Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Mathematik

On an error bound for spectral clustering in the stochastic block model

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

eingereicht von:	Thibaud Joel Hadamczik
geboren am:	16.7.1996
geboren in:	Weimar
Gutachter:	Prof. Dr. Markus Reiß Dr. Martin Wahl
eingereicht am:	....

Contents

Symbols

$\mathbb{1}_A$	Indicator function of a set A
$\mathbb{M}^{n \times K}$	Set of membership matrices for n vertices and K communities
$\|v\|$	Euclidean norm of a vector v
$\|M\|$	Operator norm of a matrix M induced by the Euclidean norm corresponding to the maximal singular value of M
$\|M\|_0$	ℓ_0-norm corresponding to the number of nonzero-entries of a matrix M
$\|M\|_{\max}$	Max norm corresponding to the maximal absolute entry of a matrix M
$\|M\|_F$	Frobenius norm of a matrix M corresponding to the square root of the sum of squared entries of M
λ_i^M	Eigenvalue of a matrix M with the i-th largest absolute value
$\lambda_{\max}^M$	Maximal absolute eigenvalue of a matrix M
$\lambda_{\min}^M$	Minimal absolute eigenvalue of a matrix M
$\widehat{\Theta}$	Estimated membership matrix
$\mathrm{Ber}(p)$	Bernoulli-distributed random variable with parameter p
$\mathrm{diag}(d_1, \ldots, d_n)$	Diagonal matrix with diagonal entries $d_1, \ldots, d_n$
$\mathrm{tr}(M)$	Trace of a matrix M, i.e. the sum of its diagonal elements
$\sin(M)$	Matrix obtained by applying the sine function to the entries of a matrix M
$\sin^{-1}(M)$	Matrix obtained by applying the arcsine function to the entries of a matrix M
E_K	Set of $K \times K$-column permutation matrices
G_k	Set of vertices which belong to community k
I_n	Identity mapping on $\mathbb{R}^n$
M^T	Transpose of a matrix M

$M_{\mathcal{I}*}$	Submatrix consisting of the rows $\mathcal{I}$ of a matrix M for a subset $\mathcal{I}$ of the rows of M
$M_{*\mathcal{J}}$	Submatrix consisting of the columns $\mathcal{J}$ of a matrix M for a subset $\mathcal{J}$ of the columns of M
$P^{\perp}$	Projection on the orthogonal complement of a subspace if P is a projection on a subspace
g_i	Community of a vertex i
n_k	Size of community k
$n_{\min}$	Size of the smallest community
$n'_{\max}$	Size of the second largest community

1 Introduction

Networks are omnipresent in the real world. Whether synaptic connections of neurons in the brain, the email correspondence in a large company or the friendship relations on social platforms. They can all be regarded as networks. One is often interested in identifying clusters within a network, i.e. groups of network members which are more connected amongst themselves than to members of the other groups. In the aforementioned examples those clusters would correspond to neurons which are tightly connected and thus more likely to "work on the same task", groups of people that work together, e.g. in the same department, and thus communicate a lot with each other and for example students from the same school that are likely more connected on social platform than they are to students of other schools.

There exist different methods to recover the communities from the observation of these relations - an overview can be found in [2] - one of which is spectral clustering. As the name suggests, spectral clustering uses the eigenstructure of the adjacency matrix of the network graph to cluster the network members.[1]

Recovering the community structure becomes difficult in sparse networks, i.e. when there are only few connections between the network members. Therefore, when analyzing methods to recover the communities it is interesting to see how they perform in relation to the level of sparsity. The thesis will be mainly based on [5], who prove a high probability bound on the relative number of misclassified vertices which also holds in comparably sparse networks. One key ingredient in their proof is a spectral bound for the deviation of the adjacency matrix from its expectation which is sharper than the one obtained by the Matrix Bernstein's inequality. We will give the proof of the bound obtained by the Matrix Bernstein's inequality and add details to the proofs of the statements quoted from [5], which are needed to obtain the error bound.

In Section 2 we will introduce the stochastic block model, briefly outline the idea of the spectral clustering algorithm and state the the error bound. Section 3 demonstrates the practical implications of the error bound by conducting a Monte Carlo simulation. Finally, in Section 4 we elaborate the results which are needed to obtain the error bound and prove the error bound.

[1] Instead of the adjacency matrix often related matrices are used, commonly referred to as Graph Laplacians. However, in this thesis we will focus on spectral clustering using the adjacency matrix.

2 Mathematical foundations

2.1 Stochastic block model

An undirected graph $G = (V, E)$ is characterized by a vertex set $V = \{1, 2, \dots, n\}$ and an edge set $E \subset \{\{i, j\} : i, j \in V\}$. Representing networks as graphs can be done by modeling the network members as vertices and the connections between network members as edges such that

$$\{i, j\} \in E \iff i \text{ is connected to } j.$$

Consider for example a network of five members represented by an undirected graph as depicted in Figure 1.

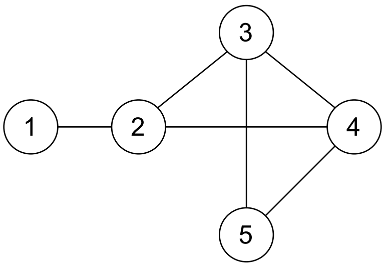


Figure 1: Network with five members represented by an undirected graph.

We obtain an equivalent representation of this graph by considering its adjacency matrix.

Definition 2.1 (Definition 3.6.2 in [8]).
Consider an undirected graph $G = (V, E)$ with n vertices. Then the **adjacency matrix** of G is the quadratic matrix

$$A = (a_{ij})_{i,j=1}^{n} \in \{0, 1\}^{n \times n} \text{ with entries } a_{ij} := \begin{cases} 0, & \text{if } \quad \{i, j\} \notin E, \\ 1, & \text{if } \quad \{i, j\} \in E. \end{cases}$$

Example 2.2.
For the graph in Figure 1 the adjacency matrix is given by

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

For undirected graphs the adjacency matrix is a symmetric matrix and therefore has a spectral decomposition $A = UDU^T$ with U being orthonormal, U^T denoting its transpose and $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ the diagonal matrix consisting of the eigenvalues of A. Our goal is to derive information about the community structure of the network from the eigenstructure of the adjacency matrix.

Remark 2.3.
We will assume that the eigenvalues are ordered according to their absolute values,

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Therefore when referring to the first/top or last/bottom eigenvalues we mean the eigenvalues with the largest or smallest absolute values, respectively.

Remark 2.4.
In this thesis we will only consider binary relations between the members of the network, i.e. the entries of the adjacency matrix are either zero or one. Another possibility is to employ weighted edges to model the strength of the relationship between different network members, see e.g. [6].

In the following we will assume that the vertex set is partitioned into K **communities** $(G_k)_{k=1}^K$, $V = \dot{\bigcup}_{k=1}^K G_k$. For a vertex $i \in \{1, \dots, n\}$ we use g_i to denote the community it belongs to, $g_i = k \iff i \in G_k$, and $n_k \coloneqq |G_k|$ denotes the **size of community** k. The community membership can be represented by a membership matrix in the following sense.

Definition 2.5 (See section 1 in [5]).
Let $\dot{\bigcup}_{k=1}^K G_k = V = \{1, \dots, n\}$ be a partition of the vertex set into K communities. Then the entries of the **membership matrix** $\Theta \in \{0,1\}^{n \times K}$ are defined by

$$\theta_{ij} = \begin{cases} 1, \text{ if } & g_i = j, \\ 0, \text{ if } & g_i \neq j. \end{cases}$$

So each row of the membership can be related to a vertex and contains a one in the column corresponding to the community the vertex belongs to. We will now include uncertainty in our setup and assume that for each pair of communities, edges between members of these communities occur with a certain probability. These probabilities will be given by the connectivity matrix.

Definition 2.6 (See section 2.1 in [5])**.**
Let $\dot{\bigcup}_{k=1}^{K} G_k = V = \{1, \dots, n\}$ be a partition of the vertex set into K communities. Then the **connectivity matrix** $B = (b_{k\ell})_{k,\ell=1}^{K} \in [0,1]^{K \times K}$ is the symmetric matrix whose entries $b_{k\ell}$ contain the probability of an edge occurring between a vertex of community k and a vertex of community ℓ.

Now, given Θ and B, we will consider an adjacency matrix with the distribution of the entries given by those two matrices in the following sense.

Definition 2.7 (See section 2.1 in [5])**.**
For a membership matrix $\Theta \in \{0,1\}^{n \times K}$ and a connectivity matrix $B \in [0,1]^{K \times K}$ a **stochastic block model (SBM)** parameterized by Θ and B is a probability distribution on the edges of a graph G, such that the entries of the associated adjacency matrix are random and given by

$$a_{ij} = \begin{cases} \mathrm{Ber}(b_{g_i g_j}), & \text{if} \quad i \leq j, \\ a_{ji}, & \text{if} \quad i > j, \end{cases}$$

where $\mathrm{Ber}(b_{g_i g_j})$ denotes a Bernoulli-distributed random variable with parameter $b_{g_i g_j}$.

For a SBM parameterized by Θ and B we define

$$P \coloneqq \Theta B \Theta^T = \mathbb{E}[A] \in \mathbb{R}^{n \times n}, \tag{1}$$

where $\mathbb{E}[A]$ is the **expected adjacency matrix** and the **noise** or **noise matrix**

$$R \coloneqq A - P$$

as the deviation of the adjacency matrix from its expectation. For each pair of vertices, i, j, the corresponding entries of P contain the probability p_{ij} that these two vertices are connected. For an adjacency matrix A, P is again a symmetric matrix and hence allows for a spectral decomposition with orthonormal matrices.

