probability matrix; now AUC scores are computed using these estimated quantities from A_2 . The k with the largest AUC score is picked and mean AUC scores of five random runs on the test graph using this k is reported.

Since in a co-authorship network, same edges tend to reappear over time, it is often possible to achieve high scores simply by predicting the edges which are already present in the training data. This is why we examine AUC scores from two experiments:

1. Training links included in the test graph.
2. Training links excluded from the test graph.

The second task is harder. We compare our methods with the Katz similarity measure between pairs of nodes ([13]). This measure simply computes a weighted sum of number of paths between two nodes, the weights decreasing exponentially as the length of the path grows. It has been shown to give competitive prediction accuracy for link prediction tasks ([14]). In both panels Normalized performs close to or better than the Katz score, and it outperforms Unnormalized consistently.

8 Summary and Discussion

Normalizing data matrices prior to Spectral Clustering is a common practice. In this paper we propose a theoretical framework to justify this seemingly heuristic choice. With a series of theoretical arguments, we show that for a large parameter regime, in the context of network block models, normalization reduces the variance of points in a given class under the spectral representation. We also present quantifiable classification tasks on simulated networks and link prediction tasks on real networks to demonstrate that normalization indeed leads to better prediction accuracy. In order to develop our theoretical argument, we also derive sharp deviation results on principal empirical eigenvalues of graphs generated from a Stochastic Blockmodel.

For the scope of this paper we consider two class blockmodels. Our proof technique for empirical eigenvalues of A can easily be generalized to more than two classes, provided k is a constant. However for \tilde{A} we construct a guess for the second eigenvector which is also orthogonal to the first. For $k > 2$, one would need to construct k eigenvectors which are orthogonal to each other. This bookkeeping may be quite cumbersome; and hence it is not obvious that the proof technique used for \tilde{A} can be extended to more than two blocks. However we believe that the final result holds for a fixed $k > 2$. In fact, for the real world graphs we learn k by cross validation, and it often exceeds two; our results show that normalization improves link prediction accuracy in those cases as well.

We conclude this paper with a note on some practical disadvantages of normalization. For example, all disconnected components contribute eigenvalue one to the eigen-spectrum of \tilde{A} . Thus, some of the top eigenvectors may be uninteresting. Unnormalized Spectral Clustering does not have this issue. Another example is a small well connected subgraph weakly connected to the rest of the graph. Here, the entries of \tilde{A} corresponding to edges in the subgraph may end up having relatively larger values than the rest of the matrix. Hence the second empirical eigenvector may have high values along this subgraph. If the subgraph is too small, this may again lead to poor clustering.

For the political blogs network, if one does not remove degree-one nodes prior to finding the largest connected component, then misclassification error rate is 50% for normalized Spectral Clustering and 40% for unnormalized Spectral Clustering. On the other hand, removing degree-one nodes drastically improves the error rate of the normalized method to 4%, while not affecting the unnormalized method's performance. We also carried out the link prediction experiments without removing the degree-one nodes; the relative behavior of the different algorithms remained essentially unchanged. Thus, for the normalized method, sparse data artifacts may rank uninteresting eigenvectors high, but the discriminating eigenvectors of \tilde{A} are often more useful than those of A .

9 Acknowledgements

This research was funded in part by NSF FRG Grant DMS-1160319.

A Supplementary Material for Analysis of A

Let A be generated from a $(\alpha_n, \beta_n, \gamma_n, Z)$ blockmodel defined in Definition 2.2. In this section we will prove the lemmas on deviation of principal eigenvalues and eigenvectors presented in Section 5. Theorem 2.1 ([17]) translates to the following:

Corollary A.1. *Define a semi-sparse stochastic blockmodel (see Definition 2.2) with parameters $(\alpha_n, \beta_n, \gamma_n, \pi)$. Assume W.L.O.G that $\pi\alpha_n > (1 - \pi)\beta_n$. Let A be the adjacency matrix of a n node graph generated from this model. Let $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ be the first, second and third largest eigenvalues (in magnitude) of A . Also, let λ_1 and λ_2 be the population versions respectively. For the principal eigenvalues, we have, for $i \in \{1, 2\}$, $\hat{\lambda}_i = \lambda_i(1 + O_P(\sqrt{\log n/n\rho_n}))$. For $\hat{\lambda}_3$ we have:*

$$\hat{\lambda}_3 = O_P(\sqrt{n\rho_n \log n}). \quad (27)$$

Proof. The first result is a straightforward application of Corollary A.1, since d, Δ and λ_i (see Lemma 5.1) are both of the form $C_0 n\rho_n$. Also Assumption 2.1 guarantees the condition on Δ . For the second result, Corollary A.1 gives us $\max_{i \leq n} |\hat{\lambda}_i - \lambda_i| \leq \|A - P\| = O_P(\sqrt{n\rho_n \log n})$. But this is an upper bound on $|\hat{\lambda}_3 - \lambda_3|$, where we have $-\max(\alpha_n, \beta_n) \leq \lambda_3 \leq -\min(\alpha_n, \beta_n)$; which readily gives the result. \square

Now we present our sharp deviation bound on the principal empirical eigenvalues. Let $Q := [\mathbf{v}_1 \ \mathbf{v}_2]$ be the matrix with column vectors as the eigenvectors of the population matrix. Also let $\hat{Q} := [\hat{\mathbf{v}}_1 \ \hat{\mathbf{v}}_2]$ be the empirical version of Q . Let C be a 2×2 matrix with $C = \hat{Q}^T Q$. We write

$$Q = \hat{Q}C + R \quad (28)$$

It is important to note that $\hat{Q}^T R$ is the 2×2 zero matrix. First we prove that the norm of R is small. Let $\hat{\Lambda}$ be the 2×2 diagonal matrix with the top two eigenvalues (in magnitude) of A along its diagonals. Left multiplying Equation (28) by A , and noting that $A\hat{Q} = \hat{Q}\hat{\Lambda}$ we get:

$$(A - P)Q = \hat{Q}\hat{\Lambda}C + AR - PQ = \hat{Q}(\hat{\Lambda}C - C\Lambda) + (AR - R\Lambda) \quad (29)$$

By left multiplying each side of Equation 28 with its transpose we also have

$$C^T C = I_{2 \times 2} - R^T R \quad (30)$$

The above equation only means that C is close to an unitary matrix. In Lemma 5.3 we show that C is indeed close to an identity matrix as well.

First we will state a few lemmas which would be required to prove the concentration results. The proofs are deferred to the appendix.

Lemma A.1. *Define a stochastic blockmodel (see Definition 2.2) with parameters $(\alpha_n, \beta_n, \gamma_n, \pi)$. Assume W.L.O.G that $\pi\alpha_n > (1 - \pi)\beta_n$. We have*

$$\|Q^T(A - P)Q\| = O_P(1) \quad (31)$$

$$\|Q^T(A - P)^2Q\| = O_P(n\rho_n) \quad (32)$$

$$\|(A - P)Q\| = O_P(\sqrt{n\rho_n}) \quad (33)$$

$$\|R\Lambda\| \geq \min(|\Lambda_{11}|, |\Lambda_{22}|)\|R\|/\sqrt{2} \quad (34)$$

$$\|R^T AR\| \leq \|R\|^2 \hat{\lambda}_3 \quad (35)$$

$$\|R^T A^2 R\| \leq \|R\|^2 \hat{\lambda}_3^2 \quad (36)$$

Lemma A.2. *Let C be defined as in Table 1. For a real diagonal 2×2 matrix L , let Γ be the 2×2 matrix of eigenvalues of $C^T LC$. If $L_{ii}, i \in \{1, 2\}$ and $|L_{11} - L_{22}|$ are all of the same order, we have:*

$$\Gamma_{ii} = L_{ii} \left(1 + O_p \left(\frac{1}{n\rho_n} \right) \right) \quad \text{For } i \in \{1, 2\}$$

