# Appendix

## A. Adding Fairness Constraints to Normalized Spectral Clustering

In this section we derive a fair version of normalized spectral clustering (similarly to how we proceeded for unnormalized spectral clustering in Sections 2 and 3 of the main paper).

Normalized spectral clustering aims at partitioning $V$ into $k$ clusters with minimum value of the NCut objective function as follows (see von Luxburg, 2007, for details): for a clustering $V = C_1 \dot\cup \ldots \dot\cup C_k$ we have

$$\mathrm{NCut}(C_1, \ldots, C_k) = \sum_{l=1}^{k} \frac{\mathrm{Cut}(C_l, V \setminus C_l)}{\mathrm{vol}(C_l)}, \tag{15}$$

where $\mathrm{vol}(C_l) = \sum_{i \in C_l} d_i = \sum_{i \in C_l, j \in [n]} W_{ij}$. Encoding a clustering $V = C_1 \dot\cup \ldots \dot\cup C_k$ by a matrix $H \in \mathbb{R}^{n \times k}$ with

$$H_{il} = \begin{cases} 1/\sqrt{\mathrm{vol}(C_l)}, & i \in C_l, \\ 0, & i \notin C_l \end{cases}, \tag{16}$$

we have $\mathrm{NCut}(C_1, \ldots, C_k) = \mathrm{Tr}(H^T L H)$. Note that any $H$ of the form (16) satisfies $H^T D H = I_k$. Normalized spectral clustering relaxes the problem of minimizing $\mathrm{Tr}(H^T L H)$ over all $H$ of the form (16) to

$$\min_{H \in \mathbb{R}^{n \times k}} \mathrm{Tr}(H^T L H) \text{ subject to } H^T D H = I_k. \tag{17}$$

Substituting $H = D^{-1/2} T$ for $T \in \mathbb{R}^{n \times k}$ (we need to assume that $G$ does not contain any isolated vertices since otherwise $D^{-1/2}$ does not exist), problem (17) becomes

$$\min_{T \in \mathbb{R}^{n \times k}} \mathrm{Tr}(T^T D^{-1/2} L D^{-1/2} T) \text{ subject to } T^T T = I_k.$$

Similarly to unnormalized spectral clustering, normalized spectral clustering computes an optimal $T$ by computing the $k$ smallest eigenvalues and some corresponding eigenvectors of $D^{-1/2} L D^{-1/2}$ and applies $k$-means clustering to the rows of $H = D^{-1/2} T$ (in practice, $H$ can be computed directly by solving the generalized eigenproblem $Lx = \lambda D x$, $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$; see von Luxburg, 2007).

Now we want to derive our fair version of normalized spectral clustering. The first step is to show that Lemma 1 holds true if we encode a clustering as in (16):

**Lemma 2** (Fairness constraint as linear constraint on $H$ for normalized spectral clustering)**.** *For $s \in [h]$, let $f^{(s)} \in \{0,1\}^n$ be the group-membership vector of $V_s$, that is $f_i^{(s)} = 1$ if $i \in V_s$ and $f_i^{(s)} = 0$ otherwise. Let $V = C_1 \dot\cup \ldots \dot\cup C_k$ be a clustering that is encoded as in (16). We have, for every $l \in [k]$,*

$$\forall s \in [h-1] : \sum_{i=1}^{n} \left( f_i^{(s)} - \frac{|V_s|}{n} \right) H_{il} = 0 \quad \Leftrightarrow \quad \forall s \in [h] : \frac{|V_s \cap C_l|}{|C_l|} = \frac{|V_s|}{n}.$$

*Proof.* This simply follows from

$$\sum_{i=1}^{n} \left( f_i^{(s)} - \frac{|V_s|}{n} \right) H_{il} = \frac{|V_s \cap C_l|}{\sqrt{\mathrm{vol}(C_l)}} - \frac{|V_s| \cdot |C_l|}{n\sqrt{\mathrm{vol}(C_l)}}$$

and $|C_l| = \sum_{s=1}^{h} |V_s \cap C_l|$. □

Lemma 2 suggests that in a fair version of normalized spectral clustering, rather than solving (17), we should solve

$$\min_{H \in \mathbb{R}^{n \times k}} \mathrm{Tr}(H^T L H) \text{ subject to } H^T D H = I_k \text{ and } F^T H = 0_{(h-1) \times k}, \tag{18}$$

**Algorithm 3** Normalized SC with fairness constraints

**Input:** weighted adjacency matrix $W \in \mathbb{R}^{n\times n}$ **(the underlying graph must not contain any isolated vertices)**; $k \in \mathbb{N}$; group-membership vectors $f^{(s)} \in \{0,1\}^n$, $s \in [h]$

**Output:** a clustering of $[n]$ into $k$ clusters

- compute the Laplacian matrix $L = D - W$ with the degree matrix $D$
- build the matrix $F$ that has the vectors $f^{(s)} - \frac{|V_s|}{n} \cdot \mathbf{1}_n$, $s \in [h-1]$, as columns
- compute a matrix $Z$ whose columns form an orthonormal basis of the nullspace of $F^T$
- compute the square root $Q$ of $Z^T D Z$
- compute some orthonormal eigenvectors corresponding to the $k$ smallest eigenvalues (respecting multiplicities) of $Q^{-1}Z^T L Z Q^{-1}$
- let $X$ be a matrix containing these eigenvectors as columns
- apply $k$-means clustering to the rows of $H = ZQ^{-1}X \in \mathbb{R}^{n\times k}$, which yields a clustering of $[n]$ into $k$ clusters

where $F \in \mathbb{R}^{n\times(h-1)}$ is the matrix that has the vectors $f^{(s)} - (|V_s|/n) \cdot \mathbf{1}_n$, $s \in [h-1]$, as columns (just as in Section 3). It is $\operatorname{rank}(F) = \operatorname{rank}(F^T) = h - 1$ and we need to assume that $k \leq n - h + 1$ since otherwise (18) does not have any solution. Let $Z \in \mathbb{R}^{n\times(n-h+1)}$ be a matrix whose columns form an orthonormal basis of the nullspace of $F^T$. We substitute $H = ZY$ for $Y \in \mathbb{R}^{(n-h+1)\times k}$, and then problem (18) becomes

$$\min_{Y\in\mathbb{R}^{(n-h+1)\times k}} \operatorname{Tr}(Y^T Z^T L Z Y) \text{ subject to } Y^T Z^T D Z Y = I_k. \tag{19}$$

Assuming that $G$ does not contain any isolated vertices, $Z^T D Z$ is positive definite and hence has a positive definite square root, that is there exists a positive definite $Q \in \mathbb{R}^{(n-h+1)\times(n-h+1)}$ with $Z^T D Z = Q^2$. We can substitute $Y = Q^{-1}X$ for $X \in \mathbb{R}^{(n-h+1)\times k}$, and then problem (19) becomes

$$\min_{X\in\mathbb{R}^{(n-h+1)\times k}} \operatorname{Tr}(X^T Q^{-1} Z^T L Z Q^{-1} X) \text{ subject to } X^T X = I_k. \tag{20}$$

A solution to (20) is given by a matrix $X$ that contains some orthonormal eigenvectors corresponding to the $k$ smallest eigenvalues (respecting multiplicities) of $Q^{-1}Z^T L Z Q^{-1}$ as columns. This gives rise to our fair version of normalized spectral clustering as stated in Algorithm 3.

## B. Computational Complexity of our Algorithms

The costs of standard spectral clustering (e.g., Algorithm 1) are dominated by the complexity of the eigenvector computations and are commonly stated to be, in general, in $\mathcal{O}(n^3)$ regarding time and $\mathcal{O}(n^2)$ regarding space for an arbitrary number of clusters $k$, unless approximations are applied (Yan et al., 2009; Li et al., 2011). In addition to the computations performed in Algorithm 1, in Algorithm 2 and Algorithm 3 we have to compute an orthonormal basis of the nullspace of $F^T$, perform some matrix multiplications, and (only for Algorithm 3) compute the square root of an $(n-h+1)\times(n-h+1)$-matrix and the inverse of this square root. All these computations can be done in $\mathcal{O}(n^3)$ regarding time and $\mathcal{O}(n^2)$ regarding space (an orthonormal basis of the nullspace of $F^T$ can be computed by means of an SVD; see, e.g., Golub & Van Loan, 2013), and hence our algorithms have the same worst-case complexity as standard spectral clustering. On the other hand, if the graph $G$, and thus the Laplacian matrix $L$, is sparse or $k$ is small, then the eigenvector computations in Algorithm 1 can be done more efficiently than with cubic running time (Bai et al., 2000). This is not the case for our algorithms as stated. However, one could apply one of the techniques suggested in the existing literature on constrained spectral clustering to speed up computation (e.g., Yu & Shi, 2004, or Xu et al., 2009; see Section 5 of the main paper). With the implementations as stated, in our experiments in Section 6 of the main paper we observe that Algorithm 2 has a similar running time as standard normalized spectral clustering while the running time of Algorithm 3 is significantly higher.

## C. Proof of Theorem 1

We split the proof of Theorem 1 into four parts. In the first part, we analyze the eigenvalues and eigenvectors of the expected adjacency matrix $\mathcal{W}$ and of the matrix $Z^T\mathcal{L}Z$, where $\mathcal{L}$ is the expected Laplacian matrix and $Z$ is the matrix computed in the execution of Algorithm 2 or Algorithm 3. In the second part, we study the deviation of the observed matrix $Z^TLZ$ from the expected matrix $Z^T\mathcal{L}Z$. In the third part, we use the results from the first and the second part to prove Theorem 1 for Algorithm 2 (unnormalized SC with fairness constraints). In the fourth part, we prove Theorem 1 for Algorithm 3 (normalized SC with fairness constraints).