Example 2.8 (SBM with two communities)**.**
For a SBM with 2 communities of the same size $n_1 = n_2 = \frac{n}{2}$ (for n even), where the probability of two vertices being connected to each other only depends on whether or not they belong to the same community and the vertices are labeled such that the first $\frac{n}{2}$

vertices belong to the first community, the connectivity and membership matrices are given by

$$B = \begin{bmatrix} p & q \\ q & p \end{bmatrix} \qquad \text{and} \qquad \Theta = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{array}{c} 1 \\ \vdots \\ 1 \end{array}}\right\} \frac{n}{2} \\ \left.\vphantom{\begin{array}{c} 0 \\ \vdots \\ 0 \end{array}}\right\} \frac{n}{2} \end{array} \tag{2}$$

and the expected adjacency matrix is given by

$$P = \Theta B \Theta^T = \begin{array}{c} \begin{array}{cc} \overbrace{\hphantom{p \cdots\cdots p}}^{\frac{n}{2}} & \overbrace{\hphantom{q \cdots\cdots q}}^{\frac{n}{2}} \end{array} \\ \begin{bmatrix} p & \cdots & p & q & \cdots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \cdots & p & q & \cdots & q \\ q & \cdots & q & p & \cdots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \cdots & q & p & \cdots & p \end{bmatrix} \end{array} \begin{array}{l} \\ \left.\vphantom{\begin{array}{c} p \\ \vdots \\ p \end{array}}\right\} \frac{n}{2} \\ \left.\vphantom{\begin{array}{c} q \\ \vdots \\ q \end{array}}\right\} \frac{n}{2} \end{array},$$

where p is the probability for an edge occurring between vertices of the same community, while q is the respective probability for vertices of different communities. So, by convenient numbering of the vertices, P has a block structure, with each block indicating the edge probabilities for the vertices of a pair of communities.

Now, given an adjacency matrix sampled from a SBM, we want to recover the community structure of the associated network. From Example 2.8 one can see that the rows of P corresponding to two vertices which belong to the same community are the same, which is also true for a higher number of communities. As we will see in the next section, the specific structure of P will carry over to its spectral decomposition and allow us to distinguish vertices of different communities by suitable eigenvectors of P. This will give rise to the spectral clustering algorithm which uses the eigenstructure of the observed adjacency matrix to recover the community membership of the vertices.

2.2 Spectral clustering algorithm

After having introduced the basic notions we now want to present the spectral clustering algorithm, state the results on the error bound and give an overview of the steps which are necessary to prove the error bound.

As described in [6], spectral clustering can be derived from a graph partitioning problem, which consists in finding partitions of a graph such that the vertices of the resulting subgraphs are highly connected within each subgraph but only loosely connected to vertices of the other subgraphs. Since solving the corresponding discrete optimization problem involves comparing all possible partitions of the vertex set, it is NP-hard. Therefore one relaxes the problem, tries to find a solution for the continuous version of it and then discretizes the solution. For an arbitrary graph one cannot be sure that the solution obtained by this approach is a good approximation to the discrete solution. However, in our case the graph we want to cluster is not arbitrary but sampled according to a SBM which will allow for statements about the accuracy of the obtained clustering.

Our goal is to recover the community membership of the vertices which was formally defined by a membership matrix Θ. We will therefore consider the error measures

$$L(\widehat{\Theta}, \Theta) := n^{-1} \min_{J \in E_K} \left\| \widehat{\Theta} J - \Theta \right\|_0 \tag{3}$$

and

$$\tilde{L}(\widehat{\Theta}, \Theta) := \min_{J \in E_K} \max_{1 \leq k \leq K} n_k^{-1} \left\| \left(\widehat{\Theta} J \right)_{G_k *} - \Theta_{G_k *} \right\|_0 \tag{4}$$

for an algorithm that returns an (estimated) membership matrix $\widehat{\Theta}$. Here the ℓ_0-norm $\|M\|_0$ counts the number of nonzero-entries of a matrix M, E_K denotes the set of $K \times K$-column permutation matrices and, for an $m \times n$ matrix M and subsets $\mathcal{I} \subset \{1, ..., m\}, \mathcal{J} \subset \{1, ..., n\}$, $M_{\mathcal{I}*}$ and $M_{*\mathcal{J}}$ denote the submatrices consisting of only the rows $\mathcal{I}$ or columns $\mathcal{J}$ of M.

$L(\widehat{\Theta}, \Theta)$ counts the fraction of entries of $\widehat{\Theta}$ which are not equal to the corresponding entries of Θ and thus measures the fraction of misclassified vertices. One estimated membership matrix $\widehat{\Theta}$ only represents one possible labeling of the communities but we are interested in the community structure and not the specific labeling. Therefore, it does not matter which specific community a vertex is assigned to but only which other vertices are assigned to the same community. Hence, we account for the different labelings by considering the minimum error over all column permutations of $\widehat{\Theta}$. Recalling that G_k is the set of all vertices in community k, we see that $\tilde{L}(\widehat{\Theta}, \Theta)$ measures the fraction of misclassified vertices in the community with the highest fraction of misclassified vertices, where we again consider the minimum over all column permutations of $\widehat{\Theta}$ to account for different labelings.

Having established how we want to measure the performance of an algorithm that returns an estimated membership matrix $\widehat{\Theta}$, we will now state a lemma that implies that the eigenstructure of P as defined in (1) allows us to retrieve the underlying community structure. We will denote the Euclidean norm of a vector v by $\|v\|$ and the operator norm induced by the Euclidean norm of a matrix M, which corresponds to the largest singular value of M, by $\|M\|$.

Lemma 2.9 (Lemma 2.1 in [5])**.**
For a stochastic block model with K communities parameterized by a membership matrix Θ and a connectivity matrix B of full rank, the spectral decomposition of $P = \Theta B \Theta^T$ can be written as $P = UDU^T$ with $U = \Theta X$ for $X \in \mathbb{R}^{K \times K}$ satisfying $\|X_{k*} - X_{\ell*}\| = \sqrt{n_k^{-1} + n_\ell^{-1}}$ for all $1 \leq k < \ell \leq K$. $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_K)$ denotes the matrix with the non-zero eigenvalues of P on the diagonal.

Remark 2.10.
Requiring B to be of full rank excludes the possibility of two vertices from different communities having exactly the same edge probabilities which would make it impossible to say which communities the vertices belong to even when knowing the respective probabilities. More formally, one implication of B not being full rank is P having rank smaller than K which is equivalent to P having less than K non-zero eigenvalues. As we will see, spectral clustering builds on relating the eigenvectors with respect to the K largest absolute eigenvalues of P to those of A which is not well-defined in the case of less than K non-zero eigenvalues.

Proof of Lemma 2.9.
According to Remark 2.10 P has K non-zero eigenvalues. By convenient ordering of the eigenvalues in the diagonal matrix D, we can therefore write the spectral decomposition of P by $P = UDU^T$ with $U \in \mathbb{R}^{n \times K}$ having the K eigenvectors with respect to the

non-zero eigenvalues of P as column vectors. Defining $\Delta := \operatorname{diag}(\sqrt{n_1}, \ldots, \sqrt{n_K})$ we can write, using symmetry of Δ,

$$P = \Theta B \Theta^T = \Theta\Delta^{-1}\Delta B \Delta(\Delta^{-1}\Theta^T) = \Theta\Delta^{-1}\Delta B\Delta(\Theta\Delta^{-1})^T. \tag{5}$$

We first check that the columns of $\Theta\Delta^{-1}$ are orthonormal, i.e. $(\Theta\Delta^{-1})^T\Theta\Delta^{-1} = I_K$, with I_K denoting the identity mapping on $\mathbb{R}^K$ (we will use I if the dimension is clear)

$$\begin{aligned}(\Theta\Delta^{-1})^T\Theta\Delta^{-1} &= \Delta^{-1}(\Theta^T\Theta)\Delta^{-1}\\ &\overset{(a)}{=} \operatorname{diag}\left(\sqrt{n_1^{-1}}, \ldots, \sqrt{n_K^{-1}}\right) \operatorname{diag}(n_1, \ldots, n_K) \operatorname{diag}\left(\sqrt{n_1^{-1}}, \ldots, \sqrt{n_K^{-1}}\right)\\ &= I_K,\end{aligned}$$

where for (a) we used that each row of the membership matrix Θ contains exactly one entry with value one while the other entries are zero and each column ℓ contains n_ℓ entries with value one while the other entries are zero. Since

$$(\Delta B\Delta)^T = (B\Delta)^T\Delta^T = \Delta^T B^T \Delta^T = \Delta B \Delta,$$

$\Delta B \Delta$ is also symmetric and has a spectral decomposition $\Delta B\Delta = ZDZ^T$ with Z orthonormal and D a diagonal matrix which will turn out to be the same as in the spectral decomposition of P. Inserting this into (5) yields

$$P = \Theta\Delta^{-1}ZDZ^T(\Theta\Delta^{-1})^T = (\Theta\Delta^{-1}Z)D(\Theta\Delta^{-1}Z)^T. \tag{6}$$

Having already verified that the columns of $\Theta\Delta^{-1}$ are orthonormal it follows that also the columns of $\Theta\Delta^{-1}Z$ are orthonormal. Thus, we can write $U = \Theta\Delta^{-1}Z$. Now, we set $X = \Delta^{-1}Z$ and show that the rows of X satisfy the claimed properties. Since Z is orthonormal its rows are orthogonal unit vectors. Multiplying with Δ^{-1} scales the i-th row by $\frac{1}{\sqrt{n_i}}$ such that the resulting matrix X has orthogonal rows with norm

$$\|X_{i*}\| = \|(\Delta^{-1}Z)_{i*}\| = \left\|\frac{1}{\sqrt{n_i}}Z_{i*}\right\| = \frac{1}{\sqrt{n_i}}\|Z_{i*}\| = \frac{1}{\sqrt{n_i}}.$$

Using this, we get

$$\begin{aligned}\|X_{k*} - X_{\ell*}\| &= \sqrt{\langle X_{k*} - X_{\ell*}, X_{k*} - X_{\ell*}\rangle} \\ &= \sqrt{\langle X_{k*}, X_{k*}\rangle - 2\langle X_{k*}, X_{\ell*}\rangle + \langle X_{\ell*}, X_{\ell*}\rangle} \\ &= \sqrt{n_k^{-1} + n_\ell^{-1}},\end{aligned}$$

which concludes the proof. □

The lemma shows that the matrix U emerging in the spectral decomposition of P can be written as the product of the underlying membership matrix and a $K \times K$-matrix X whose rows are separated by a distance depending on the community sizes $n_1, \dots, n_K$. Therefore, under the assumption of B having full rank, two rows of U corresponding to vertices of the same community are the same while they differ if they correspond to vertices of different communities.

Example 2.11 (Exercise 4.5.2 in [8])**.**
Let us take up Example 2.8 from above. In Appendix A.1 it is shown that the two non-zero eigenvalues of $P = \mathbb{E}[A]$ are

$$\lambda_1 = n\frac{(p+q)}{2} \text{ and } \lambda_2 = n\frac{(p-q)}{2} \text{ with eigenvectors } v_1 = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \text{ and } v_2 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{bmatrix}.$$

Therefore we obtain

$$U = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ \hline 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix}$$

and clustering of the rows of U into two communities and assigning each vertex according to the clustering of the corresponding row to one of the two communities leads to a correct assignment of the vertices to their respective communities.