Lemma A.3. *We have $\|R\| = O_P(\frac{1}{\sqrt{n\rho_n}})$.*

Proof. In Equation 29 we see that the two components are orthogonal in the sense that $\left(\hat{Q}(\hat{\Lambda}C - C\Lambda) \right)^T (AR - R\Lambda)$ is the 2×2 zero matrix. Therefore by pre multiplying each side of Equation 29 with its transpose we have

$$Q^T(A - P)^2Q = \left(\hat{Q}(\hat{\Lambda}C - C\Lambda) \right)^T \left(\hat{Q}(\hat{\Lambda}C - C\Lambda) \right) + (AR - R\Lambda)^T (AR - R\Lambda)$$

Hence via Weyl's identity and positive semi-definiteness of the first component on the R.H.S we have,

$$\|(A - P)Q\|^2 \geq \|AR - R\Lambda\|^2 \quad (37)$$

Since $\|R\Lambda\| \geq \lambda_2\|R\|/\sqrt{2}$, and $\|AR\| \leq \hat{\lambda}_3\|R\|$, using Lemma A.1 (Equation 34) we have $\|AR - R\Lambda\| \geq (\lambda_2/\sqrt{2} - \hat{\lambda}_3)\|R\| = \lambda_2(1 + o_P(1))\|R\|$. From Equation 37 and Lemma A.1 (Equation 34) shows that $\|(A - P)Q\| = O_P(\sqrt{n\rho_n})$ we get the following:

$$\|R\| \leq \frac{\|(A - P)Q\|(1 + o_P(1))}{\lambda_2} = O_P(1/\sqrt{n\rho_n})$$

□

PROOF OF LEMMA 5.2. We have $Q^T A Q = C^T \hat{\Lambda} C + R^T A R$. Lemma A.2 shows that $\|\text{Eig}(C^T \hat{\Lambda} C) - \hat{\Lambda}\| = O_P(1)$. Using Corollary A.1 and Lemma 5.1 we see that $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_1 + \hat{\lambda}_2$ and $\hat{\lambda}_1 - \hat{\lambda}_2$ are of the form $C_0 n \rho_n (1 + O_P(\sqrt{\log n / n \rho_n}))$, thus satisfying the condition of Lemma A.2. Using Lemma A.1 (Equation 35) and Corollary A.1 (Equation 27), we have $\|R^T A R\| \leq \|R\|^2 |\hat{\lambda}_3| = O_P(\sqrt{\log n / n \rho_n})$.

Thus for $i \in \{1, 2\}$ we have:

$$|\text{Eig}_i(Q^T A Q) - \hat{\Lambda}_{ii}| \leq |\text{Eig}_i(C^T \hat{\Lambda} C) - \hat{\Lambda}_{ii}| + \|R^T A R\| = O_P(1).$$

On the other hand, since $Q^T A Q = Q^T P Q + Q^T (A - P) Q = \Lambda + Q^T (A - P) Q$, using Weyl's identity and Lemma A.1 (Equation 31), we also have

$$\text{For } i \in \{1, 2\} \quad |\text{Eig}_i(Q^T A Q) - \Lambda_{ii}| = O_P(1)$$

The above two equations prove that for $i \in \{1, 2\}$, $\hat{\lambda}_i = \lambda_i + O_P(1)$. □

Previously we have shown that C is close to a unitary matrix; now we will prove that when $\gamma_n > 0$, C is actually close to the identity matrix.

PROOF OF LEMMA 5.3. Since $A \mathbf{v}_i = c_{ii} \hat{\lambda}_i \hat{\mathbf{v}}_i + A \mathbf{r}_i = \hat{\lambda}_i (\mathbf{v}_i - \mathbf{r}_i) + A \mathbf{r}_i$,

$$\text{For } i \in \{1, 2\} \quad (A - \hat{\lambda}_i I) \mathbf{r}_i = (A - P) \mathbf{v}_i + (\lambda_i - \hat{\lambda}_i) \mathbf{v}_i \quad (38)$$

Since \mathbf{r}_i is orthogonal to $\hat{\mathbf{v}}_i$, the R.H.S of the above equation is also orthogonal to $\hat{\mathbf{v}}_i$. For definiteness, let $i = 1$. We define the pseudo-inverse of $A - \hat{\lambda}_1 I$ as

$$(A - \hat{\lambda}_1 I)^+ := \sum_{j \neq 1} \frac{\hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^T}{\hat{\lambda}_j - \hat{\lambda}_1}; \quad \|(A - \hat{\lambda}_1 I)^+\| \leq \frac{1}{\min_{j \neq 1} |\hat{\lambda}_j - \hat{\lambda}_1|}.$$

Since $(A - \hat{\lambda}_1 I)^+ (A - \hat{\lambda}_1 I) = I - \hat{\mathbf{v}}_1 \hat{\mathbf{v}}_1^T$, and $\mathbf{r}_1^T \hat{\mathbf{v}}_1 = 0$, left multiplying Equation 38 by $(A - \hat{\lambda}_1 I)^+$ we have:

$$\mathbf{r}_1 = (A - \hat{\lambda}_1 I)^+ \left((A - P) \mathbf{v}_1 + (\lambda_1 - \hat{\lambda}_1) \mathbf{v}_1 \right)$$

Using Lemma A.1 (Equation 33)

$$\|\mathbf{r}_1\| \leq \frac{1}{\min_{j \neq 1} |\hat{\lambda}_j - \hat{\lambda}_1|} O_P(\sqrt{n\rho_n}) \quad (39)$$

For $j > 2$, we have $\min_{j>2} |\hat{\lambda}_1 - \hat{\lambda}_j| \geq |\hat{\lambda}_1| - \max_{j>2} |\hat{\lambda}_j| \geq C_0 n \rho_n (1 + o_P(1))$. Also, $|\hat{\lambda}_1 - \hat{\lambda}_2| = C_0 n \rho_n (1 + o_P(1))$. Hence $1/\min_{j \neq 1} |\hat{\lambda}_j - \hat{\lambda}_1| = O_P(1/n\rho_n)$. This along with Equation 39 gives the required bound on $\|\mathbf{r}_1\|$. An identical argument works for $\|\mathbf{r}_2\|$. Now, $c_{ii}^2 = 1 - \|\mathbf{r}_i\|^2$ and hence we have:

$$1 - c_{ii}^2 = O_P(1/n\rho_n), \text{ for } i \in \{1, 2\}.$$

Let us recall that $c_{21} = \mathbf{v}_2^T \hat{\mathbf{v}}_1$. Now

$$\mathbf{v}_2^T (A - P) \hat{\mathbf{v}}_1 = \hat{\lambda}_1 c_{21} - \lambda_2 c_{21} = (\hat{\lambda}_1 - \lambda_2) c_{21} \quad (40)$$

We also have: $\mathbf{v}_2^T (A - P) \hat{\mathbf{v}}_1 = (\mathbf{v}_2^T (A - P) \mathbf{v}_1 - \mathbf{v}_2^T (A - P) \mathbf{r}_1)/c_{11}$. Now, $\mathbf{v}_2^T (A - P) \mathbf{v}_1$ is the off diagonal term of $Q^T (A - P) Q$, and from Lemma A.1 (Equation 31), we can see that $|\mathbf{v}_2^T (A - P) \mathbf{v}_1| \leq \|Q^T (A - P) Q\| = O_P(1)$. Also, since $\|\mathbf{r}_1\| = O_P(1/\sqrt{n\rho_n})$, and $\|(A - P) \mathbf{v}_2\| = O_P(\sqrt{n\rho_n})$ (Lemma A.1 Equation 33), by the Cauchy-Schwarz inequality we have $|\mathbf{v}_2^T (A - P) \mathbf{r}_1| = O_P(1)$. Since $1 - c_{11}^2 = O_P(1/n\rho_n)$, $|\mathbf{v}_2^T (A - P) \hat{\mathbf{v}}_1| = O_P(1)$. We also have $\hat{\lambda}_1 = \lambda_1 + O_P(1)$, and hence $(\hat{\lambda}_1 - \lambda_2) \sim (\lambda_1 - \lambda_2)$. We can evaluate $\lambda_1 - \lambda_2$ exactly as $C_0 n \rho_n$, and hence the result.

$$c_{21} = \mathbf{v}_2^T (A - P) \hat{\mathbf{v}}_1 / (\hat{\lambda}_1 - \lambda_2) = O_P(1/n\rho_n)$$

A similar argument yields the result for c_{12} . \square

We now know that the norm of the residuals (\mathbf{r}_1 and \mathbf{r}_2) are small. But in order to compute the distances in Equations 2 and 3, we would also need $r_1(C_1)$, $r_2(C_1)$ etc. We now present the proof of Lemma 5.4.