**Notation** For $x \in \mathbb{R}^n$, by $\|x\|$ we denote the Euclidean norm of $x$, that is $\|x\| = \sqrt{x_1^2 + \ldots + x_n^2}$. For $A \in \mathbb{R}^{n\times m}$, by $\|A\|$ we denote the operator norm (also known as spectral norm) and by $\|A\|_F$ the Frobenius norm of $A$. It is

$$\|A\| = \max_{x\in\mathbb{R}^m:\|x\|=1} \|Ax\| = \sqrt{\lambda_{\max}(A^TA)}, \tag{21}$$

where $\lambda_{\max}(A^TA)$ is the largest eigenvalue of $A^TA$, and

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2} = \sqrt{\mathrm{Tr}(A^TA)}. \tag{22}$$

Note that for a symmetric matrix $A \in \mathbb{R}^{n\times n}$ with $A = A^T$ we have $\|A\| = \max\{|\lambda_i| : \lambda_i \text{ is an eigenvalue of } A\}$. It follows from (21) and (22) that for any $A \in \mathbb{R}^{n\times m}$ with rank at most $r$ we have

$$\|A\| \le \|A\|_F \le \sqrt{r}\|A\|. \tag{23}$$

We use $\mathrm{const}(X)$ to denote a universal constant that only depends on $X$ and that may change from line to line.

### Part 1: Eigenvalues and eigenvectors of $\mathcal{W}$ and of $Z^T\mathcal{L}Z$

Assuming the $n$ vertices $1,\ldots,n$ are sorted in a way such that

$$\begin{aligned}
1,\ldots,\frac{n}{kh} \in C_1\cap V_1, \quad \frac{n}{kh}+1,\ldots,\frac{2n}{kh} \in C_1\cap V_2, \quad \ldots\ \ldots\ \ldots, \quad \frac{(h-1)n}{kh}+1,\ldots,\frac{n}{k} \in C_1\cap V_h,\\
\frac{n}{k}+1,\ldots,\frac{n}{k}+\frac{n}{kh} \in C_2\cap V_1, \quad \ldots\ \ldots\ \ldots, \quad \frac{n}{k}+\frac{(h-1)n}{kh}+1,\ldots,\frac{2n}{k} \in C_2\cap V_h,\\
\vdots\\
\frac{(k-1)n}{k}+1,\ldots,\frac{(k-1)n}{k}+\frac{n}{kh} \in C_k\cap V_1, \quad \ldots\ \ldots\ \ldots, \quad \frac{(k-1)n}{k}+\frac{(h-1)n}{kh}+1,\ldots,n \in C_k\cap V_h,
\end{aligned} \tag{24}$$

the expected adjacency matrix $\mathcal{W} \in \mathbb{R}^{n\times n}$ is given by the block matrix

$$\mathcal{W} = \underbrace{\begin{pmatrix}
[R] & [S] & [S] & [S] & \cdots & [S] & [S]\\
[S] & [R] & [S] & [S] & \cdots & [S] & [S]\\
[S] & [S] & [R] & [S] & \cdots & [S] & [S]\\
 & \vdots & & \ddots & & \vdots & \\
[S] & [S] & [S] & [S] & \cdots & [S] & [R]
\end{pmatrix}}_{=:\widetilde{\mathcal{W}}} - aI_n, \tag{25}$$

where $[R] \in \{a,c\}^{\frac{n}{k}\times\frac{n}{k}}$ and $[S] \in \{b,d\}^{\frac{n}{k}\times\frac{n}{k}}$ are themselves block matrices

$$[R] = \begin{pmatrix} [a] & [c] & [c] & [c] & \cdots & [c] & [c] \\ [c] & [a] & [c] & [c] & \cdots & [c] & [c] \\ [c] & [c] & [a] & [c] & \cdots & [c] & [c] \\ & \vdots & & \ddots & & \vdots & \\ [c] & [c] & [c] & [c] & \cdots & [c] & [a] \end{pmatrix}, \qquad [S] = \begin{pmatrix} [b] & [d] & [d] & [d] & \cdots & [d] & [d] \\ [d] & [b] & [d] & [d] & \cdots & [d] & [d] \\ [d] & [d] & [b] & [d] & \cdots & [d] & [d] \\ & \vdots & & \ddots & & \vdots & \\ [d] & [d] & [d] & [d] & \cdots & [d] & [b] \end{pmatrix}$$

with $[a]$, $[b]$, $[c]$ and $[d]$ being matrices of size $\frac{n}{kh} \times \frac{n}{kh}$ with all entries equaling $a$, $b$, $c$ and $d$, respectively. We denote the matrix $\mathcal{W} + aI_n$ with $\mathcal{W}$ as in (25) by $\widetilde{\mathcal{W}}$. If the vertices are not ordered as in (24), the expected adjacency matrix $\mathcal{W}$ is rather given by $\mathcal{W} = P^T\widetilde{\mathcal{W}}P - aI_n$ for some permutation matrix $P \in \{0,1\}^{n\times n}$ with $PP^T = P^TP = I_n$. Note that $v \in \mathbb{R}^n$ is an eigenvector of $\widetilde{\mathcal{W}}$ with eigenvalue $\lambda$ if and only if $P^Tv$ is an eigenvector of $P^T\widetilde{\mathcal{W}}P$ with eigenvalue $\lambda$. Keeping track of the permutation matrices $P$ and $P^T$ throughout the proof of Theorem 1 does not impose any technical challenges, but makes the writing more complicated. Hence, for simplicity and without loss of generality, we assume in the following that the vertices are ordered as in (24).

The following lemma characterizes the eigenvalues and eigenvectors of $\widetilde{\mathcal{W}}$. Clearly, this also characterizes the eigenvalues and eigenvectors of $\mathcal{W}$: $v \in \mathbb{R}^n$ is an eigenvector of $\widetilde{\mathcal{W}}$ with eigenvalue $\lambda$ if and only if $v$ is an eigenvector of $\mathcal{W}$ with eigenvalue $\lambda - a$.

**Lemma 3.** *Assuming that $a > b > c > d \geq 0$, the matrix $\widetilde{\mathcal{W}}$ has rank $kh$ or rank $k+h-1$ (the latter is true if and only if $a - c = b - d$). It has the following eigenvalues $\lambda_1, \ldots, \lambda_n$:*

$$\begin{aligned} \lambda_1 &= \frac{n}{kh}\left[\big(a + (h-1)c\big) + (k-1)\big(b + (h-1)d\big)\right], \\ \lambda_2 = \lambda_3 = \ldots = \lambda_h &= \frac{n}{kh}\left[\big(a - c\big) + (k-1)\big(b - d\big)\right], \\ \lambda_{h+1} = \lambda_{h+2} = \ldots = \lambda_{h+k-1} &= \frac{n}{kh}\left[\big(a + (h-1)c\big) - \big(b + (h-1)d\big)\right], \\ \lambda_{h+k}, \lambda_{h+k+1} = \ldots = \lambda_{hk} &= \frac{n}{kh}\left[\big(a - c\big) - \big(b - d\big)\right], \\ \lambda_{hk+1} = \lambda_{hk+2} = \ldots = \lambda_n &= 0. \end{aligned} \tag{26}$$

*It is $\lambda_1 > \lambda_2 = \ldots = \lambda_h > 0$ as well as $\lambda_1 > \lambda_{h+1} = \ldots = \lambda_{h+k-1} > 0$ and $\lambda_2 = \ldots = \lambda_h > |\lambda_{h+k}| = \ldots = |\lambda_{hk}|$ as well as $\lambda_{h+1} = \ldots = \lambda_{h+k-1} > |\lambda_{h+k}| = \ldots = |\lambda_{hk}|$.*

*An eigenvector corresponding to $\lambda_1$ is given by $v_1 = \mathbf{1}_n$. The eigenspace corresponding to the eigenvalue $\lambda_2 = \ldots = \lambda_h$ is*

*spanned by the vectors $v_2, \ldots, v_h$ with*

$$v_2 = \begin{pmatrix} [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ \\ [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ \\ \vdots \\ \\ [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \end{pmatrix}, \quad v_3 = \begin{pmatrix} [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ \\ \vdots \\ \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \end{pmatrix}, \quad \cdots, \quad v_h = \begin{pmatrix} [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ \\ \vdots \\ \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \end{pmatrix},$$

*where for $z \in \mathbb{R}$, by $[z]$ we denote a block of size $\frac{n}{kh}$ with all entries equaling $z$. The eigenspace corresponding to the eigenvalue $\lambda_{h+1} = \ldots = \lambda_{h+k-1}$ is spanned by the vectors $v_{h+1}, \ldots, v_{h+k-1}$ with*

$$v_{h+1} = \begin{pmatrix} [1] \\ [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \\ \vdots \\ [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \end{pmatrix}, \quad v_{h+2} \begin{pmatrix} [-\frac{1}{k-1}] \\ [1] \\ [-\frac{1}{k-1}] \\ \vdots \\ [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \end{pmatrix}, \quad \cdots, \quad v_{h+k-1} = \begin{pmatrix} [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \\ [-\frac{1}{k-1}] \\ \vdots \\ [-\frac{1}{k-1}] \\ [1] \\ [-\frac{1}{k-1}] \end{pmatrix},$$

*where for $z \in \mathbb{R}$, by $[z]$ we denote a block of size $\frac{n}{k}$ with all entries equaling $z$. The eigenspace corresponding to the*