In general, the matrix $\widehat{U}$ containing the eigenvectors with respect to the non-zero eigenvalues of an adjacency matrix, will not have exactly similar rows for vertices of the same communities. Therefore we need to apply a clustering algorithm to the rows of $\widehat{U}$ that yields a solution to the K-means problem

$$\min_{\substack{\Theta \in \mathbb{M}^{n\times K} \\ X \in \mathbb{R}^{K\times K}}} \|\Theta X - \widehat{U}\|_F^2, \tag{7}$$

where $\|M\|_F$ denotes the Frobenius norm of a matrix M. By the structure of the membership matrices for a SBM with n vertices and K communities $\mathbb{M}^{n\times K}$, this is equivalent to clustering the rows of $\widehat{U}$ into K clusters by the K-means algorithm and then assigning the vertices to communities according to the clustering.[2]

Given a set of n vectors in $\mathbb{R}^m$ the K-means algorithm yields K mean-vectors in $\mathbb{R}^m$ and an assignment of each of the n vectors to one of the K vectors such that the sum of distances in the squared Euclidean norm between the n vectors and their assigned mean-vector becomes minimal. This is equivalent to solving (7) since $\|\cdot\|_F^2$ measures the sum of distances between the rows of ΘX and $\widehat{U}$. Formally, for $R = (r_{ij}) := \Theta X - \widehat{U} \in \mathbb{R}^{n\times K}$,

$$\|R\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{K} r_{ij}^2 = \sum_{i=1}^{n} \|R_{i*}\|^2,$$

and by definition of $\mathbb{M}^{n\times K}$, the maximal number of different rows of ΘX is K leading to the constraint of K different vectors in the algorithm. Finding a solution to (7) is NP-hard but we will use the fact that it can be implemented efficiently by the algorithm presented in [4] to find a $(1+\varepsilon)$-approximate solution. By applying the K-means algorithm to the rows of $\widehat{U}$ we obtain a membership matrix $\widehat{\Theta}$ yielding a clustering of the vertices into K communities. Summarized this yields the following algorithm:

[2] See Chapter 14.3.6 in [3] for the K-means algorithm.

Algorithm 1: Spectral Clustering Algorithm for the SBM (Algorithm 1 in [5])

```
Input : Adjacency matrix $A$; number of communities $K$; approximation
        parameter $\varepsilon$
Output : Membership matrix $\widehat{\Theta} \in \mathbb{M}^{n\times K}$
Calculate the matrix $\widehat{U} \in \mathbb{R}^{n\times K}$ where the $k$-th column of $\widehat{U}$ corresponds to the
  eigenvector with respect to the $k$-th eigenvalue of $A$
Compute a $(1+\varepsilon)$-approximate solution $(\widehat{\Theta}, \widehat{X})$ to the $K$-means problem (7) with
  $K$ clusters for the matrix $\widehat{U}$
Return $\widehat{\Theta}$
```

In Section 4.4 we will prove the following result which provides a bound on the sum of the fractions of misclassified vertices in the different communities.

Theorem 2.12 (Theorem 3.1 in [5]).
Let A be an adjacency matrix generated from a stochastic block model (Θ, B). Assume that $P = \Theta B \Theta^T$ is of rank K with smallest absolute non-zero eigenvalue at least γ_n and $\max_{k,l} b_{k\ell} = \alpha$ for some $\alpha \geq \log(n)/n$. Let $\widehat{\Theta}$ be the output of the spectral clustering algorithm 1. Then there exists an absolute constant $c > 0$, such that, if

$$(2+\varepsilon)\frac{Kn\alpha}{\gamma_n^2} < c, \tag{8}$$

with probability at least $1 - n^{-1}$, there exist subsets $S_k \subset G_k$ for $k = 1, \dots, K$ and a $K \times K$ permutation matrix J such that $\widehat{\Theta}_{G*}J = \Theta_{G*}$, where $G = \bigcup_{k=1}^{K}(G_k \setminus S_k)$, and

$$\sum_{k=1}^{K} \frac{|S_k|}{n_k} \leq c^{-1}(2+\varepsilon)\frac{Kn\alpha}{\gamma_n^2}. \tag{9}$$

Remark 2.13.
Requiring $\alpha \geq \log(n)/n$ or equivalently $n\alpha \geq \log(n)$ can be understood as a sparsity condition. We define $d := n\alpha$, as the maximal expected degree. The error bound given in the theorem only holds, when the maximal expected degree is at least of order $\log(n)$.

The existence of a permutation matrix J such that $\widehat{\Theta}_{G*}J = \Theta_{G*}$ implies that up to a different labeling, the membership matrix $\widehat{\Theta}$ estimated by the spectral clustering algorithm yields a correct clustering for the vertices contained in $G = \bigcup_{k=1}^{K}(G_k \setminus S_k)$, where S_k is the set of vertices in community k, which are potentially misclassified by the spectral clustering algorithm. Thus, (9) provides a bound on the sum of fractions of misclassified vertices.

To see which effect the sparsity has on the performance of the clustering algorithm, we will consider a connectivity matrix B and then see how clustering works when B is scaled by a scaling parameter $\alpha \in [0,1]$. The following statement is a corollary of Theorem 2.12, provides bounds on the error measures defined in (3) and (4), and relates them to α. It is also proved in Section 4.4.

Corollary 2.14 (Corollary 3.2 in [5]).
Let A be an adjacency matrix from the SBM (Θ, B_α), where $B_\alpha = \alpha B$ for some $\alpha \geq \log(n)/n$ and with B having minimum absolute eigenvalue at least $\lambda > 0$ and $\max_{k,\ell} b_{k\ell} = 1$. Let $\widehat{\Theta}$ be the output of the spectral clustering algorithm 1, $n_{\min}$ the size of the smallest community and $n'_{\max}$ the size of the second largest community. Then there exists a constant c such that if

$$(2+\varepsilon)\frac{Kn}{n_{\min}^2 \lambda^2 \alpha} < c,$$

then with probability at least $1 - n^{-1}$

$$\tilde{L}(\widehat{\Theta}, \Theta) \leq c^{-1}(2+\varepsilon)\frac{Kn'_{\max}}{n_{\min}^2 \lambda^2 \alpha} \tag{10}$$

and

$$L(\widehat{\Theta}, \Theta) \leq c^{-1}(2+\varepsilon)\frac{Kn}{n_{\min}^2 \lambda^2 \alpha}. \tag{11}$$

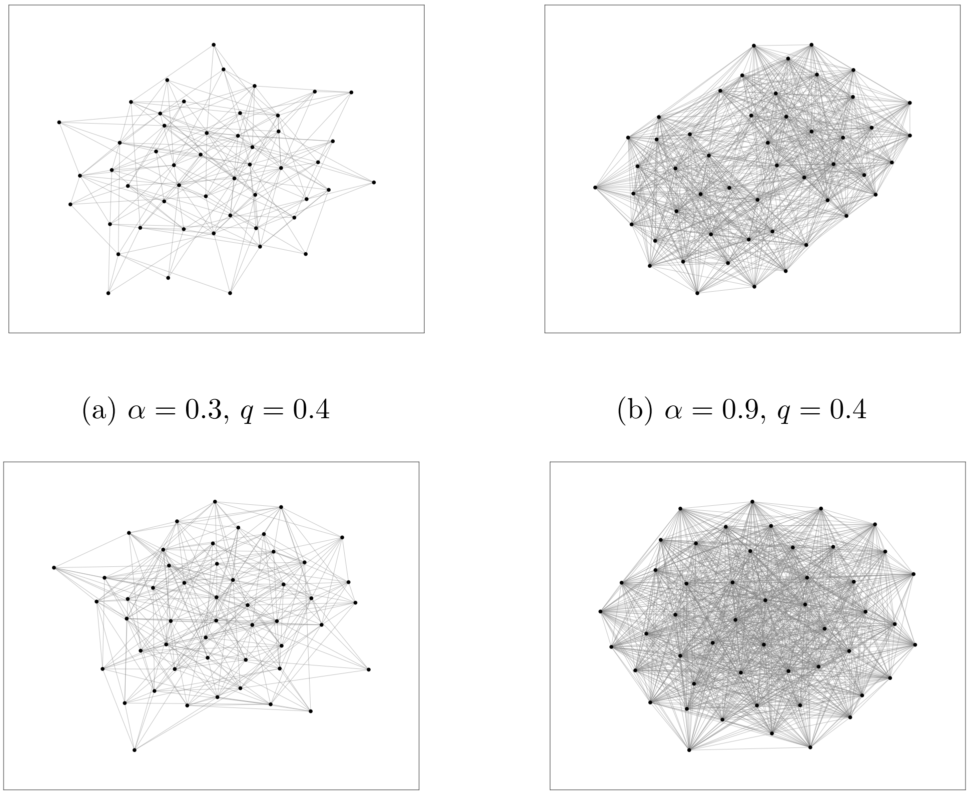
The proofs of Theorem 2.12 and its Corollary 2.14 can be decomposed into four steps:

- By Lemma 2.9, the eigenstructure of P contains information about the community structure
- The distance of the matrices containing the eigenvectors of A and P can be bounded by a factor times the distance between A and P
- A is close to P in the operator norm with high probability
- Applying an approximate K-means clustering algorithm to the eigenvectors of A yields a solution which is close to the optimal solution of the K-means problem

By Lemma 2.9 the eigenstructure of P contains information about the community structure. In particular we know the space which is spanned by the first K eigenvectors of P. In Section 4.1 we will introduce a distance on subspaces of $\mathbb{R}^n$ (in our case the spaces spanned by the first K eigenvectors) and bound the distance of the respective spaces of A and P by the distance of the noise $\|R\| = \|A - P\|$. In Section 4.2 we will argue that the noise is relatively small (with high probability) and therefore also the distance between the subspaces is small (with high probability). Then, in Section 4.3, we will show that a $(1+\varepsilon)$-approximate solution of the K-means problem (7) provides a reasonable clustering if the spaces are close to each other, i.e. $\|\widehat{U} - U\|_F$ is small.
Before deriving the error bounds in Section 4, the next section examines how the spectral algorithm performs in practice and how it behaves when varying different parameters.