PROOF OF LEMMA 5.4. Note that $\mathbf{r}_1 = c_{12} \hat{\mathbf{v}}_2 + R_1$, and $\mathbf{r}_2 = c_{21} \hat{\mathbf{v}}_1 + R_2$, where each of R_1 and R_2 are orthogonal to both $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$. Hence, $\mathbf{v}_2^T R_1 = R_2^T R_1 = O_P(1/n\rho_n)$ using the Cauchy-Schwarz inequality and Lemma A.3.

Because, \mathbf{r}_1 is orthogonal to $\hat{\mathbf{v}}_1$, $\mathbf{v}_1^T \mathbf{r}_1 = \mathbf{r}_1^T \mathbf{r}_1 = O_P(1/n\rho_n)$. Also, $\mathbf{v}_2^T \mathbf{r}_1 = \mathbf{v}_2^T (c_{12} \hat{\mathbf{v}}_2 + R_1) = c_{12} c_{22} + R_2^T R_1$. Lemma 5.3 shows that $c_{12} = O_P(1/n\rho_n)$, and hence $\mathbf{v}_2^T \mathbf{r}_1$ is $O_P(1/n\rho_n)$.

$$\begin{aligned} n\pi x_1 r_1(C_1) + n(1 - \pi) y_1 r_1(C_2) &= \mathbf{r}_1^T \mathbf{r}_1 =: W_1 \\ n\pi x_2 r_1(C_1) + n(1 - \pi) y_2 r_1(C_2) &= \mathbf{v}_2^T \mathbf{r}_1 =: W_2 \end{aligned}$$

Solving for $r_1(C_1)$ and $r_1(C_2)$ we see that, $r_1(C_1) = \frac{y_2 W_1 - y_1 W_2}{(x_1 y_2 - x_2 y_1) n \pi}$, and $r_1(C_2) = \frac{x_2 W_1 - x_1 W_2}{(x_2 y_1 - x_1 y_2) n (1 - \pi)}$. Using Lemma 5.1 we see that $x_1 y_2 - x_2 y_1 \neq 0$ for $\gamma_n > 0$. In particular, it is of the form C_0/n . Hence we have:

$$r_1(C_1) = O_P(1/n^{3/2} \rho_n) \quad r_1(C_2) = O_P(1/n^{3/2} \rho_n)$$

The result for $r_2(C_1)$ and $r_2(C_2)$ can be obtained analogously. \square

B Supplementary Material for Analysis of \widetilde{A}

PROOF OF LEMMA 6.1. Since the graph defined with adjacency matrix \widetilde{P} is connected for $\alpha_n, \beta_n, \gamma_n > 0$, there is only one principal eigenvalue. The first eigenvalue and vector are obtained by algebra. For the second eigenvector, we construct a vector orthogonal to the first eigenvector and see that it is an eigenvector with eigenvalue $1 - \gamma_n \mu / \mu_1 \mu_2$. This must be the second largest eigenvalue since we have shown that $|\widetilde{\nu}_i| = O(1/n)$ for $i > 2$. It can be shown that $\nu_2 = -1$ iff $\alpha_n = 0, \beta_n = 0, \gamma_n > 0$. The condition $\alpha_n \beta_n \neq \gamma_n^2$ ensures that ν_2 is bounded away from zero by a constant. \square

Let us recall that we approximate $\widetilde{\mathbf{u}}_2$ in two steps. First we construct a guess \mathbf{u}_g^0 orthogonal to $\widetilde{\mathbf{u}}_1$ and then take a power iteration step from it to compute the second guess \mathbf{u}_g . The next lemma computes term by term Taylor expansion of \mathbf{u}_g^0 . This will be used later in computing the error of \mathbf{u}_g .

PROOF OF LEMMA 6.2. We have $(\widetilde{A}\mathbf{u}_g^0)_i = \frac{d_i^{(1)}}{\sqrt{d_i}E_1} - \frac{d_i^{(2)}}{\sqrt{d_i}E_2}$. We recall that for $i \in C_1$, $E[d_i] = n\mu_1$, and for $i \in C_2$, $E[d_i] = n\mu_2$. In the following Taylor series expansions for $i \in C_1$, S_1 and R_i are the remainder terms.

$$\begin{aligned}\frac{1}{E_1} &= \frac{1}{n^2\pi\mu_1} \left(1 - \frac{E_1 - n^2\pi\mu_1}{n^2\pi\mu_1} (1 + S_1) \right) \\ \frac{1}{\sqrt{d_i}} &= \frac{1}{\sqrt{n\mu_1}} \left(1 - \frac{d_i - n\mu_1}{2n\mu_1} (1 + R_i) \right),\end{aligned}$$

Simple applications of the Chernoff bound (Lemmas C.2 and C.4) show that $\max_i |R_i| = o_P(1)$ and $|S| = o_P(1)$. Let $\bar{X} := X - E[X]$. Thus for $i \in C_1$, and $T_i = T_i^{(1)} + T_i^{(2)} + T_i^{(3)}$,

$$\begin{aligned}\frac{d_i^{(1)}}{\sqrt{d_i}E_1} &= \frac{d_i^{(1)}}{\sqrt{n\mu_1}} \frac{1}{n^2\pi\mu_1} \left(1 - \frac{\bar{d}_i}{2n\mu_1} - \underbrace{\frac{\bar{d}_i R_i}{2n\mu_1}}_{T_i^{(1)}} - \underbrace{\frac{\bar{E}_1(1 + S_1)}{n^2\pi\mu_1}}_{T_i^{(2)}} + \underbrace{\frac{\bar{d}_i(1 + R_i)}{2n\mu_1} \frac{\bar{E}_1(1 + S_1)}{n^2\pi\mu_1}}_{T_i^{(3)}} \right) \\ &= \frac{(n\pi - 1)\alpha_n}{\sqrt{n\mu_1}} \frac{1}{n^2\pi\mu_1} \left(1 + \underbrace{\frac{\bar{d}_i^{(1)}}{(n\pi - 1)\alpha_n}}_{G_i} - \underbrace{\frac{\bar{d}_i}{2n\mu_1}}_{H_i} + T_i' \right) \\ T_i' &= T_i + T_i G_i - G_i H_i\end{aligned}$$

We note that $G_i - H_i$ is of the form $c_1 \frac{\bar{d}_i^{(1)}}{n\rho_n} - c_2 \frac{\bar{d}_i^{(2)}}{n\rho_n}$ for constants c_1, c_2 , and using an argument similar to Lemma 5.5 (Equation 18) we see that $\|G - H\|$ as well as $\|G\|$ and $\|H\|$ are of the form $C_0/\sqrt{\rho_n}(1 + o_P(1))$. We will show that $\|T'\|$ is $o_P(1/\sqrt{\rho_n})$.