*eigenvalue $\lambda_{h+k} = \ldots = \lambda_{hk}$ is spanned by the vectors $v_{h+k}, \ldots, v_{hk}$ with*

$$\underbrace{\underbrace{\underbrace{\begin{pmatrix} [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \\ [-1] \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [0] \\ \vdots \\ [0]] \\ \\ \vdots \end{pmatrix}}_{=v_{h+k}} \underbrace{\begin{pmatrix} [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \\ [\frac{1}{h-1}] \\ [-1] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [0] \\ \vdots \\ [0] \\ \\ \vdots \end{pmatrix}}_{=v_{h+k+1}} \cdots \underbrace{\begin{pmatrix} [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [-1] \\ [\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [0] \\ \vdots \\ [0] \\ \\ \vdots \end{pmatrix}}_{=v_{h+k+(h-2)}}}_{h-1 \text{ many}} \underbrace{\begin{pmatrix} [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [-1] \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [0] \\ \vdots \\ [0] \\ \\ \vdots \end{pmatrix}}_{=v_{h+k+(h-1)}} \cdots \underbrace{\begin{pmatrix} [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [-1] \\ [\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ [0] \\ \vdots \\ [0] \\ \\ \vdots \end{pmatrix}}_{=v_{h+k+(2h-3)}}}_{h-1 \text{ many}} \cdots \underbrace{\underbrace{\begin{pmatrix} [1] \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ \vdots \\ \\ [0] \\ \vdots \\ [0] \\ \\ [-1] \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \\ \vdots \end{pmatrix}}_{=v_{hk-h+2}} \cdots \underbrace{\begin{pmatrix} [-\frac{1}{h-1}] \\ [-\frac{1}{h-1}] \\ \vdots \\ [-\frac{1}{h-1}] \\ [1] \\ [-\frac{1}{h-1}] \\ \\ [0] \\ \vdots \\ [0] \\ \\ \vdots \\ \\ [0] \\ \vdots \\ [0] \\ \\ [\frac{1}{h-1}] \\ [\frac{1}{h-1}] \\ \vdots \\ [\frac{1}{h-1}] \\ [-1] \\ [\frac{1}{h-1}] \\ \\ \vdots \end{pmatrix}}_{=v_{hk}}}_{h-1 \text{ many}}}_{(k-1)(h-1) \text{ many}},$$

*where for $z \in \mathbb{R}$, by $[z]$ we denote a block of size $\frac{n}{kh}$ with all entries equaling $z$.*

*Proof.* The matrix $\widetilde{\mathcal{W}}$ has only $kh$ different columns and hence $\operatorname{rank} \widetilde{\mathcal{W}} \leq kh$ and there are at most $kh$ many non-zero eigenvalues. The statement about the rank of $\widetilde{\mathcal{W}}$ follows from the statement about the eigenvalues of $\widetilde{\mathcal{W}}$.

It is easy to verify that any of the vectors $v_i$ is actually an eigenvector of $\widetilde{\mathcal{W}}$ corresponding to eigenvalue $\lambda_i$, $i \in [hk]$. We need to show that the vectors $v_2, \ldots, v_h$, the vectors $v_{h+1}, \ldots, v_{h+k-1}$, as well as the vectors $v_{h+k}, \ldots, v_{hk}$ are linearly independent. For example, let us show that $v_2, \ldots, v_h$ are linearly independent: assume that $\sum_{j \in \{2,\ldots,h\}} \alpha_j v_j = 0$ for some $\alpha_j \in \mathbb{R}$. We need to show that $\alpha_j = 0$, $j \in \{2, \ldots, h\}$. Looking at the $n$-th coordinate of $\sum_{j \in \{2,\ldots,h\}} \alpha_j v_j$, we infer that

$0 = -\frac{1}{h-1}\sum_{j\in\{2,\dots,h\}}\alpha_j$. Looking at a coordinate where $v_i$ is 1, we infer that

$$0 = \alpha_i - \frac{1}{h-1}\sum_{j\in\{2,\dots,h\}\setminus\{i\}}\alpha_j = \alpha_i\left(1+\frac{1}{h-1}\right) - \frac{1}{h-1}\sum_{j\in\{2,\dots,h\}}\alpha_j = \alpha_i\left(1+\frac{1}{h-1}\right)$$

and hence $\alpha_i = 0$, $i \in \{2,\dots,h\}$. Similarly, we can show that the vectors $v_{h+1},\dots,v_{h+k-1}$ as well as the vectors $v_{h+k},\dots,v_{hk}$ are linearly independent.

Since we assume that $a > b > c > d \geq 0$, we have

$$\big(a+(h-1)c\big) + (k-1)\big(b+(h-1)d\big) > \big(a-c\big) + (k-1)\big(b-d\big) > 0$$

and

$$\big(a+(h-1)c\big) + (k-1)\big(b+(h-1)d\big) > \big(a+(h-1)c\big) - \big(b+(h-1)d\big) = (a-b)+(h-1)(c-d) > 0,$$

which shows that $\lambda_1 > \lambda_2 = \ldots = \lambda_h > 0$ and $\lambda_1 > \lambda_{h+1} = \ldots = \lambda_{h+k-1} > 0$. It is

$$\big(a-c\big) + (k-1)\big(b-d\big) > (a-c) - (b-d) \qquad \text{and} \qquad \big(a-c\big) + (k-1)\big(b-d\big) > -(a-c) + (b-d),$$

and also

$$\begin{aligned}\big(a+(h-1)c\big) - \big(b+(h-1)d\big) &= (a-b)+(h-1)(c-d)\\ &\geq (a-b)+(c-d) > (a-b)+(d-c) = (a-c)-(b-d)\end{aligned}$$

and

$$\begin{aligned}\big(a+(h-1)c\big) - \big(b+(h-1)d\big) &= (a-b)+(h-1)(c-d)\\ &\geq (a-b)+(c-d) > (b-a)+(c-d) = -(a-c)+(b-d),\end{aligned}$$

which shows $\lambda_2 = \ldots = \lambda_h > |\lambda_{h+k}| = \ldots = |\lambda_{hk}|$ and $\lambda_{h+1} = \ldots = \lambda_{h+k-1} > |\lambda_{h+k}| = \ldots = |\lambda_{hk}|$. □

Note that we have

$$f^{(s)} - \frac{|V_s|}{n}\cdot\mathbf{1}_n = f^{(s)} - \frac{1}{h}\cdot\mathbf{1}_n = \frac{h-1}{h}v_{1+s}, \quad s\in[h-1], \tag{27}$$

where $f^{(s)}$ is the group-membership vector of $V_s$ and $f^{(s)} - \frac{|V_s|}{n}$ is the vector encountered in the second step of Algorithm 2 or Algorithm 3.

The next lemma provides an orthonormal basis of the eigenspace associated with eigenvalue $\lambda_{h+1} = \ldots = \lambda_{h+k-1}$.

**Lemma 4.** *An orthonormal basis of the eigenspace of $\widetilde{\mathcal{W}}$ corresponding to the eigenvalue $\lambda_{h+1} = \ldots = \lambda_{h+k-1}$ is given by $\{n_1,\dots,n_{k-1}\}$ with*

$$n_1 = \begin{pmatrix}[(k-1)q_1]\\ [-q_1]\\ [-q_1]\\ [-q_1]\\ \vdots\\ [-q_1]\\ [-q_1]\end{pmatrix}, \quad n_2 = \begin{pmatrix}[0]\\ [(k-2)q_2]\\ [-q_2]\\ [-q_2]\\ \vdots\\ [-q_2]\\ [-q_2]\end{pmatrix}, \quad \ldots, \quad n_i = \begin{pmatrix}[0]\\ \vdots\\ [0]\\ [(k-i)q_i]\\ [-q_i]\\ \vdots\\ [-q_i]\end{pmatrix}, \quad \ldots, \quad n_{k-1} = \begin{pmatrix}[0]\\ [0]\\ \vdots\\ [0]\\ [0]\\ [q_{k-1}]\\ [-q_{k-1}]\end{pmatrix}, \tag{28}$$

*where for $z\in\mathbb{R}$, by $[z]$ we denote a block of size $\frac{n}{k}$ with all entries equaling $z$ and*

$$q_i = \frac{1}{\sqrt{\big(\frac{n}{k}(k-i)^2 + \frac{n}{k}(k-i)\big)}}, \quad i\in[k-1]. \tag{29}$$

*Proof.* It is easy to verify that any $n_i$ is indeed an eigenvector of $\widetilde{\mathcal{W}}$ with eigenvalue $\lambda_{h+1} = \ldots = \lambda_{h+k-1}$, $i \in [k-1]$. Furthermore, we have

$$\|n_i\|^2 = q_i^2 \left(\frac{n}{k}(k-i)^2 + \frac{n}{k}(k-i)\right) = 1, \quad i \in [k-1],$$

and

$$\langle n_i, n_j \rangle = \frac{n}{k}\big( - q_i \cdot (k-j) q_j\big) + \frac{n}{k}(k-j)(q_i \cdot q_j) = 0, \quad 1 \leq i < j \leq n.$$

□

Let $\mathcal{L}$ be the expected Laplacian matrix. We have $\mathcal{L} = \mathcal{D} - \mathcal{W}$, where $\mathcal{D}$ is the expected degree matrix. The expected degree of vertex $i$ in a random graph constructed according to our variant of the stochastic block model equals $\sum_{j \in [n] \setminus \{i\}} \mathcal{W}_{ij} = \lambda_1 - a$ (with $\lambda_1$ defined in (26)) and hence $\mathcal{D} = (\lambda_1 - a) I_n$.

The following lemma characterizes the eigenvalues and eigenvectors of $Z^T \mathcal{L} Z$, where $Z \in \mathbb{R}^{n \times (n-h+1)}$ is the matrix computed in the execution of Algorithm 2 or Algorithm 3.