3 Simulation

In this section we want to examine how the bound given by Corollary 2.14 on the relative clustering error $L(\widehat{\Theta}, \Theta)$ defined in (3) translates into practice by picking up Example 2.8 from before and see how the spectral clustering algorithm performs depending on the different parameters.



(a) $\alpha = 0.3$, $q = 0.4$

(b) $\alpha = 0.9$, $q = 0.4$

(c) $\alpha = 0.3$, $q = 0.8$

(d) $\alpha = 0.9$, $q = 0.8$

Figure 2: Graphs sampled according to a SBM with two communities of the same size with connectivity matrix αB for different values of α and q with $n = 50$.

We set $p = 1$ to satisfy the condition on the maximal entry being one in Corollary 2.14 and obtain the connectivity matrix

$$B = \begin{bmatrix} 1 & q \\ q & 1 \end{bmatrix}. \tag{12}$$

In the following we will consider the scaled matrix

$$B_\alpha = \alpha B$$

with a scaling parameter $\alpha \in (0, 1]$. We first sample graphs according to a SBM (Θ, B_α) with Θ defined as in (2) and $n = 50$ for four different combinations of the parameters p and α to give an intuition of the effect of the different parameters. The resulting graphs are depicted in Figure 2.

The graph in Figure 2b stands out, since one can almost immediately detect the two communities, which are very densely connected while there are fewer connections between vertices of different communities. Regarding the other graphs, assigning the vertices to two communities just by looking at the graphs is a hard task. An inductive conclusion from this example would be that it is easier to recover the communities for high values of α and small values of q and that we expect spectral clustering to work better in these cases.

To formally examine the effect of the different parameters we consider the eigenvalues of B, which are $\lambda_1 = \frac{1+q}{2}$ and $\lambda_2 = \frac{1-q}{2}$ (see Appendix A.1) such that the error bound given by Corollary 2.14 applies to our setting with number of communities $K = 2$, size of the smallest community $n_{\min} = \frac{n}{2}$ and $\lambda = \frac{1-q}{2}$ the smallest absolute eigenvalue of B, yielding

$$L(\widehat{\Theta}, \Theta) \leq c^{-1}(2 + \varepsilon)\frac{64}{n(1-q)^2\alpha}$$

with probability at least $1 - n^{-1}$ for a constant $c > 0$. In our setting the number of communities and the ratio of their sizes is fixed such that the error bound will solely depend on the total number of vertices, the scaling parameter α and the value of q (and the approximation parameter ε which we assume to be fixed since we use the K-means algorithm implemented in Matlab).

The total number of vertices appears twice in the statement on the error bound, once in the error bound and once in the probability with which the error bound holds. A higher value of n thus yields a lower error bound and higher probability with which the error bound holds. Therefore we expect a lower clustering error for higher values of n when the other parameters are not varied. A smaller value of α should lead to a worse clustering outcome since it loosens the bound on the error while a lower value of q has the opposite effect. A lower value of q leading to a lower expected error is not surprising since it is the probability for edges occurring between vertices of different communities. The lower q, the greater is the difference of the probabilities for an edge occurring within and between communities, respectively, making it easier to detect the communities as is illustrated by comparing Figures 2b and 2d. Furthermore the bound is only guaranteed to hold if the maximal expected degree has value at least $\log(n)$, i.e. $n\alpha \geq \log(n)$. In our setting, due to reasonably large n, this will be satisfied unless α is close to zero.

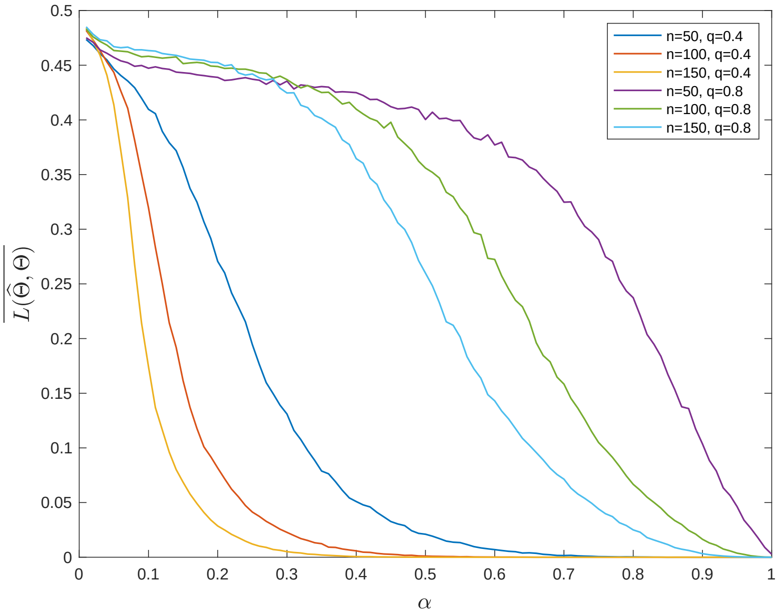


Figure 3: Mean fraction of misclassified vertices in dependence of α for different values of n and q.

To illustrate the results and thus the consistency of the error bound we considered the two values of q which were already employed in Figure 2, three different total numbers of vertices n and scaling parameters $\alpha_i = \frac{i}{100}$ for $i = 1, \cdots, 100$. To account for the randomness we conducted a Monte Carlo simulation with 1000 sampled adjacency matrices for each combination of the parameters and, for each parameter combination, calculated the mean error over all samples $\overline{L(\widehat{\Theta}, \Theta)}$ (see Appendix A.2 for the code). The results are shown in Figure 3 which shows the mean error $\overline{L(\widehat{\Theta}, \Theta)}$ corresponding to the mean fraction of misclassified vertices for the different parameter combinations.

It can be observed that the simulation results reflect the predictions of the error bound and higher values of n and lower values of q do indeed lead to a decreased clustering error as long as α is not too close to zero. The effect of the sparsity parameter α is also as predicted yielding better clustering for higher values of α with the error tending to zero for α close to one.

Now that we have exemplified the results, the next section will show how we can derive the error bounds we have stated in Section 2.2.

4 Derivation of the error bound

4.1 Subspace perturbation

In Section 2 we have established that the eigenstructure of the expected adjacency matrix P contains information about the community structure and that we can recover the community membership of a network with K communities based on the K-dimensional subspace spanned by the eigenvectors corresponding to the top K eigenvalues of P. Since the idea of spectral clustering is to recover the communities by the top K eigenvectors of an adjacency matrix A, the next step is to see whether the respective subspaces of A and P are close.

We will argue that, under certain conditions, for symmetric matrices that are close to each other in terms of the operator norm, it can also be ensured that the spaces spanned by a subset of their eigenvectors do not deviate too much from each other. Therefore, in this section we consider two subspaces S_1, S_2 of $\mathbb{R}^n$ and define a suitable distance on them. The definitions and theorems presented are based on [7] and [1].
Let us first define a distance between a single vector and a subspace.

Definition 4.1 (Definition II.4.2. in [7])**.**
Let S_1 be a subspace of $\mathbb{R}^n$ and let $s_2 \in \mathbb{R}^n$. If $\nu : \mathbb{R}^n \to \mathbb{R}_+$ is a norm on $\mathbb{R}^n$, then the **ν-distance between s_2 and S_1** is the function

$$\delta_\nu := \min_{s_1 \in S_1} \nu(s_2, s_1).$$

Generalizing from a single vector to another subspace leads to the following definition.

Definition 4.2 (Definition II.4.3. in [7])**.**
Let S_1, S_2 be K-dimensional subspaces of $\mathbb{R}^n$ and let ν be a norm on $\mathbb{R}^n$. Then the **ν-gap between S_1 and S_2** is the number

$$\rho_{g,\nu} := \max \left\{ \max_{\substack{s_1 \in S_1 \\ \nu(s_1)=1}} \delta_\nu(s_1, S_2), \max_{\substack{s_2 \in S_2 \\ \nu(s_2)=1}} \delta_\nu(s_2, S_1) \right\}.$$

According to the so called CS-decomposition (see Chapter I.5 in [7]), for orthonormal bases of two K-dimensional subspaces of $\mathbb{R}^n$ written as column vectors of matrices $U, \widehat{U}$, we can decompose these matrices in a way which allows us to extract values that can be interpreted as sines and cosines of angles between the basis vectors of the subspaces.

Theorem 4.3 (Theorem I.5.2 in [7])**.**
Let $U, \widehat{U} \in \mathbb{R}^{n \times K}$ with $U^T U = \widehat{U}^T \widehat{U} = I_n$. If $2K \leq n$, there are orthonormal matrices Q, V and W such that

$$QUV = \begin{array}{r@{}c} & K \\ \begin{array}{r} K \\ K \\ n-2K \end{array} & \begin{bmatrix} I_K \\ 0 \\ 0 \end{bmatrix} \end{array} \qquad \textit{and} \qquad Q\widehat{U}W = \begin{array}{r@{}c} & K \\ \begin{array}{r} K \\ K \\ n-2K \end{array} & \begin{bmatrix} \Gamma \\ \Sigma \\ 0 \end{bmatrix} \end{array}.$$

where

$$\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_K) \text{ and } \Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_K)$$

satisfy

$$\begin{aligned} 0 \leq \gamma_1 \leq \cdots \leq \gamma_K, \\ 0 \leq \sigma_K \leq \cdots \leq \sigma_1, \\ \gamma_i^2 + \sigma_i^2 = 1, i = 1, \dots, K. \end{aligned}$$

The case $2K > n$ permits a similar decomposition but will not be explicitly regarded since the number of communities will usually satisfy $2K \leq n$.

Since $\sigma_i^2 + \gamma_i^2 = 1$ we can interpret $\sigma_i, \gamma_i,\ i = 1, \dots, K$, as sines and cosines of angles which leads to the next definition.

Definition 4.4 (Definition I.5.3. in [7])**.**
Let S_1, S_2 be subspaces of $\mathbb{R}^n$ of the same dimension. The **canonical angles** between S_1 and S_2 are the diagonal entries of the matrix

$$\Theta\left(S_1, S_2\right) := \sin^{-1}(\Sigma),$$

where Σ is the matrix of Theorem 4.3 and $\sin^{-1}(\Sigma)$ denotes the matrix obtained by applying the arcsine function to the entries of Σ.