First we note that using applications of Chernoff bound (details in Lemmas C.2, C.4), we have:

$$\begin{aligned} \max_i R_i &= o_P(1); & \|T^{(1)}\| &= \|H\|_{o_P(1)}; \\ \|T^{(2)}\| &= O_P\left(\frac{1}{n}\right); & \|T^{(3)}\| &= \|H\|_{o_P(1)} \end{aligned} \quad (41)$$

Since, $\|H\|$ and $\|G - H\|$ are of the same order, we have $\|T\| = \|G - H\|_{o_P(1)}$. For the $G_i H_i$ term, since $\bar{d}_i = d_i^{(1)} + d_i^{(2)}$, this can be written as:

$$G_i H_i = \frac{(\bar{d}_i^{(1)})^2}{2n\mu_1(n\pi - 1)\alpha_n} + \frac{\bar{d}_i^{(1)}\bar{d}_i^{(2)}}{2n\mu_1(n\pi - 1)\alpha_n}.$$

Using Lemma C.5 we see that the vectors formed by the two summands on the R.H.S of the above equation are of norm $O_P(1/n\rho_n^2)$. Hence using the Cauchy-Schwarz inequality one can bound the norm of $\langle G_i H_i \rangle$ by $O_P(1/\sqrt{n\rho_n^2})$, which is again $\|G - H\|_{o_P(1)}$. Finally for Y_i , we have

$$T_i G_i = G_i H_i R_i + G_i T_i^{(2)} + G_i H_i (1 + R_i) T_i^{(2)}.$$

Using Equation 41 and the bound on $\langle G_i H_i \rangle$ we have $\|Y\| = o_P(G - H)$.

Similarly we can show that:

$$\begin{aligned} \frac{d_i^{(2)}}{\sqrt{d_i} E_2} &= \frac{\gamma_n}{\sqrt{n\mu_1}} \frac{1}{n\mu_2} \left(1 - \frac{\bar{d}_i}{2n\mu_1} + \frac{\bar{d}_i^{(2)}}{n(1-\pi)\gamma_n} + T_i'' \right) \\ \|T_i''\| &= o_P \left\| \left\langle -\bar{d}_i/2n\mu_1 + \bar{d}_i^{(2)}/n(1-\pi)\gamma_n \right\rangle \right\| = o_P(1/\sqrt{\rho_n}) \end{aligned}$$

Thus, for the i^{th} entry of $\tilde{A}\mathbf{u}_g^0$ we have:

$$[\tilde{A}\mathbf{u}_g^0]_i = \begin{cases} \frac{\nu_2}{n\pi\sqrt{n\mu_1}} \left(1 - \frac{\bar{d}_i}{2n\mu_1} + \frac{\bar{d}_i^{(1)}}{n\mu_1\nu_2} - \frac{\bar{d}_i^{(2)}}{n\mu_2\nu_2} \frac{\pi}{1-\pi} + M_i \right) & i \in C_1 \\ -\frac{\nu_2}{n(1-\pi)\sqrt{n\mu_2}} \left(1 - \frac{\bar{d}_i}{2n\mu_2} - \frac{\bar{d}_i^{(1)}}{n\mu_1\nu_2} \frac{1-\pi}{\pi} + \frac{\bar{d}_i^{(2)}}{n\mu_2\nu_2} + M'_i \right) & i \in C_2 \end{cases}$$

$$\|M\|, \|M'\| = o_P(1/\sqrt{\rho_n})$$

$$\|\tilde{A}\mathbf{u}_g^0\| \sim \nu_2 \sqrt{\frac{\mu}{n^2\pi(1-\pi)\mu_1\mu_2}}$$

□

PROOF OF LEMMA 6.3. We have:

$$\sqrt{d_i} = \sqrt{n\mu_1} + \frac{\bar{d}_i(1 + S_i)}{2\sqrt{n\mu_1}}; \quad \frac{1}{E_1} = \frac{1}{n^2\pi\mu_1} \left(1 - \frac{E_1 - n^2\pi\mu_1}{n^2\pi\mu_1} (1 + S_1) \right),$$

Using arguments similar to Lemma 4.3 one can show that $\max_i |S_i|$ and $|S_1| = o_P(1)$. This gives the following entrywise Taylor expansion for \mathbf{u}_g^0 .

$$u_g^0(i) = \begin{cases} \frac{1}{n\pi\sqrt{n\mu_1}} \left(1 + \frac{\bar{d}_i}{2n\mu_1} + T_i\right) & i \in C_1 \\ \frac{-1}{n(1-\pi)\sqrt{n\mu_2}} \left(1 + \frac{\bar{d}_i}{2n\mu_2} + T'_i\right) & i \in C_2 \end{cases}$$

$$\|T\|, \|T'\| = o_P(1/\sqrt{\rho_n}) \quad (42)$$

It immediately follows that

$$\|\mathbf{u}_g^0\|^2 \sim \frac{\mu}{n^2\pi(1-\pi)\mu_1\mu_2}. \quad (43)$$

Now we represent \mathbf{u}_g^0 as

$$\mathbf{u}_g^0/\|\mathbf{u}_g^0\| = c_g^0 \tilde{\mathbf{u}}_2 + \mathbf{r}_g^0 \quad (44)$$

where $c_g^0 = \tilde{\mathbf{u}}_2^T \mathbf{u}_g^0$. Since \mathbf{u}_g^0 is orthogonal to $\tilde{\mathbf{u}}_1$ (by construction), and $\tilde{\mathbf{u}}_2$ is orthogonal to $\tilde{\mathbf{u}}_1$ by definition, \mathbf{r}_g^0 is orthogonal to both $\tilde{\mathbf{u}}_1$ and $\tilde{\mathbf{u}}_2$.

$$\frac{\tilde{A}\mathbf{u}_g^0 - \nu_2\mathbf{u}_g^0}{\|\mathbf{u}_g^0\|} = c_g^0 \tilde{\mathbf{u}}_2(\tilde{\nu}_2 - \nu_2) + (\tilde{A} - \nu_2 I)\mathbf{r}_g^0$$

Now taking the squared norm of both sides yields,

$$\frac{\|\tilde{A}\mathbf{u}_g^0 - \nu_2\mathbf{u}_g^0\|^2}{\|\mathbf{u}_g^0\|^2} \geq \|(\tilde{A} - \nu_2 I)\mathbf{r}_g^0\|^2 \geq (\nu_2 - \tilde{\nu}_3)^2 \|\mathbf{r}_g^0\|^2 \quad (45)$$

since $\|(\tilde{A} - \nu_2 I)\mathbf{r}_g^0\| \geq \|\nu_2\mathbf{r}_g^0\| - \|\tilde{A}\mathbf{r}_g^0\| \geq (\nu_2 - \tilde{\nu}_3)\|\mathbf{r}_g^0\|$. With $\tilde{\nu}_3 = \max_{i>2} |\tilde{\nu}_i|$, we have $\|\tilde{A}\mathbf{r}_g^0\| \leq \tilde{\nu}_3\|\mathbf{r}_g^0\|$, since \mathbf{r}_g^0 is orthogonal to both the first and second eigenvectors of \tilde{A} .

In order to compute $\|\tilde{A}\mathbf{u}_g^0 - \nu_2\mathbf{u}_g^0\|^2$, we note that the leading term vanishes, and the first order term dominates the remainder term. Now using Lemma 6.2 and Equation 18, for some deterministic constants $\tau_{k,1}$ and $\tau_{k,2}$ ($k \in \{1, 2\}$) (which are non-zero for $\gamma_n > 0$), we have:

$$\|\tilde{A}\mathbf{u}_g^0 - \nu_2\mathbf{u}_g^0\|^2 \sim \frac{\nu_2^2}{n^3\rho_n} \sum_{k=1}^2 \sum_{i \in C_k} \left(\tau_{k,1} \frac{\bar{d}_i^{(1)}}{n\mu_1} + \tau_{k,2} \frac{\bar{d}_i^{(2)}}{n\mu_2} \right)^2 = O_P\left(\frac{1}{n^3\rho_n^2}\right) \quad (46)$$

Using Theorem 2.1 and Lemma 6.1 we know that $\tilde{\nu}_3 = O_P(\sqrt{\log n/n\rho_n})$ and hence $|\nu_2 - \tilde{\nu}_3| \sim \nu_2$, where ν_2 is a deterministic constant independent of n . Plugging these into equation 45, and using Equation 43 we have:

$$(\nu_2 - \tilde{\nu}_3)^2 \|\mathbf{r}_g^0\|^2 = O_P\left(\frac{1}{n\rho_n}\right) \rightarrow \|\mathbf{r}_g^0\|^2 = O_P\left(\frac{1}{\sqrt{n\rho_n}}\right)$$

