Supplementary material for Compressive Spectral Clustering

A. Proof of Theorem 3.2

Proof. Note that $\mathsf{H}_{\lambda_k} = \mathsf{U}_k\mathsf{U}_k^\top$, and that $\mathsf{Y}_k = \mathsf{V}_k\mathsf{U}_k$. We rewrite $\left\|\tilde{\boldsymbol{f}}_i - \tilde{\boldsymbol{f}}_j\right\|$ in a form that will let us apply the Johnson-Lindenstrauss lemma of norm conservation:

$$\begin{aligned} \left\|\tilde{\boldsymbol{f}}_i - \tilde{\boldsymbol{f}}_j\right\| &= \|\mathsf{R}^\top\mathsf{H}_{\lambda_k}^\top\mathsf{V}_k^\top(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j)\| \\ &= \|\mathsf{R}^\top\mathsf{U}_k\mathsf{U}_k^\top\mathsf{V}_k^\top(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j)\| \\ &= \|\mathsf{R}^\top\mathsf{U}_k(\boldsymbol{f}_i - \boldsymbol{f}_j)\| \end{aligned} \tag{1}$$

where the $\boldsymbol{f}_i$ are the standard SC feature vectors. Applying Theorem 1.1 of (Achlioptas, 2003) (an instance of the Johnson-Lindenstrauss lemma) to $\|\mathsf{R}^\top\mathsf{U}_k(\boldsymbol{f}_i - \boldsymbol{f}_j)\|$, the following holds. If d is larger than:

$$\frac{4+2\beta}{\epsilon^2/2 - \epsilon^3/3}\log N, \tag{2}$$

then with probability at least $1 - N^{-\beta}$, we have, $\forall(i,j) \in \{1,\ldots,N\}^2$:

$$(1-\epsilon)\,\|\mathsf{U}_k(\boldsymbol{f}_i - \boldsymbol{f}_j)\| \leqslant \tilde{D}_{ij} \leqslant (1+\epsilon)\,\|\mathsf{U}_k(\boldsymbol{f}_i - \boldsymbol{f}_j)\|\,.$$

As the columns of U_k are orthonormal, we end the proof:

$$\forall(i,j) \in [1,N]^2 \quad \|\mathsf{U}_k(\boldsymbol{f}_i - \boldsymbol{f}_j)\| = \|\boldsymbol{f}_i - \boldsymbol{f}_j\| = D_{ij}.$$

□

B. Proof of Theorem 4.1

Proof. Recall that: $\tilde{D}^r_{ij} := \left\|\tilde{\boldsymbol{f}}_{\omega_i} - \tilde{\boldsymbol{f}}_{\omega_j}\right\| = \left\|\mathsf{R}^\top\tilde{\mathsf{H}}_{\lambda_k}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|$, where $\boldsymbol{\delta}^r_{ij} = \boldsymbol{\delta}^r_i - \boldsymbol{\delta}^r_j$. Given that $\tilde{\mathsf{H}}_{\lambda_k} = \mathsf{H}_{\lambda_k} + \mathsf{E}$ and using the triangle inequality in the definition of $\tilde{D}^r_{ij}$, we obtain

$$\begin{aligned} &\left\|\mathsf{R}^\top\mathsf{H}_{\lambda_k}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| - \\ &\quad\left\|\mathsf{R}^\top\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| \leqslant \tilde{D}^r_{ij} \leqslant \left\|\mathsf{R}^\top\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| + \\ &\qquad\qquad\left\|\mathsf{R}^\top\mathsf{H}_{\lambda_k}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|, \end{aligned} \tag{3}$$

We continue the proof by bounding $\left\|\mathsf{R}^\top\mathsf{H}_{\lambda_k}^\top\mathsf{V}_k^\top\mathsf{M}^\top\,\boldsymbol{\delta}^r_{ij}\right\|$ and $\left\|\mathsf{R}^\top\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\,\boldsymbol{\delta}^r_{ij}\right\|$ separately.

Let $\delta \in]0,1]$. To bound $\left\|\mathsf{R}^\top\mathsf{H}_{\lambda_k}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|$, we set $\epsilon = \delta/2$ in Theorem 3.2. This proves that if d is larger than

$$d_0 = \frac{16(2+\beta)}{\delta^2 - \delta^3/3}\log n,$$

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

$$\left(1 - \frac{\delta}{2}\right) D^r_{ij} \leqslant \left\|\mathsf{R}^\top\mathsf{H}_{\lambda_k}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| \leqslant \left(1 + \frac{\delta}{2}\right) D^r_{ij},$$

for all $(i,j) \in \{1,\ldots,n\}^2$. To bound $\left\|\mathsf{R}^\top\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|$, we use Theorem 1.1 in (Achlioptas, 2003). This theorem proves that if $d > d_0$, then with probability at least $1-n^{-\beta}$,

$$\left\|\mathsf{R}^\top\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| \leqslant \left(1 + \frac{\delta}{2}\right)\left\|\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|,$$

for all $(i,j) \in \{1,\ldots,n\}^2$. Using the union bound and (3), we deduce that, with probability at least $1 - 2n^{-\beta}$,

$$\begin{aligned} &\left(1 - \frac{\delta}{2}\right) D^r_{ij} - \left(1 + \frac{\delta}{2}\right)\left\|\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| \\ &\qquad\leqslant \quad \tilde{D}^r_{ij} \quad \leqslant \\ &\left(1 + \frac{\delta}{2}\right)\left\|\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\| + \left(1 + \frac{\delta}{2}\right) D^r_{ij}, \end{aligned} \tag{4}$$

for all $(i,j) \in \{1,\ldots,n\}^2$ provided that $d > d_0$.