Note that we are also using Θ to denote the membership matrix of a SBM. Its intended use will be clear from the context. Let us recall that for a subspace $S \subset \mathbb{R}^n$ the **orthogonal projection onto** S is given by

$$P_S = Q_S Q_S^T, \tag{13}$$

with the columns of Q_S forming an orthonormal basis for S and the projection on the orthogonal complement $S^\perp$ of S is given by

$$P_S^\perp = I - P_S.$$

The next theorem shows that the singular values of $P_{S_1}(I - P_{S_2})$ correspond to the sines of the canonical angles between S_1 and S_2.

Theorem 4.5 (Theorem I.5.5 in [7]).
Let S_1 and S_2 be K-dimensional subspaces of $\mathbb{R}^n$. Let

$$k = \begin{cases} K & , \text{ if } 2K \leq n \\ n - K & , \text{ if } 2K > n. \end{cases}$$

Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ be the sines of the canonical angles between S_1 and S_2. Then the singular values of $P_{S_1}(I - P_{S_2})$ are

$$\sigma_1, \sigma_2, \ldots, \sigma_k, \overbrace{0, \ldots, 0}^{n-k}.$$

It can be shown that the ν-gap $\rho_{g,\nu}$ is a metric (see Theorem II.4.7 in [7]). With the decomposition in Theorem 4.3 we can characterize the $\|\cdot\|$-gap as the sine of the largest angle by the following theorem.

Theorem 4.6 (Theorem II.4.5. in [7]).
Let S_1, S_2 be K-dimensional subspaces of $\mathbb{R}^n$ and let $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_K)$, where $\theta_1 \geq \cdots \geq \theta_K$ are the canonical angles between S_1 and S_2. Then

$$\rho_{g,2}\left(S_1, S_2\right) = \sin\theta_1 = \|\sin\Theta\|,$$

where $\sin\Theta$ is the matrix obtained by applying the sine function to the entries of Θ.

Now we want to estimate the distance between two subspaces by the following result.

Theorem 4.7 (Theorem VII.3.1 in [1])**.**
Let A, P be normal operators. Let S_1 and S_2 be two subsets of the complex plane that are separated either by an annulus of width δ or a strip of width δ. Let $E = P_A(S_1)$ and $F = P_P(S_2)$. Then,

$$\|EF\| \leq \frac{1}{\delta}\|E(A-P)F\| \leq \frac{1}{\delta}\|A-P\|. \tag{14}$$

Here $P_A(S_1)$ denotes the orthogonal projection onto the subspace spanned by the eigenvectors of A corresponding to the eigenvalues contained in S_1.

Since the adjacency matrix of a random undirected graph and its expectation are both symmetric and thus normal, we can use Theorem 4.7 in combination with Theorems 4.5 and 4.6 to bound the distance of the spaces spanned by their first K eigenvectors.

Lemma 4.8 (Lemma 5.1 in [5])**.**
Assume that $P \in \mathbb{R}^{n\times n}$ is a rank K symmetric matrix with smallest non-zero absolute eigenvalue γ_n. Let A be any symmetric matrix and let $\widehat{U}, U \in \mathbb{R}^{n\times K}$ be the K leading (unit) eigenvectors of A and P, respectively. Then there exists a $K\times K$ orthogonal matrix Q such that

$$\|\widehat{U} - UQ\|_F \leq \frac{2\sqrt{2K}}{\gamma_n}\|A-P\|. \tag{15}$$

Proof of Lemma 4.8.
By Proposition 2.2 in [9] and using the fact that

$$\forall A \in \mathbb{R}^{n\times m} \text{ of rank } K : \|A\|_F^2 = \operatorname{tr}(A^T A) \leq K\|A\|^2, \tag{16}$$

with $\operatorname{tr}(A^T A)$ denoting the trace of $A^T A$, we can bound the minimum distance (measured by the squared Frobenius norm) between the orthonormal bases of two subspaces, S_1, S_2, by K times the squared $\|\cdot\|$-gap of the subspaces which, by Theorems 4.6 and 4.5, is given by the largest singular value of $P_{S_1}(I - P_{S_2})$. We conclude that there exists an orthogonal matrix $Q \in \mathbb{R}^{K\times K}$ such that

$$\frac{1}{2}\|\widehat{U} - UQ\|_F^2 \leq \|(I - \widehat{U}\widehat{U}^T)UU^T\|_F^2 \leq K\|(I - \widehat{U}\widehat{U}^T)UU^T\|^2.$$

Multiplying by 2 and taking the square root yields

$$\|\widehat{U} - UQ\|_F \leq 2\|(I - \widehat{U}\widehat{U}^T)UU^T\|_F \leq \sqrt{2K}\|(I - \widehat{U}\widehat{U}^T)UU^T\|. \tag{17}$$

We will now show that

$$\|(I - \widehat{U}\widehat{U}^T)UU^T\| \leq 2\frac{\|A - P\|}{\gamma_n}. \tag{18}$$

Therefore consider first the case that $\|A - P\| \leq \gamma_n/2$. Since A and P are both symmetric and thus normal, the idea is to apply Theorem 4.7 identifying $E = I - \widehat{U}\widehat{U}^T$ as the projection onto the subspace spanned by the eigenvectors corresponding to the bottom $n - K$ eigenvalues $\lambda^A_{K+1}, \ldots, \lambda^A_n$ of A and $F = UU^T$ as the projection onto the subspace spanned by the eigenvectors $\lambda^P_1, \ldots, \lambda^P_K$ corresponding to the top K eigenvalues of P. Now we need to find subsets of $S_1, S_2 \subset \mathbb{R}$ such that $\lambda^A_{K+1}, \ldots, \lambda^A_n \in S_1$ and $\lambda^P_1, \ldots, \lambda^P_K \in S_2$. If the distance δ between S_1 and S_2 satisfies $\delta \geq \gamma_n - \|A - P\|$, Theorem 4.7 yields

$$\|(I - \widehat{U}\widehat{U}^T)UU^T\| = \|EF\| \leq \frac{\|A - P\|}{\gamma_n - \|A - P\|} \leq 2\frac{\|A - P\|}{\gamma_n}.$$

By assumption, P is of rank K and therefore has K non-zero eigenvalues, for which it holds that their absolute value is bounded below by γ_n. Hence, we can define $S_2 = \mathbb{R} \setminus [-\gamma_n, \gamma_n]$. Since the bottom $n - K$ eigenvalues of P vanish, by Weyl's inequality (Corollary IV.4.10 in [7]), we can bound the absolute value of the bottom $n - K$ eigenvalues of A by the norm of the noise and get

$$|\lambda^A_i| = |\lambda^A_i - \lambda^P_i| \leq \|A - P\|, i = K + 1, \ldots, n.$$

Therefore we can define $S_1 = [-\|A - P\|, \|A - P\|]$ and see that S_1 and S_2 are indeed separated by $\delta = \gamma_n - \|A - P\|$, yielding (18) for the first case.

Now consider the case

$$\|A - P\| > \gamma_n/2 \iff 1 < 2\|A - P\|/\gamma_n. \tag{19}$$

By Theorem 4.5 (identifying S_1 as the space spanned by the top K eigenvectors of P and S_2 as the subspace spanned by the top K eigenvectors of A), we have (using Theorem

4.6)

$$\|(I - \widehat{U}\widehat{U}^T)UU^T\| = \|P_{S_1}(I - P_{S_2})\| = \|\sin\Theta\| = |\sin\theta_1| \leq 1, \tag{20}$$

where θ_1 is the first canonical angle between S_1 and S_2. Now putting together (19) and (20) yields (18) also in the second case, which combined with (17) shows (15), concluding the proof. □

4.2 Spectral bound for the noise

In the previous section we have shown that the distance defined as the sine of the largest canonical angle between the spaces spanned by the eigenvectors corresponding to the top K eigenvalues of two symmetric matrices A and P can be bounded by in terms of the distance between A and P. Thus, to bound the distance between the two subspaces it suffices to bound $\|A - P\|$. We will present two bounds, the first one being based on the Matrix Bernstein's inequality which we will state next.

Theorem 4.9 (Matrix Bernstein's inequality, Theorem 5.4.1 in [8])**.**
Let $X_1, \ldots, X_n$ be independent, mean-zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all $i \in \{1, \ldots, n\}$. Then for $\sigma^2 = \|\sum_{i=1}^n \mathbb{E}(X_i^2)\|$ and every $t > 0$ we have

$$\mathbb{P}\left(\left\{\left\|\sum_{i=1}^n X_i\right\| \geq t\right\}\right) \leq 2n\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right). \tag{21}$$

The bound obtained by using the Matrix Bernstein's inequality is stated in the next theorem.

Theorem 4.10 (See Exercise 5.5.1 in [8])**.**
Let $A = (a_{ij})_{i,j=1}^n$ be the adjacency matrix of a random graph on n nodes in which edges occur independently with expectation $\mathbb{E}[A] = P = (p_{ij})_{i,j=1}^n$. Set $d = n\max_{i,j} p_{ij}$ and assume that $d \geq c_0 \log(n)$ for a constant $c_0 > 0$. Then, for any $r > 0$ there exists a constant $C = C(r, c_0)$ such that

$$\|A - P\| \leq C\sqrt{d\log(n)}$$

with probability at least $1 - n^{-r}$.