From equation 44 we have the following ortho-normal decomposition of \mathbf{u}_g :

$$\mathbf{u}_g = \frac{c_g^0 \tilde{\nu}_2}{\|\tilde{A}\mathbf{u}_g^0\|/\|\mathbf{u}_g^0\|} \tilde{\mathbf{u}}_2 + \frac{\tilde{A}\mathbf{r}_g^0}{\|\tilde{A}\mathbf{u}_g^0\|/\|\mathbf{u}_g^0\|} = c_g \tilde{\mathbf{u}}_2 + \mathbf{r}_g$$

where $c_g := (\tilde{\mathbf{u}}_2)^T \mathbf{u}_g$ and $\mathbf{r}_g := \frac{\tilde{A}\mathbf{r}_g^0}{\|\tilde{A}\mathbf{u}_g^0\|/\|\mathbf{u}_g^0\|}$. From Lemma 6.2 we have $\|A\mathbf{u}_g^0\| \sim \|\nu_2 \mathbf{u}_g^0\|$. This shows that the residual \mathbf{r}_g has norm of the following order:

$$\|\mathbf{r}_g\| = \frac{\|\tilde{A}\mathbf{r}_g^0\|}{\|\tilde{A}\mathbf{u}_g^0\|/\|\mathbf{u}_g^0\|} = O_P \left(\frac{\tilde{\nu}_3 \|\mathbf{r}_g^0\|}{\|\tilde{A}\mathbf{u}_g^0\|/\|\mathbf{u}_g^0\|} \right) = O_P \left(\sqrt{\frac{\log n}{n^2 \rho_n^2}} \right)$$

But $c_g^2 = 1 - \|\mathbf{r}_g\|^2 = 1 - o_P(1)$. □

C Proofs of Ancillary Results

PROOF OF COROLLARY 2.2. Theorem 2.1 shows that indeed $\|A - P\| \leq 4\sqrt{\Delta \log(n/\delta)}$ for A generated from a semi-sparse Stochastic Blockmodel. In fact, for $\gamma_n > 0$, and $\alpha_n \beta_n \neq \gamma_n^2$, Lemma 5.1 shows that $|\lambda_1 - \lambda_2| = C_0 n \rho_n$. Let $x := |\lambda_1 - \lambda_2|/4$, $a_i := \lambda_i - 2x$, and $b_i := \lambda_i + 2x$, for $i \in \{1, 2\}$; there are no eigenvalues in $(a_i - x, a_i + x) \cup (b_i - x, b_i + x)$. Now, an application of Corollary 2.2 proves the result for eigenvectors of A .

In spirit of the analysis of misclassification error in [19], we construct a matrix Ψ such that the i^{th} row is the center of the cluster assigned to node i , also denoted by Ψ_i . Let O denote a 2×2 diagonal matrix with $O_{ii} = 1/\hat{\mathbf{v}}_i^T \mathbf{v}_i$. Then a node i is correctly classified if for some $j \neq i$, $\|\Psi_i O - Q_j\| > \|\Psi_i O - Q_i\|$.

We have $\|Q_i - Q_j\| = \frac{1}{\sqrt{n\pi(1-\pi)}} := s$. We will first show that $\|\Psi_i O - Q_i\| < s/2$ is a sufficient condition for correct classification. So for i and j belonging to different clusters, if $\|\Psi_i O - Q_i\| < s/2$ we have:

$$\|\Psi_i O - Q_j\| \geq \|Q_i - Q_j\| - \|\Psi_i O - Q_i\| > s/2 > \|\Psi_i O - Q_i\|,$$

which implies that i is correctly classified. Now let $\mathcal{M} := \{i : \|\Psi_i O - Q_i\| \geq s/2\}$. The set of misclassified nodes is a subset of this set. We have:

$$|\mathcal{M}| = \sum_{i \in \mathcal{M}} 1 \leq \frac{1}{4s^2} \sum_{i \in \mathcal{M}} \|\Psi_i O - Q_i\|^2 \leq n\pi(1-\pi) \frac{\|\Psi - QO^{-1}\|_F^2 \|O\|_F^2}{4}$$

The **kmeans** algorithm minimizes $\|\Psi - \hat{Q}\|_F$ over all $n \times k$ dimensional real matrices Ψ with at most two unique rows, we have $\|\Psi - \hat{Q}\|_F \leq \|\hat{Q} - QO^{-1}\|_F$, which also gives $\|\Psi - QO^{-1}\|_F \leq 2\|\hat{Q} - QO^{-1}\|_F$. However Corollary 2.2 shows that, the operator norm $\|\mathbf{v}_i \mathbf{v}_i^T - \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T\| = O_P(\sqrt{\log n / n \rho_n})$. Now, $\|\mathbf{v}_i - (\mathbf{v}_i^T \hat{\mathbf{v}}_i) \hat{\mathbf{v}}_i\| = \|(\mathbf{v}_i \mathbf{v}_i^T - \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T) \mathbf{v}_i\| = O_P(\sqrt{\log n / n \rho_n})$ for $i \in \{1, 2\}$; hence $\|\hat{Q} - QO^{-1}\|_F = O_P(\sqrt{\log n / n \rho_n})$. Moreover, $|1 - (\mathbf{v}_i^T \hat{\mathbf{v}}_i)^2| = |\mathbf{v}_i^T (\mathbf{v}_i - (\mathbf{v}_i^T \hat{\mathbf{v}}_i) \hat{\mathbf{v}}_i)| = O_P(\sqrt{\log n / n \rho_n})$, giving $\|O\|_F = 1 + o_P(1)$.

Hence we have, $|\mathcal{M}|/n = O_P(\log n/n\rho_n)$, which upper bounds the fraction of misclassified nodes. In order to analyze misclassification error of \tilde{A} , one can use Lemma 6.1 and Corollary 2.2 to first show that the eigenvectors of \tilde{A} converge in the appropriate sense. Then, an identical argument yields the same rate for \tilde{A} . \square

PROOF OF LEMMA 4.2. We have, $x_n(i) - \sum_i x_n(i)/n = y_n(i) + z_n(i)$, where \mathbf{y}_n and \mathbf{z}_n are the centered \mathbf{x}_n^1 and \mathbf{R}_n vectors respectively. Now, $\|\mathbf{z}_n\|^2 \leq \|\mathbf{R}_n\|^2 = o_P(\|\mathbf{x}_n^1\|^2)$ and $\|\mathbf{y}_n\|^2 = \|\mathbf{x}_n^1\|^2(1 + o_P(1))$ by assumption $|\sum_i x_n^1(i)/n| = o_P(\|\mathbf{x}_n^1\|/\sqrt{n})$. In fact, this assumption is stronger than one obtained by applying the Cauchy-Schwarz inequality; it implies that the average is actually of a smaller order. Hence, $\|\mathbf{z}_n\| = o_P(\|\mathbf{y}_n\|)$ and an application of the Cauchy-Schwarz inequality yields the result. \square

PROOF OF LEMMA 4.3. We will use the fact that $\|\tilde{\mathbf{r}}\|^2 = 1 - \tilde{c}_{11}^2 = \sum_i (\tilde{\mathbf{u}}_i - \sum_i \tilde{\mathbf{u}}_i/n)^2$. Since one can explicitly obtain the expression of u_i , the basic idea is to use term by term Taylor approximation to obtain the norm. However, the issue is that we are summing over n elements where n is going to infinity, and extra care is required for the remainder terms; in particular, we will bound them uniformly over n .