**Lemma 5.** *Let $Z \in \mathbb{R}^{n \times (n-h+1)}$ be any matrix whose columns form an orthonormal basis of the nullspace of $F^T$, where $F$ is the matrix that has the vectors $f^{(s)} - \frac{|V_s|}{n} \cdot \mathbf{1}_n$, $s \in [h-1]$, as columns. Then the eigenvalues of $Z^T \mathcal{L} Z$ are*

$$\lambda_1 - \lambda_1, \lambda_1 - \lambda_{h+1}, \lambda_1 - \lambda_{h+2}, \ldots, \lambda_1 - \lambda_n$$

*with $\lambda_i$ defined in (26). It is*

$$\begin{aligned} &\lambda_1 - \lambda_1 = 0, \\ &\lambda_1 - \lambda_{h+1} = \lambda_1 - \lambda_{h+2} = \ldots = \lambda_1 - \lambda_{h+k-1}, \\ &\lambda_1 - \lambda_{h+k} = \lambda_1 - \lambda_{h+k+1} = \ldots = \lambda_1 - \lambda_{hk}, \\ &\lambda_1 - \lambda_{hk+1} = \lambda_1 - \lambda_{hk+2} = \ldots = \lambda_1 - \lambda_n = \lambda_1 \end{aligned} \tag{30}$$

*with*

$$\lambda_1 - \lambda_1 < \lambda_1 - \lambda_{h+1} < \min\{\lambda_1 - \lambda_{h+k}, \lambda_1 - \lambda_{hk+1}\}, \tag{31}$$

*so that the $k$ smallest eigenvalues of $Z^T \mathcal{L} Z$ are $\lambda_1 - \lambda_1, \lambda_1 - \lambda_{h+1}, \lambda_1 - \lambda_{h+2}, \ldots, \lambda_1 - \lambda_{h+k-1}$.*

*Furthermore, there exists an orthonormal basis $\{r_1, r_{h+1}, r_{h+2}, \ldots, r_n\}$ of eigenvectors of $Z^T \mathcal{L} Z$ with $r_i$ corresponding to eigenvalue $\lambda_1 - \lambda_i$ such that*

$$Z r_1 = \mathbf{1}_n / \sqrt{n} \qquad \textit{and} \qquad Z r_{h+i} = n_i, \quad i \in [k-1],$$

*with $n_i$ defined in (28).*

*Proof.* Because of $Z^T Z = I_{(n-h+1)}$ we have

$$Z^T \mathcal{L} Z = Z^T (\mathcal{D} - \mathcal{W}) Z = Z^T \mathcal{D} Z - Z^T (\widetilde{\mathcal{W}} - a I_n) Z = (\lambda_1 - a) I_n - Z^T \widetilde{\mathcal{W}} Z + a I_n = \lambda_1 I_n - Z^T \widetilde{\mathcal{W}} Z.$$

Let $\{u_1, \ldots, u_n\}$ be an orthonormal basis of eigenvectors of $\widetilde{\mathcal{W}}$ with $u_i$ corresponding to eigenvalue $\lambda_i$. According to Lemma 3 and Lemma 4 we can choose $u_1 = \mathbf{1}_n / \sqrt{n}$ and $u_{h+i} = n_i$ for $i \in [k-1]$. We can write $\widetilde{\mathcal{W}}$ as $\widetilde{\mathcal{W}} = \sum_{i=1}^n \lambda_i u_i u_i^T$.

The nullspace of $F^T$, where $F$ is the matrix that has the vectors $f^{(s)} - \frac{|V_s|}{n} \cdot \mathbf{1}_n$, $s \in [h-1]$, as columns, equals the orthogonal complement of $\{f^{(s)} - (|V_s|/n) \cdot \mathbf{1}_n, s \in [h-1]\}$. According to (27), the orthogonal complement of $\{f^{(s)} - (|V_s|/n) \cdot \mathbf{1}_n, s \in [h-1]\}$ equals the orthogonal complement of $\{v_{1+s}, s \in [h-1]\}$, with $v_i$ defined in Lemma 3 and being an eigenvalue of $\widetilde{\mathcal{W}}$ with eigenvalue $\lambda_i$. According to Lemma 3, $\{v_{1+s}, s \in [h-1]\}$ is a basis of the eigenspace of $\widetilde{\mathcal{W}}$ corresponding to eigenvalue $\lambda_2 = \lambda_3 = \ldots = \lambda_h$, and hence the orthogonal complement of $\{v_{1+s}, s \in [h-1]\}$

equals the orthogonal complement of $\{u_2, \dots, u_h\}$, which is the subspace spanned by $\{u_1, u_{h+1}, u_{h+2}, \dots, u_n\}$. Let $U \in \mathbb{R}^{n\times(n-h+1)}$ be a matrix that has the vectors $u_1, u_{h+1}, u_{h+2}, \dots, u_n$ as columns (in this order). It follows that $U = ZR$ for some $R \in \mathbb{R}^{(n-h+1)\times(n-h+1)}$ with $R^TR = RR^T = I_{(n-h+1)}$. It is

$$Z^T\mathcal{L}Z = \lambda_1 I_n - Z^T\widetilde{\mathcal{W}}Z = \lambda_1 I_n - RU^T\left(\sum_{i=1}^{n}\lambda_i u_i u_i^T\right)UR^T.$$

Let $r_1$ be the first column of $R$, $r_{h+1}$ be the second column of $R$, $r_{h+2}$ be the third column of $R$, and so on. We have

$$\begin{aligned}
Z^T\mathcal{L}Zr_1 &= \left[\lambda_1 I_n - RU^T\left(\sum_{i=1}^{n}\lambda_i u_i u_i^T\right)UR^T\right]r_1\\
&= \lambda_1 r_1 - RU^T\left(\sum_{i=1}^{n}\lambda_i u_i u_i^T\right)Ue_1\\
&= \lambda_1 r_1 - RU^T\left(\sum_{i=1}^{n}\lambda_i u_i u_i^T\right)u_1\\
&= \lambda_1 r_1 - \lambda_1 RU^T u_1\\
&= \lambda_1 r_1 - \lambda_1 Re_1\\
&= (\lambda_1 - \lambda_1)r_1,
\end{aligned}$$

where $e_1$ denotes the first natural basis vector. Similarly, we obtain $Z^T\mathcal{L}Zr_{h+i} = (\lambda_1 - \lambda_{h+i})r_{h+i}$, $i \in [n-h]$. This proves that the eigenvalues of $Z^T\mathcal{L}Z$ are $\lambda_1 - \lambda_1, \lambda_1 - \lambda_{h+1}, \lambda_1 - \lambda_{h+2}, \dots, \lambda_1 - \lambda_n$. The claims in (30) and (31) immediately follow from Lemma 3. Clearly, it is $Zr_1 = u_1 = \mathbf{1}_n/\sqrt{n}$ and $Zr_{h+i} = u_{h+i} = n_i$ for $i \in [k-1]$. □

We need one more simple lemma.

**Lemma 6.** *Let $T \in \mathbb{R}^{n\times k}$ be a matrix that contains the vectors $\mathbf{1}_n/\sqrt{n}, n_1, n_2, \dots, n_{k-1}$, with $n_i$ defined in (28), as columns. For $i \in [n]$, let $t_i$ denote the $i$-th row of $T$. For all $i, j \in [n]$, we have $t_i = t_j$ if and only if the vertices $i$ and $j$ are in the same cluster $C_l$. If the vertices $i$ and $j$ are not in the same cluster, then $\|t_i - t_j\| = \sqrt{2k/n}$.*

*Proof.* This simply follows from the structure of the vectors $n_i$. It is, up to a permutation of the entries,

$$t_i = \left(\frac{1}{\sqrt{n}}, -q_1, -q_2, \dots, -q_{l-1}, (k-l)q_l, 0, 0, \dots, 0\right),$$

with $q_l$ defined in (29), for all $i \in [n]$ such that vertex $i$ is in cluster $C_l$, $l \in [k-1]$, and

$$t_i = \left(\frac{1}{\sqrt{n}}, -q_1, -q_2, \dots, -q_{k-1}\right)$$

for all $i \in [n]$ such that vertex $i$ is in cluster $C_k$. It is easy to verify that $\|t_i - t_j\|^2 = 2k/n$ for all $i, j \in [n]$ such that the vertices $i$ and $j$ are not in the same cluster. □

**Part 2: Deviation of $Z^TLZ$ from $Z^T\mathcal{L}Z$**

We want to obtain an upper bound on $\|Z^TLZ - Z^T\mathcal{L}Z\|$. Because of $Z^TZ = I_{(n-h+1)}$, it is $\|Z\| = \|Z^T\| = 1$ and hence

$$\|Z^TLZ - Z^T\mathcal{L}Z\| \le \|Z^T\| \cdot \|L - \mathcal{L}\| \cdot \|Z\| \le \|L - \mathcal{L}\|. \tag{32}$$

We have

$$\|L - \mathcal{L}\| = \|(D - W) - (\mathcal{D} - \mathcal{W})\| \le \|D - \mathcal{D}\| + \|W - \mathcal{W}\|,$$

with $\mathcal{D} = (\lambda_1 - a)I_n$ as we have seen in Part 1. We bound both terms separately.