Proof of Theorem 4.10.
Denoting the indicator function of a set A as $\mathbb{1}_A$, we define the matrices $Z^{ij} = (z^{ij}_{k\ell})^n_{k,\ell=1}$ for $1 \leq i \leq j \leq n$ by

$$z^{ij}_{k\ell} := \mathbb{1}_{\{i,j\}^2}((k,\ell))(a_{k\ell} - \mathbb{E}[a_{k\ell}]), \text{ for } 1 \leq k, \ell \leq n,$$

such that the only nonzero entries of Z^{ij} are z^{ij}_{ij} and z^{ij}_{ji} which both assume the value $a_{ij} - \mathbb{E}[a_{ij}] = a_{ji} - \mathbb{E}[a_{ji}]$ of the respective entries of the noise matrix (for $i = j$, Z^{ij} has only one non-zero entry). We can then write

$$A - P = \sum_{1 \leq i \leq j \leq n} Z_{ij}$$

as a sum over $(n^2 + n)/2$ mean-zero, independent and symmetric random matrices which satisfy $\|Z^{ij}\| \leq |a_{ij} - \mathbb{E}[a_{ij}]| \leq 1$ since the only non-zero eigenvalue of Z^{ij} is $a_{ij} - \mathbb{E}[a_{ij}]$ and $a_{ij} \in [0,1]$. Furthermore it holds by $\|M\| \leq n\|M\|_{\max}$ for $M \in \mathbb{R}^{n \times n}$, $\|M\|_{\max}$ denoting the max norm of M, that

$$\left\| \sum_{1 \leq i \leq j \leq n} \mathbb{E}\left[(Z^{ij})^2\right] \right\| \leq n \max_{i,j} p_{ij},$$

since by the nature of Z^{ij}, the two non-zero entries of $(Z^{ij})^2$ are the ones in the i-th row/j-th column and j-th row/i-th column assuming the value z^2_{ij} with expectation $p_{ij}(1 - p_{ij})$. So for each element of the sum it holds $p_{ij}(1 - p_{ij}) \leq p_{ij}$ resulting in the upper bound. Hence, we can represent $A - P$ as a sum of mean-zero, symmetric and independent random matrices Z^{ij} and therefore apply the Matrix Bernstein's inequality (21) with $K = 1$, $\sigma^2 \leq d = n \max_{i,j} p_{ij}$ and $t = \sqrt{Cd \log(n)}$, with an absolute constant

$C = C(c_0, r)$ only depending on r and c_0 but not on n, resulting in

$$\begin{aligned}
\mathbb{P}\left(\|A-P\| \le t\right) &\ge \mathbb{P}\left(\| \sum_{1\le i\le j\le n} Z^{ij}\| < t\right) \\
&\ge 1 - 2n\exp\left(-\frac{Cd\log(n)/2}{d+\sqrt{Cd\log(n)}/3}\right) \\
&\ge 1 - 2n\exp\left(-\frac{Cd\log(n)/2}{d+\sqrt{Cd\log(n)}}\right) \\
&= 1 - 2n\exp\left(-\frac{Cd\log(n)}{2(d+\sqrt{\frac{C}{c_0}dc_0\log(n)})}\right).
\end{aligned}$$

By using the assumption that $d \ge c_0\log(n)$ we can further estimate

$$\begin{aligned}
\mathbb{P}\left(\|A-P\| \le t\right) &\ge 1 - 2n\exp\left(-\frac{Cd\log(n)}{2\left(d+\sqrt{d^2\frac{C}{c_0}}\right)}\right) \\
&= 1 - 2n\exp\left(-\frac{Cd\log(n)}{2(d+d\sqrt{\frac{C}{c_0}})}\right) \\
&= 1 - 2n\exp\left(-\frac{C\log(n)}{2(1+\sqrt{\frac{C}{c_0}})}\right) \\
&= 1 - 2nn^{-C/2\left(1+\sqrt{\frac{C}{c_0}}\right)}.
\end{aligned}$$

To prove the claim we need to find $C(c_0, r)$ such that

$$1 - 2nn^{-C/2\left(1+\sqrt{\frac{C}{c_0}}\right)} \ge 1 - n^{-r}$$

which is equivalent to

$$n^{-C/2\left(1+\sqrt{\frac{C}{c_0}}\right)} \le n^{-(r+1)}/2. \tag{22}$$

Given r, c_0, choosing $C > c_0$ allows us to estimate

$$n^{-C/2\left(1+\sqrt{\frac{C}{c_0}}\right)} \leq n^{-C/4\sqrt{\frac{C}{c_0}}} = n^{-\frac{\sqrt{C}c_0}{4}}.$$

For large C the last term will now become arbitrarily small. Therefore, there exists $C > 0$ such that (22) is satisfied, completing the proof. □

However, in [5] a tighter bound for the noise is proved which requires some technical work but will also lead to a tighter bound on the fraction of misclassified vertices. We will now state the theorem.

Theorem 4.11 (Theorem 5.2 in [5])**.**
Let A be the adjacency matrix of a random graph on n nodes in which edges occur independently. Set $\mathbb{E}[A] = P = (p_{ij})_{1\leq i,j\leq n}$, $d = n \max_{i,j} p_{ij}$ and assume that $d \geq c_0 \log(n)$ for $c_0 > 0$. Then, for any $r > 0$ there exists a constant $C = C(r, c_0)$ such that

$$\|A - P\| \leq C\sqrt{d}$$

with probability at least $1 - n^{-r}$.

Remark 4.12.
In [5] the diagonal entries of the random adjacency matrix were defined to take a constant value instead of being Bernoulli-distributed such that the diagonal entries of A and P coincide. Since

$$\|A - P\| \leq \|A - P - \operatorname{diag}(A - P)\| + \|\operatorname{diag}(A - P)\| \leq \|A - P - \operatorname{diag}(A - P)\| + 1,$$

the norm of the resulting noise matrices differ at most by one implying that the bound also holds in our setting.

Comparing this to the bound obtained by applying the Matrix Bernstein's inequality in Theorem 4.10 we see that a factor of $\sqrt{\log(n)}$ can be omitted, thus substantially improving the result. The proof, however, is technically involved and thus not presented here. It can be found in the supplement of [5].

4.3 K-means approximation bound

The last step for obtaining the error bound of the spectral clustering algorithm is to reason that an output of the K-means clustering algorithm which is in the range of $(1+\varepsilon)$ times the exact solution of the K-means problem (7) still yields a good clustering of the vertices if the optimal solution yields a good clustering.

Lemma 4.13 (Lemma 5.3 in [5])**.**
For $\varepsilon > 0$ and any two matrices $\widehat{U}, U \in \mathbb{R}^{n\times K}$ such that $U = \Theta X$ with $\Theta \in \mathbb{M}^{n\times K}$ a membership matrix and $X \in \mathbb{R}^{K\times K}$, let $(\widehat{\Theta}, \widehat{X})$ be a $(1+\varepsilon)$-approximate solution to the k-means problem (7) and $\bar{U} = \widehat{\Theta}\widehat{X}$. For any $\delta_k \leq \min_{\ell\neq k} \|X_{\ell *} - X_{k*}\|$, define $S_k = \{i \in G_k(\Theta) : \|\bar{U}_{i*} - U_{i*}\| \geq \delta_k/2\}$ then

$$\sum_{k=1}^{K} |S_k|\delta_k^2 \leq 4(4+2\varepsilon)\|\widehat{U} - U\|_F^2. \tag{23}$$

Moreover, if

$$(16+8\varepsilon)\|\widehat{U} - U\|_F^2/\delta_k^2 < n_k \textit{ for all } k, \tag{24}$$

then there exists a $K \times K$ permutation matrix J such that $\widehat{\Theta}_{G*} = \Theta_{G*}J$, where $G = \bigcup_{k=1}^{K}(G_k \setminus S_k)$.

Proof of Lemma 4.13.
We estimate

$$\begin{aligned}\|\bar{U} - U\|_F^2 &\leq \Big(\|\bar{U} - \widehat{U}\|_F + \|\widehat{U} - U\|_F\Big)^2\\ &= \|\bar{U} - \widehat{U}\|_F^2 + 2\|\bar{U} - \widehat{U}\|_F\|\widehat{U} - U\|_F + \|\widehat{U} - U\|_F^2\\ &\leq 2\|\bar{U} - \widehat{U}\|_F^2 + 2\|\widehat{U} - U\|_F^2,\end{aligned} \tag{25}$$

where we used, that $a^2 + b^2 \geq 2|ab|$ for $a, b \in \mathbb{R}$ for the last inequality. Since $\bar{U} = \widehat{\Theta}\widehat{X}$ with $(\widehat{\Theta}, \widehat{X})$ being a $(1+\varepsilon)$-approximate solution to (7), i.e.

$$\|\bar{U} - \widehat{U}\|_F^2 = \|\widehat{\Theta}\widehat{X} - \widehat{U}\|_F^2 \leq (1+\varepsilon)\|\Theta X - \widehat{U}\|_F^2 = (1+\varepsilon)\|U - \widehat{U}\|_F^2,$$

(25) can be further estimated by

$$\begin{aligned}\|\bar{U} - U\|_F^2 &\leq 2(1+\varepsilon)\|U - \widehat{U}\|_F^2 + 2\|\widehat{U} - U\|_F^2 \\ &= (4+2\varepsilon)\|\widehat{U} - U\|_F^2. \end{aligned} \tag{26}$$

By equivalently defining S_k as $S_k = \{i \in G_k(\Theta) : \|\bar{U}_{i*} - U_{i*}\|_F^2 \geq \delta_k^2/4\}$ we see that

$$|S_k|\delta_k^2/4 = \sum_{i \in S_k} \frac{\delta_k^2}{4} \leq \sum_{i \in S_k} \|\bar{U}_{i*} - U_{i*}\| = \|\bar{U}_{S_k*} - U_{S_k*}\|_F^2.$$

Therefore, (using (26) for the last inequality) we obtain

$$\sum_{k=1}^{K} |S_k|\delta_k^2/4 \leq \sum_{k=1}^{K} \|\bar{U}_{S_k*} - U_{S_k*}\|_F^2 \leq \|\bar{U} - U\|_F^2 \leq (4+2\varepsilon)\|\widehat{U} - \Theta X\|_F^2,$$

which is equivalent to

$$\sum_{k=1}^{K} |S_k|\delta_k^2 \leq 4(4+2\varepsilon)\|\widehat{U} - \Theta X\|_F^2, \tag{27}$$

proving the first part of the lemma. For the second part we first conclude from (27) that also

$$|S_k| \leq (16+8\varepsilon)\|\widehat{U} - \Theta X\|_F^2/\delta_k^2 \text{ for all } k.$$

Combining this with (24), we have that $|S_k| < n_k$ for all k. Now define $T_k \coloneqq \{i \in G_k(\Theta) : \|\bar{U}_{i*} - U_{i*}\| < \delta_k/2\} = G_k \setminus S_k$ for $k = 1, \ldots, K$. Since $|S_k| < n_k$ these are nonempty sets. We will now show

$$\bar{U}_{i*} = \bar{U}_{j*} \iff i, j \in T_k \text{ for some } k \in \{1, \ldots, K\}. \tag{28}$$