It is easy to check that the vector $\langle \sqrt{d_i/E} \rangle$ is an eigenvector of \tilde{A} with eigenvalue one. Assumption 2.1 guarantees that A is connected with high probability, and so the principal eigenvalue has multiplicity one. Thus $\tilde{\mathbf{u}}(i) = \sqrt{d_i/E}$. We have

$$\sqrt{x} = \sqrt{x_0} + \frac{x - x_0}{2\sqrt{x_0}}(1 + S) \quad \frac{1}{\sqrt{y}} = \frac{1}{\sqrt{y_0}} - \frac{y - y_0}{2y_0^{3/2}}(1 + S'). \quad (47)$$

where $S = \sqrt{x_0} \int_0^1 ((x_0 + \tau(x - x_0))^{-1/2} - x_0^{-1/2}) d\tau$. For $\tau \in [0, 1]$, $|(x_0 + \tau(x - x_0))^{-1/2} - x_0^{-1/2}| \leq |x^{-1/2} - x_0^{-1/2}|$, and hence $|S| \leq |(x_0/x)^{1/2} - 1|$. Similarly $|S'| \leq |(E_0/E)^{3/2} - 1| = O_P(\sqrt{\log n/n^2\rho_n})$ applying the Chernoff bound (Lemma C.3). Setting $x = d_i, x_0 = d_0$, where $d_0 := (n-1)p$, $y = E, y_0 = E_0$ where $E_0 := n(n-1)p$ we get:

$$\begin{aligned} \tilde{u}_i &= \frac{\sqrt{d_0}}{\sqrt{E_0}} + \frac{\bar{d}_i}{2\sqrt{d_0 E_0}} + R \\ R &= \frac{\bar{d}_i}{2\sqrt{d_0 E_0}} S_i - \frac{E - E_0}{2} \sqrt{\frac{d_0}{E_0^3}} (1 + S') - \frac{E - E_0}{2\sqrt{d_0 E_0^3}} \bar{d}_i (1 + S_i) (1 + S') \\ |S'| &= o_P(1), \quad \max_i |S_i| = o_P(1) \end{aligned}$$

We will now invoke Lemma 4.2. Let c_n be the vector of constants $1/\sqrt{n}$, and $x_n^1 := \frac{\bar{d}_i}{2\sqrt{d_0 E_0}}$. Hence $\|x_n^1\| \sim C_0/\sqrt{n\rho_n}$, and the mean of x_n^1 is $O_P(1/\sqrt{n^3\rho_n}) = o_P(\|x_n^1/n\|/\sqrt{n})$. Now an application of Chernoff bound (details in lemmas C.1, C.3 and C.4) we can show that $\|R\| = o_P(\|1/\sqrt{n\rho_n}\|)$. \square

Lemma C.1. Let $d_o := (n-1)p$, $\max_i \left| \left(\frac{d_i}{d_o} \right)^{-1/2} - 1 \right| = O_P(\sqrt{\frac{\log n}{np}})$

Proof. Since $(d_i - d_0)^{-1/2}$ is well defined w.h.p and Lipschitz continuous, one can show that for $\epsilon := \sqrt{\frac{6 \log n}{np}}$, for large enough n , $P(U_i^{-1/2} \in [1 - \epsilon, 1 + \epsilon]) \geq 1 - 1/n^2$. Now a simple union bound yields the result. \square

The following three lemmas can be proven in an analogous way.

Lemma C.2. $\max_i \left| \left(\frac{d_i}{(n-1)p} \right)^{-3/2} - 1 \right| = O_P(\sqrt{\frac{\log n}{np}})$

Lemma C.3. Let $E := \sum_j d_j$, $E_0 := n(n-1)p$. $\left| \left(\frac{E}{E_0} \right)^{-3/2} - 1 \right| = O_P(\sqrt{\frac{\log n}{n^2 \rho_n}})$

Lemma C.4. We have $\frac{1}{E} \sim \frac{1}{n(n-1)p}$.

Lemma C.5. Define $\bar{d}_i := d_i - (n-1)p$, we have $\sum_i \bar{d}_i^4 = O_P(n^3 \rho_n^2)$.

Proof. First we compute $\sum_i E[\bar{d}_i^4]$. Let $\bar{A}_{ij} := A_{ij} - p$, where A is the adjacency matrix.

$$E[\bar{d}_i^4] = \sum_{j_1, j_2, j_3, j_4} E[\bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{ij_3} \bar{A}_{ij_4}] = O(n^2 \rho_n^2)$$

Since $E[\bar{A}_{ij}] = 0$, and \bar{A}_{ij} and \bar{A}_{ik} are independent for $j \neq k$, all terms with exactly one occurrence of \bar{A}_{ij} will contribute zero. Thus the above is obtained by setting $j_1 = j_2$ and $j_3 = j_4 \neq j_1$ and considering different permutations to achieve similar settings. We have:

$$\text{var}(\bar{d}_i^4) \leq E[\bar{d}_i^8] = \sum_{j_1, j_2, j_3, j_4, \ell_1, \ell_2, \ell_3, \ell_4} E[\bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{ij_3} \bar{A}_{ij_4} \bar{A}_{i\ell_1} \bar{A}_{i\ell_2} \bar{A}_{i\ell_3} \bar{A}_{i\ell_4}] = O(n \rho_n)^4.$$

As for $\text{cov}[\bar{d}_i^4, \bar{d}_k^4] = E[\bar{d}_i^4 \bar{d}_k^4] - E[\bar{d}_i^4]E[\bar{d}_k^4]$, we will first compute $E[\bar{d}_i^4 \bar{d}_k^4]$ which upper bounds the covariance.

$$\sum_{j_1, j_2, j_3, j_4, \ell_1, \ell_2, \ell_3, \ell_4} E[\bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{ij_3} \bar{A}_{ij_4} \bar{A}_{k\ell_1} \bar{A}_{k\ell_2} \bar{A}_{k\ell_3} \bar{A}_{k\ell_4}] = O(n \rho_n)^4$$

Thus $\text{var}(\sum_i \bar{d}_i^4) = O(n^6 \rho_n^4)$ and $E[\sum_i \bar{d}_i^4] = O(n^3 \rho_n^2)$ leading to the result. \square

PROOF OF LEMMA A.1. In order to prove this lemma, we will first state a simple result which will be used heavily. Let B be a symmetric 2×2 matrix with $B_{11} = a$, $B_{12} = c$ and $B_{22} = b$. The two eigenvalues of B are given by $\frac{a+b+\sqrt{(a-b)^2+4c^2}}{2}$ and $\frac{a+b-\sqrt{(a-b)^2+4c^2}}{2}$. If $a, b \geq 0$, we have:

$$\|B\| = \frac{a + b + \sqrt{(a-b)^2 + 4c^2}}{2}. \quad (48)$$

Now we present the proofs of Equations 31 to 36.

PROOF OF EQUATION 31: . Let $B := Q^T(A - P)Q$. We see that the entries of B are of the form

$$a \sum_{i \in C_1} (d_i^{(1)} - n\pi\alpha_n) + b \sum_{i \in C_2} (d_i^{(2)} - n(1 - \pi)\beta_n) + c \sum_{i \in C_1} (d_i^{(2)} - n(1 - \pi)\gamma_n).$$

For B_{11} , $a = x_1^2$, $b = y_1^2$ and $c = 2x_1y_1$; for B_{22} , $a = x_2^2$, $b = y_2^2$ and $c = 2x_2y_2$ and for B_{12} $a = x_1x_2$, $b = y_1y_2$, and $c = (x_2y_1 + x_1y_2)$. Now $\sum_{i \in C_1} (d_i^{(1)} - n\pi\alpha_n)$ is the sum of $\binom{n\pi}{2}$ centered Bernoulli random variables, and hence is $O_P(n\sqrt{\rho_n})$. One can show that $\sum_{i \in C_2} (d_i^{(2)} - n(1 - \pi)\beta_n)$ and $\sum_{i \in C_1} (d_i^{(2)} - n(1 - \pi)\gamma_n)$ are also $O_P(n\sqrt{\rho_n})$. From Lemma 5.1 x_1, x_2, y_1 and y_2 are all of the form C_0/\sqrt{n} . Hence B_{11}, B_{12} and B_{22} are all $O_P(1)$. Hence $\|B\| \leq \sqrt{2}B_\infty = O_P(1)$.