- Upper bound on $\|W - \mathcal{W}\|$:

  Theorem 5.2 of Lei & Rinaldo (2015) provides a bound on $\|W - \mathcal{W}\|$: assuming that $a \geq C \ln n/n$ for some $C > 0$, for every $r > 0$ there exists a constant $\mathrm{const}(C, r)$ such that

  $$\|W - \mathcal{W}\| \leq \mathrm{const}(C, r)\sqrt{a \cdot n} \tag{33}$$

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

- Upper bound on $\|D - \mathcal{D}\|$:

  The matrix $D - \mathcal{D}$ is a diagonal matrix and hence $\|D - \mathcal{D}\| = \max_{i\in[n]} |D_{ii} - \mathcal{D}_{ii}| = \max_{i\in[n]} |D_{ii} - (\lambda_1 - a)|$. The random variable $D_{ii} = \sum_{j\in[n]\setminus\{i\}} \mathbb{1}[i \sim j]$, where $\mathbb{1}[i \sim j]$ denotes the indicator function of the event that there is an edge between vertices $i$ and $j$, is a sum of independent Bernoulli random variables. It is $\mathbb{E}[D_{ii}] = \lambda_1 - a$. For a fixed $i \in [n]$, we want to obtain an upper bound on $|D_{ii} - (\lambda_1 - a)| = |D_{ii} - \mathbb{E}[D_{ii}]|$ and distinguish two cases:

  1. $a > \frac{1}{2}$:

     Hoeffding's inequality (e.g., Boucheron et al., 2004, Theorem 1) yields

     $$\Pr[|D_{ii} - (\lambda_1 - a)| \geq t] \leq 2\exp\left(-\frac{2t^2}{n}\right)$$

     for any $t > 0$. Choosing $t = \sqrt{2(r+1)}\sqrt{a \cdot n \ln n}$ for $r > 0$, we have with $\mathrm{const}(r) = \sqrt{2(r+1)}$ that

     $$\Pr\left[|D_{ii} - (\lambda_1 - a)| \geq \mathrm{const}(r) \cdot \sqrt{a \cdot n \ln n}\right] \leq 2\exp\left(-4(r+1)a \ln n\right) \leq n^{-(r+1)}. \tag{34}$$

  2. $a \leq \frac{1}{2}$:

     Bernstein's inequality (e.g., Boucheron et al., 2004, Theorem 3) yields

     $$\Pr\left[|D_{ii} - (\lambda_1 - a)| > tn\right] \leq 2\exp\left(-\frac{nt^2}{2\left(\frac{\mathrm{Var}[D_{ii}]}{n} + \frac{t}{3}\right)}\right)$$

     for any $t > 0$. It is

     $$\mathrm{Var}[D_{ii}] = \sum_{j\in[n]\setminus\{i\}} \mathrm{Var}\left[\mathbb{1}[i \sim j]\right] = \sum_{j\in[n]\setminus\{i\}} \Pr[\mathbb{1}[i \sim j]](1 - \Pr[\mathbb{1}[i \sim j]]) \leq na(1-a) \leq na$$

     since the function $x \mapsto x(1-x)$ is monotonically increasing on $[0, 1/2]$. If we choose $t = \mathrm{const} \cdot \frac{\sqrt{a \cdot n \ln n}}{n}$ for some $\mathrm{const} > 0$, assuming that $a \geq C \ln n/n$ for some $C > 0$, we have

     $$\frac{\mathrm{Var}[D_{ii}]}{n} + \frac{t}{3} \leq a\left(1 + \frac{\mathrm{const}}{3\sqrt{C}}\right)$$

     and hence

     $$\Pr\left[|D_{ii} - (\lambda_1 - a)| > \mathrm{const} \cdot \sqrt{a \cdot n \ln n}\right] \leq 2\exp\left(-\frac{\mathrm{const}^2 \cdot \ln n}{2 + \frac{2\,\mathrm{const}}{3\sqrt{C}}}\right).$$

     Because of

     $$\frac{\mathrm{const}^2}{2 + \frac{2\,\mathrm{const}}{3\sqrt{C}}} \to \infty \quad \text{as const} \to \infty,$$

     for every $r > 0$ we can choose $\mathrm{const} = \mathrm{const}(C, r)$ large enough such that $\mathrm{const}^2 / \left(2 + \frac{2\,\mathrm{const}}{3\sqrt{C}}\right) \geq 2(r+1)$ and

     $$\Pr\left[|D_{ii} - (\lambda_1 - a)| > \mathrm{const}(C, r) \cdot \sqrt{a \cdot n \ln n}\right] \leq n^{-(r+1)}. \tag{35}$$

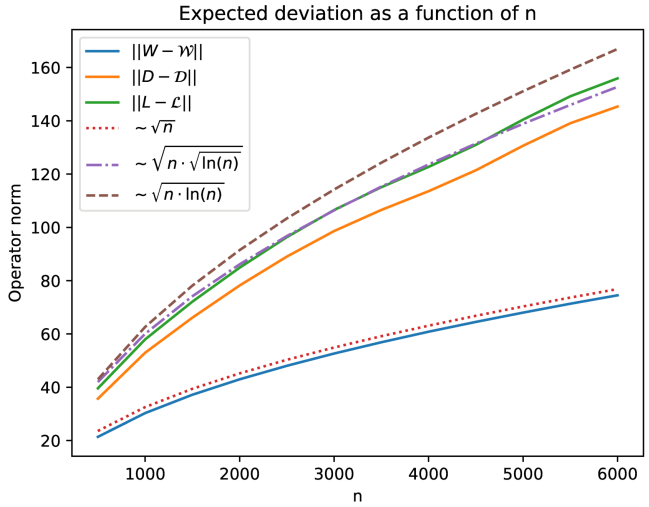


*Figure 6.* Average deviations $\|W - \mathcal{W}\|$, $\|D - \mathcal{D}\|$ and $\|L - \mathcal{L}\|$ as a function of $n$ when $a = 0.6, b = 0.5, c = 0.4, d = 0.3$ are constant, $k = 5$ and $h = 2$. The average is computed over sampling the graph for 100 times.

Choosing $\mathrm{const}(C, r)$ as the maximum of $\mathrm{const}(r)$ encountered in (34) and $\mathrm{const}(C, r)$ encountered in (35), we see that there exists $\mathrm{const}(C, r)$ such that

$$\Pr\left[|D_{ii} - (\lambda_1 - a)| > \mathrm{const}(C, r) \cdot \sqrt{a \cdot n \ln n}\right] \leq n^{-(r+1)},$$

no matter whether $a > 1/2$ or $1/2 \geq a \geq C \ln n/n$. Applying a union bound we obtain

$$\Pr\left[\max_{i \in [n]} |D_{ii} - (\lambda_1 - a)| > \mathrm{const}(C, r) \cdot \sqrt{a \cdot n \ln n}\right] \leq n \cdot n^{-(r+1)} = n^{-r},$$

and hence with probability at least $1 - n^{-r}$ we have

$$\|D - \mathcal{D}\| \leq \mathrm{const}(C, r)\sqrt{a \cdot n \ln n}. \tag{36}$$

From (33) and (36) we see that for every $r > 0$ there exists $\mathrm{const}(C, r)$ such that with probability at least $1 - n^{-r}$ we have

$$\|W - \mathcal{W}\| \leq \mathrm{const}(C, r)\sqrt{a \cdot n} \qquad \text{and} \qquad \|D - \mathcal{D}\| \leq \mathrm{const}(C, r)\sqrt{a \cdot n \ln n} \tag{37}$$

and hence

$$\|Z^T L Z - Z^T \mathcal{L} Z\| \leq \|L - \mathcal{L}\| \leq \|D - \mathcal{D}\| + \|W - \mathcal{W}\| \leq \mathrm{const}(C, r)\sqrt{a \cdot n \ln n}. \tag{38}$$

For illustrative purposes, we show empirically that, in general, our upper bounds on $\|W - \mathcal{W}\|$, $\|D - \mathcal{D}\|$ and $\|L - \mathcal{L}\|$ in (37) and (38), respectively, are tight, up to a factor of at most $\sqrt[4]{\ln n}$ in case of $\|D - \mathcal{D}\|$ and $\|L - \mathcal{L}\|$. The plot in Figure 6 shows the observed deviations $\|W - \mathcal{W}\|$, $\|D - \mathcal{D}\|$ and $\|L - \mathcal{L}\|$ as a function of $n$ when $a = 0.6, b = 0.5, c = 0.4, d = 0.3$ are constant, $k = 5$ and $h = 2$. The shown curves are average results, obtained from sampling the graph for 100 times.

### Part 3: Proving Theorem 1 for Algorithm 2 (unnormalized SC with fairness constraints)

In the last step of Algorithm 2 we apply $k$-means clustering to the rows of the matrix $ZY$, where $Y \in \mathbb{R}^{(n-h+1)\times k}$ contains some orthonormal eigenvectors corresponding to the $k$ smallest eigenvalues of $Z^T L Z$ as columns. We want to show that up to some orthogonal transformation, the rows of $ZY$ are close to the rows of $Z\mathcal{Y}$, where $\mathcal{Y} \in \mathbb{R}^{(n-h+1)\times k}$ contains some orthonormal eigenvectors corresponding to the $k$ smallest eigenvalues of $Z^T \mathcal{L} Z$ as columns. According to Lemma 5, we

can choose $\mathcal{Y}$ in such a way that $Z\mathcal{Y} = T$ with $T$ as in Lemma 6, that is $T$ contains the vectors $\mathbf{1}_n/\sqrt{n}, n_1, n_2, \ldots, n_{k-1}$, with $n_i$ defined in (28), as columns.