First consider $i \in T_k$ and $j \in T_\ell$ for $k \neq \ell$. Assuming $\bar{U}_{i*} \neq \bar{U}_{j*}$ would not hold, then

$$\begin{aligned}\max(\delta_k, \delta_\ell) &\leq \|X_{k*} - X_{\ell*}\| = \|U_{i*} - U_{j*}\| \\ &\leq \|U_{i*} - \bar{U}_{i*}\| + \|\bar{U}_{i*} - \bar{U}_{j*}\| + \|\bar{U}_{j*} - U_{j*}\| \\ &= \|U_{i*} - \bar{U}_{i*}\| + \|\bar{U}_{j*} - U_{j*}\| \\ &< \delta_k/2 + \delta_\ell/2,\end{aligned}$$

where we used the assumption on δ_k, the fact that $U_{i*} = X_{g_i*} = X_k*$ and $\bar{U}_{i*} = \bar{U}_{j*}$ as well as the definition of T_k. Since

$$\max(\delta_k, \delta_l) < \delta_k/2 + \delta_l/2$$

is a contradiction the assumption cannot be true, thus showing

$$\bar{U}_{i*} \neq \bar{U}_{j*}.$$

Now consider $i, j \in T_k$ with $i \neq j$ (the case $i = j$ follows immediately). If $\bar{U}_{i*} = \bar{U}_{j*}$ would not hold, $\bar{U}$ had more than K mutually different rows since for each k, $T_k \neq \emptyset$ and therefore by above argumentation, $\bar{U}$ already has at least K different rows corresponding to the elements in T_k for $k = 1, \ldots, K$. However, by construction each row of $\bar{U}$ corresponds to one row of $\widehat{X} \in \mathbb{R}^{K \times K}$ and thus the maximal number of different rows is K. Hence, we get that

$$\bar{U}_{i*} - \bar{U}_{j*} \text{ for } i, j \in T_k,$$

concluding the proof of (28). Writing

$$G = \bigcup_{k=1}^{K} T_k = \bigcup_{k=1}^{K} (G_k \setminus S_k),$$

we get that, if the rows i and j of $\bar{U}_{G*} = \left(\widehat{\Theta}\widehat{X}\right)_{G*}$ are the same, this implies that the vertices i and j belong to the same community. Hence the community assignment of $\widehat{\Theta}_{G*}$ is correct up to the labeling of the communities and there exists a (column) permutation matrix $J \in E_K$ such that $\widehat{\Theta}_{G*} = \Theta_{G*}J$ which concludes the proof. □

The sets S_k are defined as the vertices in community k for which the distance of the corresponding row of the matrix $\bar{U}$ that is obtained as a $(1+\varepsilon)$-approximate solution to the K-means problem and the respective row of U, which contains the first K eigenvectors of the expected adjacency matrix P and thus leads to a perfect clustering, exceeds a threshold. Under (24), (23) thus provides a bound on the sum of misclassified vertices which depends on the distance of the matrix $\widehat{U}$, which is obtained by solving the K-means-problem optimally, to U, the approximation accuracy ε and the thresholds δ_k. A higher accuracy, i.e. lower ε, leads to a lower bound while higher thresholds lead to a higher bound since they allow the rows of $\bar{U}$, which are used to cluster the vertices, to deviate a lot from the rows of U which provide a perfect clustering.

4.4 Proof of the error bound

Proof of Theorem 2.12.
Let $\widehat{U}, U$ be the matrices containing the eigenvectors with respect to the top K eigenvalues of A and P, respectively. By Lemma 4.8 there exists an orthonormal matrix $Q \in \mathbb{R}^{K\times K}$ such that

$$\left\|\widehat{U} - UQ\right\|_F \leq \frac{2\sqrt{2K}}{\gamma_n} \left\|A - P\right\|. \tag{29}$$

Now we can apply Theorem 4.11 to $\|A-P\|$ with $r = 1, c_0 = 1$ and $d = n \max_{i,j} p_{i,j} = n\alpha$ to estimate

$$\|A - P\| \leq C\sqrt{n\alpha} \tag{30}$$

with probability at least $1 - n^{-1}$ for an absolute constant C. Combining (29) with (30) yields that

$$\left\|\widehat{U} - UQ\right\|_F \leq \frac{2\sqrt{2K}}{\gamma_n} C\sqrt{n\alpha} \tag{31}$$

with probability at least $1 - n^{-1}$. By Lemma 2.9 we can write $U = \Theta X$, with $X \in \mathbb{R}^{K\times K}$ satisfying

$$\forall 1 \leq k < \ell \leq K := \|X_{k*} - X_{\ell *}\| = \sqrt{\frac{1}{n_k} + \frac{1}{n_\ell}}.$$

Therefore we can also write

$$UQ = \Theta XQ = \Theta X',$$

with $X' \in \mathbb{R}^{K\times K}$, by orthonormality of Q, having the property that

$$\forall 1 \leq k < \ell \leq K : \|X'_{k*} - X'_{\ell *}\| = \|X_{k*}Q - X_{\ell *}Q\| = \|X_{k*} - X_{\ell *}\| = \sqrt{\frac{1}{n_k} + \frac{1}{n_\ell}}.$$

Now we want to apply Lemma 4.13 to $\widehat{U}$ and UQ and therefore choose

$$\delta_k \coloneqq \sqrt{\frac{1}{n_k} + \frac{1}{\max\{n_\ell : \ell \neq k\}}} = \min_{\ell \neq k} \|X'_{k*} - X'_{\ell *}\|, k = 1, \ldots, K,$$

satisfying

$$n_k \delta_k^2 > 1.$$

Next, we need condition (24) to hold and will show

$$(16 + 8\varepsilon)\|\widehat{U} - UQ\|_F^2 \leq 1 < n_k \delta_k^2,$$

which then implies (24). Using (31), we can estimate

$$(16 + 8\varepsilon)\|\widehat{U} - UQ\|_F^2 \leq (16 + 8\varepsilon)\left(\frac{2\sqrt{2K}}{\gamma_n} C\sqrt{n\alpha}\right)^2 = 64C^2(2+\varepsilon)\frac{Kn\alpha}{\gamma_n^2} \tag{32}$$

which is less than one if we set $c = \frac{1}{64C^2}$ in (8).

Therefore (24) holds with probability at least $1 - n^{-1}$ and we can apply Lemma 4.13 yielding that with probability $1 - n^{-1}$, there exist subsets $S_k \subset G_k$ and a $K \times K$ permutation matrix J such that $\widehat{\Theta}_{G*}J = \Theta_{G*}$ for $G = \bigcup_{k=1}^K (G_k \setminus S_k)$ and

$$\sum_{k=1}^K \frac{|S_k|}{n_k} \leq \sum_{k=1}^K |S_k|\delta_k^2 \leq 4(4 + 2\varepsilon)\left\|\widehat{U} - UQ\right\|_F^2,$$

which by (32) and with $c = \frac{1}{64C^2}$ implies that

$$\begin{aligned}\sum_{k=1}^{K} \frac{|S_k|}{n_k} \leq \sum_{k=1}^{K} |S_k|\delta_k^2 \leq 64C^2(2+\varepsilon)\frac{Kn\alpha}{\gamma_n^2} \\ = \frac{2+\varepsilon}{c}\frac{Kn\alpha}{\gamma_n^2},\end{aligned} \tag{33}$$

holds with probability $1 - n^{-1}$, concluding the proof. □

Proof of Corollary 2.14.

From the proof of Lemma 2.9, in particular (6), we see that the non-zero eigenvalues of $P = \Theta B_\alpha \Theta^T$ coincide with the eigenvalues of $\Delta\alpha B\Delta$. Since Δ and B are symmetric, their absolute eigenvalues correspond to their singular values. Therefore, the minimum absolute eigenvalue γ_n of P can be bounded below by

$$\gamma_n = \alpha\lambda_{\min}^{\Delta B\Delta} \geq \lambda_{\min}^{\Delta}\alpha\lambda\lambda_{\min}^{\Delta} = \alpha\sqrt{n_{\min}}\lambda\sqrt{n_{\min}} = n_{\min}\lambda\alpha,$$

where $\lambda_{\min}^{\Delta}$, the minimum absolute eigenvalue of Δ corresponds to the square root of the smallest community size $n_{\min}$ and we used the assumption that the minimum eigenvalue of B is greater than λ. Therefore, (33) in the proof of Theorem 2.12 applies with $\gamma_n = n_{\min}\lambda\alpha$, implying

$$\sum_{k=1}^{K} |S_k| \left(\frac{1}{n_k} + \frac{1}{max\{n_\ell : \ell \neq k\}}\right) \leq 64C^2(2+\varepsilon)\frac{Kn}{n_{\min}^2\lambda^2\alpha}. \tag{34}$$

By the definition of $\tilde{L}\left(\widehat{\Theta}, \Theta\right)$ being the maximal fraction of misclassified vertices throughout all communities we can estimate

$$\tilde{L}\left(\widehat{\Theta}, \Theta\right) \leq \max_{1\leq k\leq K} \frac{|S_k|}{n_k} \leq \sum_{k=1}^{K} \frac{|S_k|}{n_k} \leq 64C^2(2+\varepsilon)\frac{Kn}{n_{\min}^2\lambda^2\alpha},$$

where the last inequality follows from (34) and holds with probability at least $1 - n^{-1}$. Again, by definition of S_k - being the set of vertices in community k for which a correct clustering cannot be guaranteed - and $L\left(\widehat{\Theta}, \Theta\right)$, we can estimate

$$L\left(\widehat{\Theta}, \Theta\right) \leq \frac{1}{n}\sum_{k=1}^{K} |S_k|. \tag{35}$$

Recalling that $n'_{\max}$ is the second largest community size we get that

$$1 < n'_{\max}\left(\frac{1}{n_k} + \frac{1}{\max\{n_\ell : \ell \neq k\}}\right)$$

for all $k = 1, \ldots, K$ and combining this with (35) implies

$$L\left(\widehat{\Theta}, \Theta\right) \leq n'_{\max}\frac{1}{n}\sum_{k=1}^{K}|S_k|\left(\frac{1}{n_k} + \frac{1}{\max\{n_\ell : \ell \neq k\}}\right)$$

which, using estimate (34) yields

$$L\left(\widehat{\Theta}, \Theta\right) \leq 64C^2(2+\varepsilon)\frac{Kn'_{\max}}{n^2_{\min}\lambda^2\alpha}$$

with probability at least $1 - n^{-1}$ and therefore concludes the proof. □

A Appendix

A.1 Eigenstructure of a SBM with two communities

By equation (5) in the proof of Lemma 2.9 we know that the non-zero eigenvalues of P are exactly the eigenvalues of $\Delta B \Delta$, where in our case $\Delta = \begin{bmatrix} \sqrt{\frac{n}{2}} & 0 \\ 0 & \sqrt{\frac{n}{2}} \end{bmatrix}$ and $B = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$. Hence, $\Delta B \Delta = \frac{n}{2} B$ and we compute

$$\det(\Delta B \Delta - \lambda I_2) = \left(\frac{pn}{2} - \lambda\right)^2 - \left(\frac{qn}{2}\right)^2$$

which is equal to zero for

$$\lambda = \frac{n}{2}(p+q) \vee \lambda = \frac{n}{2}(p-q). \tag{36}$$

To obtain the corresponding eigenvectors we compute the kernel of $P - \lambda I_n$. Defining $p' = p - \lambda$ and using the representation of P from Example 2.8, we get

$$P - \lambda I_n = \begin{array}{r|cccccccc|}
I & p' & p & \cdots & p & q & \cdots & \cdots & q \\
II & p & p' & p \cdots & p & q & q & \cdots & q \\
\vdots & \vdots & & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & p & \ddots & p & \vdots & \vdots & & \vdots \\
(*) & q & q & \cdots q & p' & p & p & \cdots & p \\
III & q & \cdots & \cdots & q & p' & p & \cdots & p \\
IV & p & p & \cdots & p & q & p' & q \cdots & q \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & p & \ddots & \vdots \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & p \\
(**) & q & q & \cdots & q & p & p & \cdots p & p'
\end{array}.$$