PROOF OF EQUATION 32: . Let $B := Q^T(A - P)^2Q$; we have:

$$B_{11} = \|(A - P)v_1\|^2 \quad B_{22} = \|(A - P)v_2\|^2 \quad B_{12} = B_{21} \leq \sqrt{B_{11}B_{22}}$$

Hence, Equation 48 gives $\|B\| \leq B_{11} + B_{22}$, where we have:

$$\begin{aligned} B_{11} &= \sum_{i \in C_1} (x_1(d_i^{(1)} - n\pi\alpha_n) + y_1(d_i^{(2)} - n(1 - \pi)\gamma_n))^2 + \sum_{i \in C_2} (x_1(d_i^{(1)} - n\pi\gamma_n) + y_1(d_i^{(2)} - n(1 - \pi)\beta_n))^2 \\ B_{22} &= \sum_{i \in C_1} (x_2(d_i^{(1)} - n\pi\alpha_n) + y_2(d_i^{(2)} - n(1 - \pi)\gamma_n))^2 + \sum_{i \in C_2} (x_2(d_i^{(1)} - n\pi\gamma_n) + y_2(d_i^{(2)} - n(1 - \pi)\beta_n))^2 \end{aligned}$$

We note that we have $E[\sum_{i \in C_1} (d_i^{(1)} - n\pi\alpha_n)^2] = (n\pi)^2\alpha_n(1 - \alpha_n)$, and $\text{var}(\sum_{i \in C_1} (d_i^{(1)} - n\pi\alpha_n)^2) = O((n\pi)^3\alpha_n^2(1 - \alpha_n)^2)$ [8]. But the coefficients x_1, x_2, y_1 and y_2 are all of the form C_0/\sqrt{n} ; now the Cauchy-Schwarz inequality gives: B_{11} and B_{22} are $O_P(n\rho_n)$, and hence Equation (32) is proven.

PROOF OF EQUATION 33. Note that $\|(A - P)Q\| = \sqrt{\|Q^T(A - P)^2Q\|}$. This along with Equation 32 gives Equation 33.

PROOF OF EQUATION 34. $\|R\Lambda\| = \sqrt{\|\Lambda R^T R \Lambda\|}$. Let $B := \Lambda R^T R \Lambda$. Let λ_1 and λ_2 be the diagonal elements of Λ . Also WLOG let $|\lambda_1| \geq |\lambda_2|$. If the columns of R are denoted by r_1 and r_2 , then we have:

$$B_{11} = \lambda_1^2 r_1^T r_1 \quad B_{12} = \lambda_1 \lambda_2 r_1^T r_2 \quad B_{22} = \lambda_2^2 r_2^T r_2$$

From Equation 48 we have, $\|B\| \geq (B_{11} + B_{22})/2 \geq \lambda_2^2(r_1^T r_1 + r_2^T r_2)/2$. Also, using the Cauchy-Schwarz inequality and Equation 48 we have: $\|R^T R\| \leq (r_1^T r_1 + r_2^T r_2)$. Hence $\|R\Lambda\| \geq \sqrt{\lambda_2^2 \|R^T R\|}/2$, which gives us the result in Equation 34.

PROOF OF EQUATION 35. Let the eigenvalues of A be ordered as $|\hat{\lambda}_1| \geq |\hat{\lambda}_2| \geq \dots \geq |\hat{\lambda}_n|$. The adjacency matrix A can be written as $\widehat{Q}\widehat{\Lambda}\widehat{Q}^T + \widehat{Q}'\widehat{\Lambda}'\widehat{Q}'^T$, where \widehat{Q}' is

the $n \times n - 2$ matrix with columns as eigenvectors corresponding to the eigenvalues $i > 2$. $\widehat{\Lambda}'$ is the $n - 2$ by $n - 2$ diagonal matrices with eigenvalues $\widehat{\lambda}_3, \dots, \widehat{\lambda}_n$ along the diagonal. We have $\|\widehat{\Lambda}'\| = \|\lambda_3\|$. Hence, $\|R^T A R\| = \|R^T \widehat{Q}' \widehat{\Lambda}' \widehat{Q}'^T R\| \leq \|R\|^2 |\widehat{\lambda}_3|$.

PROOF OF EQUATION 36. This is a simple consequence of the proof of Equation 35. \square

Lemma C.6. Let C be defined as in section A. $\|CC^T - I_{2 \times 2}\| = O_p(1/n\rho_n)$.

Proof. Equation 30 gives $C^T C = I_{2 \times 2} - R^T R$. The eigenvalues of CC^T and $C^T C$ are the same; let us denote them by c_1 and c_2 ($|c_1| \geq |c_2|$). Since $\|R^T R\| = O_p(1/n\rho_n)$ (Lemma A.3), Weyl's inequality yields $c_i \leq 1 \leq c_i + \|R^T R\|$, i.e. $c_i \in (1 - O_p(1/n(1 - \pi)\beta_n), 1)$, thus proving the result. \square

PROOF OF LEMMA A.2. Let $\det(\cdot)$ be the determinant of the square matrix (\cdot) , and let $\text{trace}(\cdot)$ denote the trace of the square matrix (\cdot) . The product of the eigenvalues of $C^T L C$ is given by $\Gamma_{11}\Gamma_{22} = \det(C^T L C)$. The sum of the eigenvalues is given by $\Gamma_{11} + \Gamma_{22} = \text{trace}(C^T L C)$. Let $M := CC^T - I_{2 \times 2}$. From Lemma C.6 we have $\|M\| = O_p(1/n\rho_n)$. Every element of M is also $O_p(1/n\rho_n)$ (using $|M_{ij}| \leq \|M\|_F \leq \sqrt{2}\|M\|$). Hence for the sum we have: $\Gamma_{11} + \Gamma_{22} = \text{trace}(C^T L C) = \text{trace}(CC^T L) = \text{trace}(L + ML) = \text{trace}(L) + R$, where $|R| = \|L\|O_p(1/n\rho_n)$. Also, for the product,

$$\begin{aligned} \Gamma_{11}\Gamma_{22} &= \det(C^T L C) = \det(C^T C) \det(L) = \det(I_{2 \times 2} - R^T R) \det(L) \\ &= ((1 - r_1^T r_1)(1 - r_2^T r_2) - (r_1^T r_2)^2) \det(L) = \det(L) (1 + S), \end{aligned}$$

where S is $O_p(1/n\rho_n)$. Now, Γ_{ii} for $i \in \{1, 2\}$ are solutions to the quadratic equation $x^2 - (L_{11} + L_{22} + R)x + L_{11}L_{22}(1 + S)$. The solutions are given by

$$\begin{aligned} x &= \frac{L_{11} + L_{22} + R \pm \sqrt{(L_{11} + L_{22} + R)^2 - 4L_{11}L_{22}(1 + S)}}{2} \\ &= \frac{L_{11} + L_{22} + R \pm \sqrt{(L_{11} - L_{22})^2 + R'}}{2} \quad \text{where } |R'| = \|L\|^2 O_p(1/n\rho_n) \end{aligned}$$

If $L_{11} - L_{22}$ is the same order as $L_{ii}, i \in \{1, 2\}$, then a little algebra yields the result. \square

PROOF OF LEMMA 5.5. First we prove Equation 17.

PROOF OF EQUATION 17. We see that $\text{var}[\bar{d}_i^{(1)}] = (n\pi - 1)\alpha_n(1 - \alpha_n)$. Also $\text{var}((\bar{d}_i^{(1)})^2) = E[(\bar{d}_i^{(1)})^4] - E[(\bar{d}_i^{(1)})^2]^2$. Now $E[(\bar{d}_i^{(1)})^4] = \sum_{j_1, j_2, j_3, j_4} \bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{ij_3} \bar{A}_{ij_4}$, where $j_1, j_2, j_3, j_4 \in C_1$. Because of pairwise independence we have,

$$\begin{aligned} E\left[\sum_{j_1, j_2, j_3, j_4} \bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{ij_3} \bar{A}_{ij_4}\right] &= C_0 \sum_{j_1 \neq j_2} E[\bar{A}_{ij_1}^2] E[\bar{A}_{ij_2}^2] + \sum_j E[\bar{A}_{ij}^4] \\ &= C_0 \binom{n\pi}{2} \alpha_n^2 (1 - \alpha_n)^2 + n\pi E[\bar{A}_{ij}^4] = O_p(n^2 \rho_n^2) \end{aligned}$$

As for the covariance terms we have for $i \neq k \in C_1$,

$$\begin{aligned} E[(\bar{d}_i^{(1)})^2 (d_k^{(1)})^2] &= E\left[\sum_{j_1, j_2, \ell_1, \ell_2} \bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{k\ell_1} \bar{A}_{k\ell_2}\right] = \sum_{j_1, \ell_1} E[\bar{A}_{ij_1}^2] E[\bar{A}_{k\ell_1}^2] + E[\bar{A}_{ik}^4] \\ &= \left(\sum_{j_1 \in C_1} E[\bar{A}_{ij_1}^2]\right)^2 + E[\bar{A}_{ik}^4] \end{aligned}$$

Hence $\text{cov}((\bar{d}_i^{(1)})^2, (d_k^{(1)})^2) = E[\bar{A}_{ik}^4]$. Thus $\text{var}(\sum_{i \in C_1} \bar{d}_i^2) = O_P(n^3 \rho_n^2)$, and $\sum_{i \in C_1} (\bar{d}_i^{(1)})^2 \sim \sum_{i \in C_1} E[(\bar{d}_i^{(1)})^2]$, proving the first result.