We want to obtain an upper bound on $\min_{U\in\mathbb{R}^{k\times k}:U^TU=UU^T=I_k} \|Z\mathcal{Y} - ZYU\|_F$. For any $U \in \mathbb{R}^{k\times k}$ with $U^TU = UU^T = I_k$, because of $Z^TZ = I_{(n-h+1)}$ we have

$$\|Z\mathcal{Y} - ZYU\|_F^2 = \|Z(\mathcal{Y} - YU)\|_F^2 = \mathrm{Tr}((\mathcal{Y} - YU)^TZ^TZ(\mathcal{Y} - YU)) = \|\mathcal{Y} - YU\|_F^2$$

and hence

$$\min_{U\in\mathbb{R}^{k\times k}:U^TU=UU^T=I_k} \|Z\mathcal{Y} - ZYU\|_F = \min_{U\in\mathbb{R}^{k\times k}:U^TU=UU^T=I_k} \|\mathcal{Y} - YU\|_F. \tag{39}$$

We proceed similarly to Lei & Rinaldo (2015). According to Proposition 2.2 and Equation (2.6) in Vu & Lei (2013) we have (note that the set of all orthogonal matrices $U \in \mathbb{R}^{k\times k}$ is a compact subset of $\mathbb{R}^{k\times k}$ and hence the infimum is indeed a minimum)

$$\min_{U\in\mathbb{R}^{k\times k}:U^TU=UU^T=I_k} \|\mathcal{Y} - YU\|_F \le \sqrt{2}\|\mathcal{Y}\mathcal{Y}^T(I_{(n-h+1)} - YY^T)\|_F \overset{(23)}{\le} \sqrt{2}\sqrt{k}\,\|\mathcal{Y}\mathcal{Y}^T(I_{(n-h+1)} - YY^T)\|. \tag{40}$$

According to Lemma 5 the eigenvalues of $Z^T\mathcal{L}Z$ are $\lambda_1 - \lambda_1, \lambda_1 - \lambda_{h+1}, \lambda_1 - \lambda_{h+2}, \ldots, \lambda_1 - \lambda_n$. The $k$ smallest eigenvalues are $\lambda_1 - \lambda_1, \lambda_1 - \lambda_{h+1}, \lambda_1 - \lambda_{h+2}, \ldots, \lambda_1 - \lambda_{h+k-1}$ and the $(k+1)$-th smallest eigenvalue is either $\lambda_1 - \lambda_{h+k}$ or $\lambda_1$. Hence, for the eigengap $\gamma$ between the $k$-th and the $(k+1)$-th smallest eigenvalue we have

$$\gamma = \min\left\{(\lambda_1 - \lambda_{h+k}) - (\lambda_1 - \lambda_{h+k-1}), \lambda_1 - (\lambda_1 - \lambda_{h+k-1})\right\} = \min\left\{\frac{n}{k}(c-d), \frac{n}{kh}((a-b) + (h-1)(c-d))\right\}.$$

It is

$$\frac{n}{2k}(c-d) \le \frac{n(h-1)}{hk}(c-d) \le \gamma \le \frac{n}{k}(c-d). \tag{41}$$

We want to show that

$$\|\mathcal{Y}\mathcal{Y}^T(I_{(n-h+1)} - YY^T)\| \le \frac{4}{\gamma}\|Z^T\mathcal{L}Z - Z^TLZ\|. \tag{42}$$

If $\|Z^T\mathcal{L}Z - Z^TLZ\| > \frac{\gamma}{4}$, then (42) holds trivially because of

$$\|\mathcal{Y}\mathcal{Y}^T(I_{(n-h+1)} - YY^T)\| \le \|\mathcal{Y}\mathcal{Y}^T\| \cdot \|I_{(n-h+1)} - YY^T\| = 1 \cdot 1 = 1.$$

Assume that $\|Z^T\mathcal{L}Z - Z^TLZ\| \le \frac{\gamma}{4}$ and let $\mu_1 \le \mu_2 \le \ldots \le \mu_{n-h+1}$ be the eigenvalues of $Z^TLZ$. Since $L$ is positive semi-definite, so is $Z^TLZ$, and hence $\mu_1 \ge 0$. Let $\lambda'_1 \le \lambda'_2 \le \ldots, \lambda'_{n-h+1}$ be the eigenvalues $\lambda_1 - \lambda_1, \lambda_1 - \lambda_{h+1}, \lambda_1 - \lambda_{h+2}, \ldots, \lambda_1 - \lambda_n$ of $Z^T\mathcal{L}Z$ in ascending order. According to Weyl's Perturbation Theorem (e.g., Bhatia, 1997, Corollary III.2.6) it is

$$|\mu_i - \lambda'_i| \le \|Z^T\mathcal{L}Z - Z^TLZ\| \le \frac{\gamma}{4}, \quad i \in [n-h+1].$$

In particular, we have

$$\mu_1, \ldots, \mu_k \in \left[0, \lambda'_k + \frac{\gamma}{4}\right], \quad \mu_{k+1}, \ldots, \mu_n \in \left[\lambda'_{k+1} - \frac{\gamma}{4}, \infty\right)$$

with $\left(\lambda'_{k+1} - \frac{\gamma}{4}\right) - \left(\lambda'_k + \frac{\gamma}{4}\right) = \frac{\gamma}{2}$. The Davis-Kahan sinΘ Theorem (e.g., Bhatia, 1997, Theorem VII.3.1) yields that

$$\|\mathcal{Y}\mathcal{Y}^T(I_{(n-h+1)} - YY^T)\| \le \frac{2}{\gamma}\|Z^T\mathcal{L}Z - Z^TLZ\|$$

and hence (42). Combining (39) to (42), we end up with

$$\min_{U\in\mathbb{R}^{k\times k}:U^TU=UU^T=I_k} \|Z\mathcal{Y} - ZYU\|_F \le \frac{16\sqrt{k^3}}{n(c-d)}\|Z^T\mathcal{L}Z - Z^TLZ\|. \tag{43}$$

Using (38) from Part 2, we see that with probability at least $1 - n^{-r}$

$$\min_{U \in \mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|Z\mathcal{Y} - ZYU\|_F \leq \text{const}(C, r) \cdot \frac{\sqrt{k^3}}{c-d} \cdot \sqrt{\frac{a \cdot \ln n}{n}}. \tag{44}$$

We use Lemma 5.3 in Lei & Rinaldo (2015) to complete the proof of Theorem 1 for Algorithm 2. Assume that (44) holds and let $U \in \mathbb{R}^{k\times k}$ be an orthogonal matrix attaining the minimum, that is we have

$$\|Z\mathcal{Y}U^T - ZY\|_F = \|Z\mathcal{Y} - ZYU\|_F \leq \text{const}(C, r) \cdot \frac{\sqrt{k^3}}{c-d} \cdot \sqrt{\frac{a \cdot \ln n}{n}}. \tag{45}$$

As we have noted above, we can choose $\mathcal{Y}$ in such a way that $Z\mathcal{Y} = T$ with $T$ as in Lemma 6. According to Lemma 6, if we denote the $i$-th row of $T$ by $t_i$, then $t_i = t_j$ if the vertices $i$ and $j$ are in the same cluster and $\|t_i - t_j\| = \sqrt{2k/n}$ if the vertices $i$ and $j$ are not in the same cluster. Since multiplying $T$ by $U^T$ from the right side has the effect of applying an orthogonal transformation to the rows of $T$, the same properties are true for the matrix $TU^T$. Lemma 5.3 in Lei & Rinaldo (2015) guarantees that for any $\delta \leq \sqrt{2k/n}$, if

$$\frac{16 + 8M}{\delta^2}\|TU^T - ZY\|_F^2 < \underbrace{|C_l|}_{=\frac{n}{k}}, \quad l \in [k], \tag{46}$$

with $|C_l|$ being the size of cluster $C_l$, then a $(1 + M)$-approximation algorithm for $k$-means clustering applied to the rows of the matrix $ZY$ returns a clustering that misclassifies at most

$$\frac{4(4 + 2M)}{\delta^2}\|TU^T - ZY\|_F^2 \tag{47}$$

many vertices. If we choose $\delta = \sqrt{2k/n}$, then for a small enough $\widehat{C}_1 = \widehat{C}_1(C, r)$ the condition (11) implies (46) because of (45). Also, for a large enough $\widetilde{C}_1 = \widetilde{C}_1(C, r)$, the expression (47) is upper bounded by the expression (12).

**Part 4: Proving Theorem 1 for Algorithm 3 (normalized SC with fairness constraints)**

According to Part 2, for every $r > 0$ there exists $\text{const}(C, r)$ such that with probability at least $1 - n^{-r}$ we have

$$\|D - \mathcal{D}\| \leq \text{const}(C, r)\sqrt{a \cdot n \ln n}. \tag{48}$$

Condition (13), with a suitable $\widehat{C}_2 = \widehat{C}_2(C, r)$, implies that in this case we also have

$$\|D - \mathcal{D}\| \leq \frac{\lambda_1 - a}{2}. \tag{49}$$

Let $\mu'_1, \ldots, \mu'_{n-h+1}$ denote the eigenvalues of $Z^TDZ$. It is $\mathcal{D} = (\lambda_1 - a)I_n$ (see Part 1) and because of $Z^TZ = I_{(n-h+1)}$ we have $Z^T\mathcal{D}Z = (\lambda_1 - a)I_{(n-h+1)}$. According to Weyl's Perturbation Theorem (e.g., Bhatia, 1997, Corollary III.2.6) it is

$$|\mu'_i - (\lambda_1 - a)| \leq \|Z^TDZ - Z^T\mathcal{D}Z\| \leq \|D - \mathcal{D}\|, \quad i \in [n - h + 1], \tag{50}$$

where the second inequality follows analogously to (32). It follows from (49) that

$$\mu'_i \geq \frac{\lambda_1 - a}{2} \overset{(13)}{>} 0, \quad i \in [n - h + 1], \tag{51}$$

In particular, this shows that $Z^TDZ$ is positive definite and hence Algorithm 3 is well-defined.