Subtracting row I from each of the rows $II, \dots, (*)$ and row III from each of the rows $IV, \dots, (**)$ leads to

$$\rightsquigarrow \begin{array}{r|cccccccc|}
I & p' & p & \cdots & p & q & q & \cdots & q \\
II & \lambda & -\lambda & 0 \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & 0 & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\
(*) & \lambda & 0 & \cdots 0 & -\lambda & 0 & 0 & \cdots & 0 \\
III & q & q & \cdots & q & p' & p & \cdots & p \\
IV & 0 & 0 & \cdots & 0 & \lambda & -\lambda & 0 \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & 0 & \ddots & 0 \\
(**) & 0 & 0 & \cdots & 0 & \lambda & 0 & \cdots 0 & -\lambda
\end{array}.$$

Now, subtracting each of the rows $II, \ldots, (*)$ from row I, each of the rows $IV, \ldots, (**)$ from row III and dividing rows $II, \ldots, (*)$ and $IV, \ldots, (**)$ by λ (noting that $\lambda \neq 0$ since we assumed B to have full rank, i.e. $p \neq q$ in (36)) leads to

$$\rightsquigarrow \begin{bmatrix}
\frac{np}{2} - \lambda & 0 & \cdots & 0 & \frac{nq}{2} & 0 & \cdots & 0 \\
1 & -1 & 0 \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & 0 & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \cdots 0 & -1 & 0 & 0 & \cdots & 0 \\
\frac{nq}{2} & 0 & \cdots & 0 & \frac{np}{2} - \lambda & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 1 & -1 & 0 \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & 0 & \ddots & 0 \\
0 & 0 & \cdots & 0 & 1 & 0 & \cdots 0 & -1
\end{bmatrix}. \tag{37}$$

The kernel of (37) corresponds to the span of

$$v_1 = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1 \end{bmatrix}, \text{ for } \lambda = \frac{n}{2}(p+q) \text{ and } v_2 = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{array}\right], \text{ for } \lambda = \frac{n}{2}(p-q),$$

showing that the eigenvectors of $P = \mathbb{E}[A]$ are of the claimed shape.

A.2 Code for conducting the simulation

The simulation was implemented in *MATLAB* by creating the function *sparsity_error()* given in Listing 1 and then calling it for the different total numbers of vertices n and values of q and plotting the results by the script which is given in Listing 2.

Listing 1: The function *sparsity_error()*.

```
function mean_errors=sparsity_error(alphas, n1, n2, q, samplesize)
% Computes mean error for a stochastic block model
%           with 2 communities and connectivity matrix
%           alpha*B, where B=[1 q;
%                             q 1]
%           for values of alpha given as function input. For each
%           alpha a sample of sample size specified as input
%           is drawn from the model.
%
%     Input parameters
%       alphas - vector of scaling parameters alpha
%       n1 - size of the first community
%       n2 - size of the second community
%       q - edge probability for vertices of different communities in
%           the unscaled connectivity matrix
%       samplesize - number of adjacency matrices that are sampled for
%           each scaling parameter
%
%     Output value
%       mean_errors - an array containing the mean fraction of
%           misclassified vertices of the sampled adjacency matrices
%           for each scaling parameter in alphas
%
    %% Set parameters
    % random number generator for reproducible results
    rng('default');
    % compute total number of vertices
    n = n1 + n2;

    % vector for community assignment
    Assignment = zeros(n,1);
    % first n1 vertices are assigned to community 1
    Assignment(1:n1) = 1;
    % vertices n1+1 through n are assigned to community 2
    Assignment(n1+1:n) = 2;

    % connectivity matrix B containing the inter-community edge
    % probabilities
    B= [1 q;
        q 1];

    % vector collecting the relative clustering errors
    % one row consists of the sample outcomes for one alpha
    relative_errors_total = zeros(length(alphas),samplesize);


    % counter for indexing
    counter = 0;

    % go through vector of scaling parameters
    for alpha=alphas

        counter = counter + 1;

        % scale B by the scaling parameter alpha
        B_scaled = alpha*B;

        % vector collecting the relative errors for each sampled
        % adjacency matrix
        rel_errors_alpha = zeros(1, samplesize);

        % sampling of adjacency matrix and error computation
        for sample_index=1:samplesize
            %% sample adjacency matrix
            % initialize an nxn matrix
            A=zeros(n,n);

            % sample entries on and above the diagonal as Bernoulli(p)-
            % distributed random variables with p given by the
            % connectivity matrix B
            for i=1:n
                for j=i:n
                    A(i,j) = random('Binomial',1, ...
                        B_scaled(Assignment(i),Assignment(j)));

                    % set entries below the diagonal such that a
                    % symmetric matrix is obtained
                    if i ~= j
                        A(j,i) = A(i,j);
                    end
                end
            end

            %% Run spectral clustering algorithm
            % compute 2 largest absolute eigenvalues and corresponding
            % eigenvectors
            [V,D] = eigs(A,2);

            % classify vertices based on k-means applied to the matrix
            % with columns consisting of the 2 eigenvectors
            classification = kmeans(V,2);

            %% error is smallest errors over possible labelings
            % labeling 1
            error_vector_1 = (Assignment == classification);

            % labeling 2
            error_vector_2 = (Assignment ~= classification);

            % absolute error is minimum of absolute errors
            % over 2 possible labelings
            abs_error = min(sum(error_vector_1), sum(error_vector_2));

            % relative error as fraction of absolute misclassified
            % vertices
            rel_error = abs_error/n;

            % add error for sampled adjacency matrix to vector
            % for current alpha
            rel_errors_alpha(sample_index) = rel_error;
        end
        % add relative errors obtained for current alpha to matrix
        % collecting the errors for all alphas
        relative_errors_total(counter,:) = rel_errors_alpha;

    end

    % compute the mean fraction of misclassified vertices
    % over all sampled adjacency matrices
    mean_errors = mean(relative_errors_total,2);
end
```

Listing 2: Script for conducting the simulation.

```
% vector of scaling parameters
alphas = (1:100)/100;
% edge probabilities for vertices of different communities in the unscaled connectivity matrix
q_s = [0.4 0.8];

% number of sampled adjacency matrices for each alpha
samplesize = 1000;

% size of community 1 and community 2 for n=50
[n1, n2] = deal(25, 25);
%compute mean error for n=50
mean_50_q_04 = sparsity_error(alphas, n1, n2, q_s(1), samplesize);
mean_50_q_08 = sparsity_error(alphas, n1, n2, q_s(2), samplesize);

% size of community 1 and community 2 for n=100
[n1, n2] = deal(50, 50);
%compute mean error for n=100
mean_100_q_04 = sparsity_error(alphas, n1, n2, q_s(1), samplesize);
mean_100_q_08 = sparsity_error(alphas, n1, n2, q_s(2), samplesize);

% size of community 1 and community 2 for n=150
[n1, n2] = deal(75, 75);
%compute mean error for n=150
mean_150_q_04 = sparsity_error(alphas, n1, n2, q_s(1), samplesize);
mean_150_q_08 = sparsity_error(alphas, n1, n2, q_s(2), samplesize);

% plot the mean errors for the different n's
plot(alphas, [mean_50_q_04, mean_100_q_04, mean_150_q_04,...
              mean_50_q_08, mean_100_q_08, mean_150_q_08],'LineWidth', 1)

legend('n=50, q=0.4', 'n=100, q=0.4', 'n=150, q=0.4',...
       'n=50, q=0.8', 'n=100, q=0.8', 'n=150, q=0.8');
xlabel('\alpha','fontsize',14);
ylabel('$\overline{L(\widehat{\Theta},\Theta)}$','fontsize',14,'interpreter','latex');
set(gcf,'Units','Inches');
pos = get(gcf,'Position');
set(gcf,'PaperPositionMode','Auto','PaperUnits','Inches','PaperSize',[pos(3), pos(4)]);
saveas(gcf, 'mean_error.pdf');
```

References

[1] R. Bhatia. *Matrix Analysis.* New York: Springer New York, 1997.

[2] S. Fortunato and D. Hric. „Community Detection in Networks: A User Guide“. *Physics Reports* 659 (2016), pp. 1–44.

[3] J. Friedman, R. Tibshirani, and T. Hastie. *The Elements of Statistical Learning.* New York: Springer New York, 2001.

[4] A. Kumar, Y. Sabharwal, and S. Sen. „A Simple Linear Time $(1 + \varepsilon)$-approximation Algorithm for k-means Clustering in any Dimensions“. *45th Annual IEEE Symposium on Foundations of Computer Science.* 2004, pp. 454–462.

[5] J. Lei and A. Rinaldo. „Consistency of Spectral Clustering in Stochastic Block Models“. *The Annals of Statistics* 43.1 (2015), pp. 215–237.

[6] U. von Luxburg. „A Tutorial on Spectral Clustering“. *Statistics and Computing* 17.4 (2007), pp. 395–416.

[7] G. W. Stewart and J. Sun. *Matrix Perturbation Theory.* Boston: Academic Press, 1990.

[8] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge: Cambridge University Press, 2018.

[9] V. Q. Vu and J. Lei. „Minimax Sparse Principal Subspace Estimation in High Dimensions“. *The Annals of Statistics* 41.6 (2013), pp. 2905–2947.