The second result is very similar, with the only difference that, $i \in C_1$ and we are dealing with $\sum_j \bar{A}_{ij}$ with $j \in C_2$. Thus, $E[(\bar{d}_i^{(2)})^2] = n(1 - \pi)\gamma_n(1 - \gamma_n)$. As before $\text{var}(\bar{d}_i^{(2)})^2 = O_P(n^2 \rho_n^2)$. The difference is in the covariance terms; for $i \neq k \in C_1$ we see that $(\bar{d}_i^{(2)})^2$ and $(d_k^{(2)})^2$ are independent, since none of the edges considered in the two can be the same; this leads to zero covariance and gives the result.

PROOF OF EQUATION 18.

$$\sum_{i \in C_1} (x_1 \bar{d}_i^{(1)} + y_1 \bar{d}_i^{(2)})^2 = \sum_{i \in C_1} (x_1^2 (\bar{d}_i^{(1)})^2 + y_1^2 (\bar{d}_i^{(2)})^2 + 2x_1 y_1 \bar{d}_i^{(1)} \bar{d}_i^{(2)})$$

We will show that $\sum_{i \in C_1} (\bar{d}_i^{(1)})^2$ and $\sum_{i \in C_1} (\bar{d}_i^{(2)})^2$ are a larger order than $\sum_{i \in C_1} \bar{d}_i^{(1)} \bar{d}_i^{(2)}$.

Lemma 5.5 shows that $\sum_{i \in C_1} (\bar{d}_i^{(1)})^2$ is $C_0 n^2 \rho_n (1 + o_P(1))$. Using similar arguments as in Lemma 5.5, one can show that $\sum_{i \in C_1} (\bar{d}_i^{(2)})^2$ is of the same order. Now, for $i \in C_1$, $E[\bar{d}_i^{(1)} \bar{d}_i^{(2)}] = 0$, since $\bar{d}_i^{(1)}$ and $\bar{d}_i^{(2)}$ are independent and have zero expectation. As for $\text{var}(\bar{d}_i^{(1)} \bar{d}_i^{(2)}) = E[(\bar{d}_i^{(1)})^2] E[(\bar{d}_i^{(2)})^2] = O_P(n^2 \rho_n^2)$. For $i \neq k \in C_1$,

$$\text{cov}(\bar{d}_i^{(1)} \bar{d}_i^{(2)}, d_k^{(1)} d_k^{(2)}) = \sum_{j_1, \ell_1 \in C_1, j_2, \ell_2 \in C_2} E[\bar{A}_{ij_1} \bar{A}_{ij_2} \bar{A}_{k\ell_1} \bar{A}_{k\ell_2}] = 0$$

The last step is true because $i \neq k \in C_1$, $j_1, \ell_1 \in C_1$, whereas $j_2, \ell_2 \in C_2$. Hence \bar{A}_{ij_2} and $\bar{A}_{k\ell_2}$ will always be unpaired and hence contribute zero to the expectation. Thus $\sum_{i \in C_1} \text{var}(\bar{d}_i^{(1)} \bar{d}_i^{(2)}) = O_P(n^3 \rho_n^2)$. This is of smaller order than $\sum_{i \in C_1} (\bar{d}_i^{(1)})^2$ and

$\sum_{i \in C_1} (\bar{d}_i^{(2)})^2$, and hence we have our result. \square

PROOF OF LEMMA 5.1. Let x and y be $v(i)$ for $i \in C_1$ and $i \in C_2$ respectively. The coefficients (x, y, λ) satisfy the following set of equations:

$$(n\pi\alpha_n - \alpha_n)x + n(1 - \pi)\gamma_n y = \lambda x; \quad n\pi\gamma_n x + (n(1 - \pi)\beta_n - \beta_n)y = \lambda y$$

If x or y is zero, then $\gamma_n = 0$, and we get the zero communication setting. When $\gamma_n > 0$, x, y are nonzero and writing $\xi = x/y$ and eliminating λ gives the following quadratic equation.

$$\pi\gamma_n \xi^2 - (\pi\alpha_n - (1 - \pi)\beta_n - (\alpha_n - \beta_n)/n)\xi - (1 - \pi)\gamma_n = 0 \quad (49)$$

Also, we have $\lambda = n(\pi\gamma_n\xi + (1 - \pi)\beta_n - \beta_n/n)$. If $\alpha_n\beta_n = \gamma_n^2$, then some algebra shows that $\lambda_2 = O_P(1)$; since this is not of the form $C_0n\rho_n$, our subsequent results will not hold. An example of this is the case where $\alpha_n = \beta_n = \gamma_n$, because then we just have a Erdős-Rényi graph on hand.

For $i \in \{1, 2\}$, $x_i = \frac{\xi_i}{\sqrt{n(\pi\xi_i^2 + (1-\pi))}}$ and $y_i = \frac{1}{\sqrt{n(\pi\xi_i^2 + (1-\pi))}}$. Some algebraic manipulation gives the first two expressions in Equation 15. From Equation 49, we also have:

$$\pi\gamma_n x^2 - (\pi\alpha_n - (1 - \pi)\beta_n - (\alpha_n - \beta_n)/n)xy - (1 - \pi)\gamma_n y^2 = 0,$$

which gives $x_1y_1 + x_2y_2 = (\pi(x_1^2 + x_2^2) - (1 - \pi)(y_1^2 + y_2^2))\frac{\gamma_n}{\pi\alpha_n - (1-\pi)\beta_n} = 0$. \square

References

- [1] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd intl. workshop on Link discovery*, 2005.
- [2] Peter J. Bickel and Aiyu Chen. A nonparametric view of network models and newman girvan and other modularities. *PNAS*, 106(50), 2009.
- [3] Bélla Bollobás. *Modern Graph Theory*. Springer-Verlag, 1998.
- [4] Kamalika Chaudhuri, Fan Chung Graham, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research - Proceedings Track*, 23:35.1–35.23, 2012.
- [5] W.E. Donath and A.J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.
- [6] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, 2005.
- [7] Miroslav Fiedler. Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.
- [8] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981.
- [9] Fan Chung Graham and Mary Radcliffe. On the spectra of general random graphs. *Electr. J. Comb.*, 18(1), 2011.
- [10] Lars W. Hagen and Andrew B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 11(9):1074–1085, 1992.
- [11] Bruce Hendrickson and Robert Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comput.*, 16(2):452–469, 1995.
- [12] P. W. Holland and S. Leinhardt. A dynamic model for social networks. *Social Networks*, 1977.
- [13] L. Katz. A new status index derived from sociometric analysis. In *Psychometrika*, volume 18, pages 39–43, 1953.
- [14] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Conference on Information and Knowledge Management*. ACM, 2003.

- [15] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [16] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.
- [17] Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *Preprint*, 2009.
- [18] Alex Pothén, Horst D. Simon, and Kan-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11(3):430–452, 1990.
- [19] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39:1878–1915, 2011.
- [20] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
- [21] Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107:1119–1128, 2012.
- [22] U. Von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.