Now we proceed similarly to Part 3. In the last step of Algorithm 3 we apply $k$-means clustering to the rows of the matrix $ZQ^{-1}X$, where $Q \in \mathbb{R}^{(n-h+1)\times(n-h+1)}$ is the positive definite square root of $Z^TDZ$ and $X \in \mathbb{R}^{(n-h+1)\times k}$ contains some orthonormal eigenvectors corresponding to the $k$ smallest eigenvalues of $Q^{-1}Z^TLZQ^{-1}$ as columns. We want to show that up to some orthogonal transformation, the rows of $ZQ^{-1}X$ are close to the rows of $Z\mathcal{Q}^{-1}\mathcal{X}$, where

$\mathcal{Q} \in \mathbb{R}^{(n-h+1)\times(n-h+1)}$ is the positive definite square root of $Z^T\mathcal{D}Z$ and $\mathcal{X} \in \mathbb{R}^{(n-h+1)\times k}$ contains some orthonormal eigenvectors corresponding to the $k$ smallest eigenvalues of $\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1}$ as columns. It is $Z^T\mathcal{D}Z = (\lambda_1 - a)I_{(n-h+1)}$. Consequently, $\mathcal{Q} = \sqrt{\lambda_1 - a} \cdot I_{(n-h+1)}$ and $\mathcal{Q}^{-1} = \frac{1}{\sqrt{\lambda_1 - a}} \cdot I_{(n-h+1)}$ and it is $\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1} = \frac{1}{\lambda_1 - a} \cdot Z^T\mathcal{L}Z$. Hence, the eigenvalues of $\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1}$ are the eigenvalues of $Z^T\mathcal{L}Z$ rescaled by $(\lambda_1 - a)^{-1}$ with the same eigenvectors as for $Z^T\mathcal{L}Z$. According to Lemma 5, we can choose $\mathcal{X}$ in such a way that $Z\mathcal{Q}^{-1}\mathcal{X} = \frac{1}{\sqrt{\lambda_1 - a}} \cdot Z\mathcal{X} = \frac{1}{\sqrt{\lambda_1 - a}} \cdot T$ with $T$ as in Lemma 6, that is $T$ contains the vectors $\mathbf{1}_n/\sqrt{n}, n_1, n_2, \ldots, n_{k-1}$, with $n_i$ defined in (28), as columns.

We want to obtain an upper bound on $\min_{U\in\mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|Z\mathcal{Q}^{-1}\mathcal{X} - ZQ^{-1}XU\|_F$. Analogously to (39) we obtain

$$\min_{U\in\mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|Z\mathcal{Q}^{-1}\mathcal{X} - ZQ^{-1}XU\|_F = \min_{U\in\mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|\mathcal{Q}^{-1}\mathcal{X} - Q^{-1}XU\|_F.$$

The rank of both $\mathcal{Q}^{-1}\mathcal{X}$ and $Q^{-1}XU$ equals $k$ and hence the rank of $\mathcal{Q}^{-1}\mathcal{X} - Q^{-1}XU$ is not greater than $2k$. We have

$$\|\mathcal{Q}^{-1}\mathcal{X} - Q^{-1}XU\|_F \overset{(23)}{\leq} \sqrt{2k} \cdot \|\mathcal{Q}^{-1}\mathcal{X} - Q^{-1}XU\| \leq \sqrt{2k} \cdot \|\mathcal{Q}^{-1}\| \cdot \|\mathcal{X} - XU\| + \sqrt{2k} \cdot \|\mathcal{Q}^{-1} - Q^{-1}\| \cdot \|XU\|$$

with $\|\mathcal{Q}^{-1}\| = \frac{1}{\sqrt{\lambda_1 - a}}$ and $\|XU\| = 1$ because of $X^TX = I_k$ and $U^TU = I_k$. Hence

$$\min_{U:\, U^TU = UU^T = I_k} \|Z\mathcal{Q}^{-1}\mathcal{X} - ZQ^{-1}XU\|_F \leq \frac{\sqrt{2k}}{\sqrt{\lambda_1 - a}} \cdot \min_{U:\, U^TU = UU^T = I_k} \|\mathcal{X} - XU\| + \sqrt{2k} \cdot \|\mathcal{Q}^{-1} - Q^{-1}\|. \quad (52)$$

Because of (23) we have

$$\min_{U\in\mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|\mathcal{X} - XU\| \leq \min_{U\in\mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|\mathcal{X} - XU\|_F \quad (53)$$

and similarly to how we obtained the bound (43) in Part 2, we can show that

$$\min_{U\in\mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|\mathcal{X} - XU\|_F \leq \frac{16\sqrt{k^3}(\lambda_1 - a)}{n(c-d)} \|\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1} - Q^{-1}Z^TLZQ^{-1}\|. \quad (54)$$

Before looking at $\|\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1} - Q^{-1}Z^TLZQ^{-1}\|$ let us first look at the second term in (52). Because $Q^{-1}$ is symmetric and $\mathcal{Q}^{-1} = \frac{1}{\sqrt{\lambda_1 - a}} \cdot I_{(n-h+1)}$ we have

$$\|\mathcal{Q}^{-1} - Q^{-1}\| = \max\left\{\left|\nu_i - \frac{1}{\sqrt{\lambda_1 - a}}\right| : \nu_i \text{ is an eigenvalue of } Q^{-1}\right\}.$$

It is $Q^2 = Z^TDZ$. Denoting the eigenvalues of $Z^TDZ$ by $\mu'_1, \ldots, \mu'_{n-h+1}$ (note that all of them are greater than zero according to (51)), the eigenvalues of $Q^{-1}$ are $1/\sqrt{\mu'_1}, \ldots, 1/\sqrt{\mu'_{n-h+1}}$. For any $z_1, z_2 > 0$ we have

$$|\sqrt{z_1} - \sqrt{z_2}| = \frac{|(\sqrt{z_1} - \sqrt{z_2})(\sqrt{z_1} + \sqrt{z_2})|}{\sqrt{z_1} + \sqrt{z_2}} = \frac{|z_1 - z_2|}{\sqrt{z_1} + \sqrt{z_2}} \leq \frac{|z_1 - z_2|}{\sqrt{z_2}} \quad (55)$$

and

$$\left|\frac{1}{\sqrt{z_1}} - \frac{1}{\sqrt{z_2}}\right| = \frac{|\sqrt{z_1} - \sqrt{z_2}|}{\sqrt{z_1}\sqrt{z_2}} \overset{\text{for } z_1 \geq \frac{z_2}{2}}{\leq} \frac{\sqrt{2} \cdot |\sqrt{z_1} - \sqrt{z_2}|}{z_2} \overset{(55)}{\leq} \frac{\sqrt{2} \cdot |z_1 - z_2|}{\sqrt{z_2^3}}. \quad (56)$$

According to (51) we have $\mu'_i \geq \frac{\lambda_1 - a}{2} > 0$, $i \in [n-h+1]$, and hence

$$\left|\frac{1}{\sqrt{\mu'_i}} - \frac{1}{\sqrt{\lambda_1 - a}}\right| \overset{(56)}{\leq} \frac{\sqrt{2} \cdot |\mu'_i - (\lambda_1 - a)|}{\sqrt{(\lambda_1 - a)^3}} \overset{(50)}{\leq} \frac{\sqrt{2} \cdot \|D - \mathcal{D}\|}{\sqrt{(\lambda_1 - a)^3}}, \quad i \in [n-h+1],$$

and

$$\|\mathcal{Q}^{-1} - Q^{-1}\| \leq \frac{\sqrt{2} \cdot \|D - \mathcal{D}\|}{\sqrt{(\lambda_1 - a)^3}}. \quad (57)$$

Let us now look at $\|\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1} - Q^{-1}Z^TLZQ^{-1}\|$. It is

$$\begin{aligned}\|\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1} - Q^{-1}Z^TLZQ^{-1}\| \leq \|\mathcal{Q}^{-1} - Q^{-1}\| \cdot \|Z^T\mathcal{L}Z\| \cdot \|\mathcal{Q}^{-1}\| + \\ \|Q^{-1}\| \cdot \|Z^T\mathcal{L}Z - Z^TLZ\| \cdot \|\mathcal{Q}^{-1}\| + \|Q^{-1}\| \cdot \|Z^TLZ\| \cdot \|\mathcal{Q}^{-1} - Q^{-1}\|.\end{aligned} \tag{58}$$

It is $\|\mathcal{Q}^{-1}\| = \frac{1}{\sqrt{\lambda_1 - a}}$. According to Lemma 5, the largest eigenvalue of $Z^T\mathcal{L}Z$ is $\lambda_1$ or $\lambda_1 - \lambda_{hk}$, where $\lambda_1 - \lambda_{hk} \leq 2\lambda_1$ according to Lemma 3. Consequently, $\|Z^T\mathcal{L}Z\| \leq 2\lambda_1$. It is

$$\|Q^{-1}\| \leq \|\mathcal{Q}^{-1} - Q^{-1}\| + \|\mathcal{Q}^{-1}\| \overset{(57)}{\leq} \frac{\sqrt{2} \cdot \|D - \mathcal{D}\|}{\sqrt{(\lambda_1 - a)^3}} + \frac{1}{\sqrt{\lambda_1 - a}}$$

and

$$\|Z^TLZ\| \leq \|Z^TLZ - Z^T\mathcal{L}Z\| + \|Z^T\mathcal{L}Z\| \overset{(32)}{\leq} \|L - \mathcal{L}\| + 2\lambda_1.$$

It follows that

$$\begin{aligned}\|\mathcal{Q}^{-1}Z^T\mathcal{L}Z\mathcal{Q}^{-1} - Q^{-1}Z^TLZQ^{-1}\| \leq \frac{4\lambda_1 \cdot \|D - \mathcal{D}\|}{(\lambda_1 - a)^2} + \left(\frac{\sqrt{2} \cdot \|D - \mathcal{D}\|}{(\lambda_1 - a)^2} + \frac{1}{\lambda_1 - a}\right) \cdot \|\mathcal{L} - L\| + \\ \left(\frac{2 \cdot \|D - \mathcal{D}\|^2}{(\lambda_1 - a)^3} + \frac{\sqrt{2} \cdot \|D - \mathcal{D}\|}{(\lambda_1 - a)^2}\right) \cdot (\|L - \mathcal{L}\| + 2\lambda_1) \\ \leq \frac{8\lambda_1 \cdot \|D - \mathcal{D}\|}{(\lambda_1 - a)^2} + \left(\frac{4 \cdot \|D - \mathcal{D}\|}{(\lambda_1 - a)^2} + \frac{1}{\lambda_1 - a}\right) \cdot \|\mathcal{L} - L\| + \frac{2 \cdot \|D - \mathcal{D}\|^2}{(\lambda_1 - a)^3} \cdot (\|L - \mathcal{L}\| + 2\lambda_1).\end{aligned} \tag{59}$$

If (37) and (38) hold, then, after combining (52), (53), (54), (57), (59) and using that $\lambda_1 - a > \lambda_1/2$, which follows from (13), we end up with

$$\begin{aligned}\min_{U \in \mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|Z\mathcal{Q}^{-1}\mathcal{X} - ZQ^{-1}XU\|_F \leq \frac{\text{const}(C,r) \cdot k^2}{n(c-d)\sqrt{\lambda_1 - a}} \left(\sqrt{a \cdot n \ln n} + \frac{a \cdot n \ln n}{\lambda_1 - a} + \frac{(a \cdot n \ln n)^{3/2}}{(\lambda_1 - a)^2}\right) + \\ \frac{\text{const}(C,r)}{\sqrt{\lambda_1 - a}} \cdot \frac{\sqrt{k} \cdot \sqrt{a \cdot n \ln n}}{\lambda_1 - a}\end{aligned}$$

for some $\text{const}(C,r)$. Using that $\sqrt{a \cdot n \ln n} \leq \sqrt{k \cdot a \cdot n \ln n} \leq \frac{\widehat{C}_2}{1+M}(\lambda_1 - a) \leq \widehat{C}_2(\lambda_1 - a)$ due to (13), for some $\widehat{C}_2 = \widehat{C}_2(C,r)$ that we will specify shortly (we will choose it smaller than 1), we can simplify this bound such that

$$\min_{U \in \mathbb{R}^{k\times k}: U^TU = UU^T = I_k} \|Z\mathcal{Q}^{-1}\mathcal{X} - ZQ^{-1}XU\|_F \leq \frac{\text{const}(C,r) \cdot k^2}{n(c-d)\sqrt{\lambda_1 - a}} \cdot \sqrt{a \cdot n \ln n} + \frac{\text{const}(C,r)}{\sqrt{\lambda_1 - a}} \cdot \frac{\sqrt{k} \cdot \sqrt{a \cdot n \ln n}}{\lambda_1 - a}. \tag{60}$$

Similarly to Part 3, we use Lemma 5.3 in Lei & Rinaldo (2015) to complete the proof of Theorem 1 for Algorithm 3. Assume that (60) holds and let $U \in \mathbb{R}^{k\times k}$ be an orthogonal matrix attaining the minimum, that is we have

$$\begin{aligned}\|Z\mathcal{Q}^{-1}\mathcal{X}U^T - ZQ^{-1}X\|_F = \|Z\mathcal{Q}^{-1}\mathcal{X} - ZQ^{-1}XU\|_F \\ \leq \frac{\text{const}(C,r) \cdot k^2}{n(c-d)\sqrt{\lambda_1 - a}} \cdot \sqrt{a \cdot n \ln n} + \frac{\text{const}(C,r)}{\sqrt{\lambda_1 - a}} \cdot \frac{\sqrt{k} \cdot \sqrt{a \cdot n \ln n}}{\lambda_1 - a}.\end{aligned} \tag{61}$$

As we have noted above, we can choose $\mathcal{X}$ in such a way that $Z\mathcal{Q}^{-1}\mathcal{X} = \frac{1}{\sqrt{\lambda_1 - a}} \cdot T$ with $T$ as in Lemma 6. According to Lemma 6, if we denote the $i$-th row of $\frac{1}{\sqrt{\lambda_1 - a}} \cdot T$ by $\tilde{t}_i$, then $\tilde{t}_i = \tilde{t}_j$ if the vertices $i$ and $j$ are in the same cluster and $\|\tilde{t}_i - \tilde{t}_j\| = \sqrt{\frac{2k}{n(\lambda_1 - a)}}$ if the vertices $i$ and $j$ are not in the same cluster. Since multiplying $\frac{1}{\sqrt{\lambda_1 - a}} \cdot T$ by $U^T$ from the right

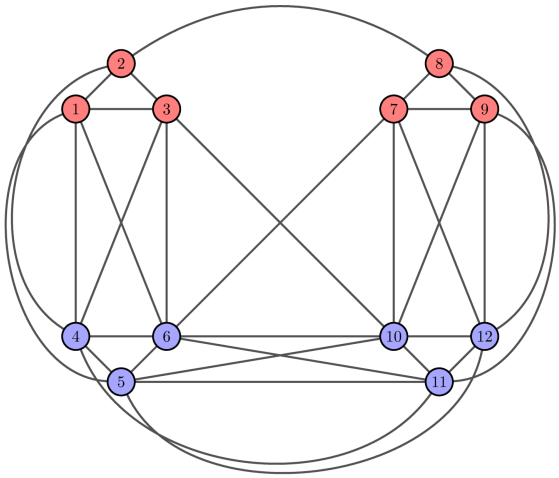


*Figure 7.* Example of a graph for which both standard spectral clustering and our fair versions are able to recover the fair meaningful ground-truth clustering while a naive approach that runs standard spectral clustering on each group separately fails to do so. It is $V = [12]$, $V_1 = \{1, 2, 3, 7, 8, 9\}$, $V_2 = \{4, 5, 6, 10, 11, 12\}$ and the fair ground-truth clustering is $V = \{1, 2, 3, 4, 5, 6\} \dot\cup \{7, 8, 9, 10, 11, 12\}$.

side has the effect of applying an orthogonal transformation to the rows of $\frac{1}{\sqrt{\lambda_1 - a}} \cdot T$, the same properties are true for the matrix $\frac{1}{\sqrt{\lambda_1 - a}} \cdot TU^T$. Lemma 5.3 in Lei & Rinaldo (2015) guarantees that for any $\delta \leq \sqrt{\frac{2k}{n(\lambda_1 - a)}}$, if

$$\frac{16 + 8M}{\delta^2} \left\| \frac{1}{\sqrt{\lambda_1 - a}} \cdot TU^T - ZQ^{-1}X \right\|_F^2 < \underbrace{|C_l|}_{=\frac{n}{k}}, \quad l \in [k], \tag{62}$$

with $|C_l|$ being the size of cluster $C_l$, then a $(1 + M)$-approximation algorithm for $k$-means clustering applied to the rows of the matrix $ZQ^{-1}X$ returns a clustering that misclassifies at most

$$\frac{4(4 + 2M)}{\delta^2} \left\| \frac{1}{\sqrt{\lambda_1 - a}} \cdot TU^T - ZQ^{-1}X \right\|_F^2 \tag{63}$$

many vertices. If we choose $\delta = \sqrt{\frac{2k}{n(\lambda_1 - a)}}$, then for a small enough $\widehat{C}_2 = \widehat{C}_2(C, r)$ (chosen smaller than 1 and also so small that (48) implies (49)), the condition (13) implies (62) because of (61). Also, for a large enough $\widetilde{C}_2 = \widetilde{C}_2(C, r)$ the expression (63) is upper bounded by the expression (14).

## D. Why Running Standard Spectral Clustering on Each Group $V_s$ Separately is not a Good Idea

One might think that the following was a good idea for partitioning $V = V_1 \dot\cup \ldots \dot\cup V_h$ into $k$ clusters such that every cluster has a high balance value: we could try to run standard spectral clustering with $k$ clusters on each of the groups $V_s$, $s \in [h]$, separately and then to merge the $k \cdot h$ many clusters to end up with $k$ clusters.

The graph shown in Figure 7 illustrates that such an approach, in general, fails to recover an underlying fair ground-truth clustering, even when standard spectral clustering succeeds. We have $V = [12]$ and two groups $V_1 = \{1, 2, 3, 7, 8, 9\}$ (shown in red) and $V_2 = \{4, 5, 6, 10, 11, 12\}$ (shown in blue). We want to partition $V$ into two clusters. It can be verified that a clustering with minimum RatioCut value is given by $V = \{1, 2, 3, 4, 5, 6\} \dot\cup \{7, 8, 9, 10, 11, 12\}$ and that this clustering is found by running standard spectral clustering. This clustering is perfectly fair with $\mathrm{balance}(\{1, 2, 3, 4, 5, 6\}) = \mathrm{balance}(\{7, 8, 9, 10, 11, 12\}) = 1$ and is also returned by our fair versions of spectral clustering. Let us now look at the idea of running standard spectral clustering on $V_1$ and $V_2$ separately: when running spectral clustering on the subgraph induced by $V_1$, we obtain the clustering $V_1 = \{1, 2, 3\} \dot\cup \{7, 8, 9\}$ as we would hope for. However, in the subgraph induced by $V_2$ the clustering $V_2 = \{4, 5, 6\} \dot\cup \{10, 11, 12\}$ does not have minimum RatioCut value and is not returned by spectral

clustering. Consequently, no matter how we merge the two clusters for $V_1$ and the two clusters for $V_2$, we do not end up with the clustering $V = \{1, 2, 3, 4, 5, 6\} \dot{\cup} \{7, 8, 9, 10, 11, 12\}$.

Note that for these findings to hold we do not require the specific graph shown in Figure 7. The key is its structure: let $V_1 = \{1, 2, 3, 7, 8, 9\}$, $V_2 = \{4, 5, 6, 10, 11, 12\}$, $C_1 = \{1, 2, 3, 4, 5, 6\}$ and $C_2 = \{7, 8, 9, 10, 11, 12\}$. Then the graph looks like a realization of the following random graph model: as in our variant of the stochastic block model introduced in Section 4, two vertices $i$ and $j$ are connected with an edge with a certain probability $\Pr(i, j)$, which is now given by

$$\Pr(i,j) = \begin{cases} a, & i,j \in C_1 \ \vee \ i,j \in C_2 \ \vee \ i,j \in V_2, \\ b, & \text{else}, \end{cases}$$

with $a$ large and $b$ small.