Then, as e is bounded by e_1 on the first k eigenvalues of the spectrum and by e_2 on the remaining ones, we have

$$\begin{aligned} \left\|\mathsf{E}^\top\mathsf{V}_k^\top\mathsf{M}^\top\,\boldsymbol{\delta}^r_{ij}\right\|^2 &= \left\|\mathsf{U}e(\Lambda)\mathsf{U}^\top\mathsf{V}_k^\top\mathsf{M}^\top\,\boldsymbol{\delta}^r_{ij}\right\|^2 \\ &= \left\|e(\Lambda)\mathsf{U}^\top\mathsf{V}_k^\top\mathsf{M}^\top\,\boldsymbol{\delta}^r_{ij}\right\|^2 \\ &= \sum_{l=1}^{N} e(\lambda_l)^2\,\left|(\mathsf{M}\mathsf{V}_k\boldsymbol{u}_l)^\top\boldsymbol{\delta}^r_{ij}\right|^2 \\ &\leqslant\; e_1^2\sum_{l=1}^{k}\left|(\mathsf{M}\mathsf{V}_k\boldsymbol{u}_l)^\top\boldsymbol{\delta}^r_{ij}\right|^2 \\ &\qquad + e_2^2\sum_{l=k+1}^{N}\left|(\mathsf{M}\mathsf{V}_k\boldsymbol{u}_l)^\top\boldsymbol{\delta}^r_{ij}\right|^2 \\ &= e_1^2\,\left\|\mathsf{U}_k^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|^2 \\ &\quad + e_2^2\,\left(\left\|\mathsf{U}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|^2 - \left\|\mathsf{U}_k^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|^2\right) \\ &= \;(e_1^2 - e_2^2)\,\left\|\mathsf{U}_k^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|^2 \\ &\qquad + e_2^2\,\left\|\mathsf{U}^\top\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|^2 \\ &= (e_1^2 - e_2^2)\,(D^r_{ij})^2 + e_2^2\,\left\|\mathsf{V}_k^\top\mathsf{M}^\top\boldsymbol{\delta}^r_{ij}\right\|^2 \\ &\leqslant (e_1^2 - e_2^2)\,(D^r_{ij})^2 + \frac{2\,e_2^2}{\min_i\{v_k(i)^2\}}. \end{aligned}$$

The last step follows from the fact that

$$\left\|\mathsf{V}_k^{\intercal}\mathsf{M}^{\intercal}\boldsymbol{\delta}_{ij}^r\right\|^2 = \sum_{l=1}^{N}\frac{1}{v_k(l)^2}\left|(\mathsf{M}^{\intercal}\boldsymbol{\delta}_{ij}^r)(l)\right|^2 = \frac{1}{v_k(\omega_i)^2}+\frac{1}{v_k(\omega_j)^2} \leqslant \frac{2}{\min_i\{v_k(i)\}^2}$$

Define, for all $(i,j)\in\{1,\dots,n\}^2$:

$$e_{ij} := \sqrt{|e_1^2-e_2^2|D_{ij}^r} + \frac{\sqrt{2}e_2}{\min_i\{v_k(i)\}}.$$

Thus, the above inequality may be rewritten as:

$$\left\|\mathsf{E}^{\intercal}\mathsf{V}_k^{\intercal}\mathsf{M}^{\intercal}\boldsymbol{\delta}_{ij}^r\right\| \leqslant e_{ij},$$

for all $(i,j)\in\{1,\dots,n\}^2$, which combined with (4) yields

$$\left(1-\frac{\delta}{2}\right)D_{ij}^r - \left(1+\frac{\delta}{2}\right)e_{ij} \leqslant \tilde{D}_{ij}^r \leqslant \left(1+\frac{\delta}{2}\right)e_{ij} + \left(1+\frac{\delta}{2}\right)D_{ij}^r, \quad (5)$$

for all $(i,j)\in\{1,\dots,n\}^2$, with probability at least $1-2n^{-\beta}$ provided that $d>d_0$.

Let us now separate two cases. In the case where $D_{ij}^r \geqslant D_{min}^r > 0$, we have

$$e_{ij} = \frac{e_{ij}}{D_{ij}^r}D_{ij}^r = \left(\sqrt{|e_1^2-e_2^2|} + \frac{\sqrt{2}e_2}{D_{ij}^r\,\min_i\{v_k(i)\}}\right)D_{ij}^r \leqslant \left(\sqrt{|e_1^2-e_2^2|} + \frac{\sqrt{2}e_2}{D_{min}^r\,\min_i\{v_k(i)\}}\right)D_{ij}^r \leqslant \frac{\delta}{2+\delta}D_{ij}^r.$$

provided that Eq. (7) of the main paper holds. Combining the last inequality with (5) proves the first part of the theorem.

In the case where $D_{ij}^r < D_{min}^r$, we have

$$e_{ij} < \sqrt{|e_1^2-e_2^2|D_{min}^r} + \frac{\sqrt{2}\,e_2}{\min_i\{v_k(i)\}} \leqslant \frac{\delta}{2+\delta}D_{min}^r.$$

provided that Eq. (7) of the main paper holds. Combining the last inequality with (5) terminates the proof. □

C. Experiments on the SBM with heterogeneous community sizes

We perform experiments on a SBM with $N=10^3, k=20, s=16$ and hetereogeneous community sizes. More specifically, the list of community sizes is chosen to be: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 50, 55, 60, 65, 70, 75, 80, 85, 90 and 95 nodes. In this scenario, there is no theoretical value of ϵ over which it is proven that recovery is impossible in the large N limit. Instead, we vary ϵ between 0 and 0.2 and show the recovery performance results with respect to n, d, p and γ in Fig. 1. Results are similar to the homogeneous case presented in Fig. 1(a-d) of the main paper.

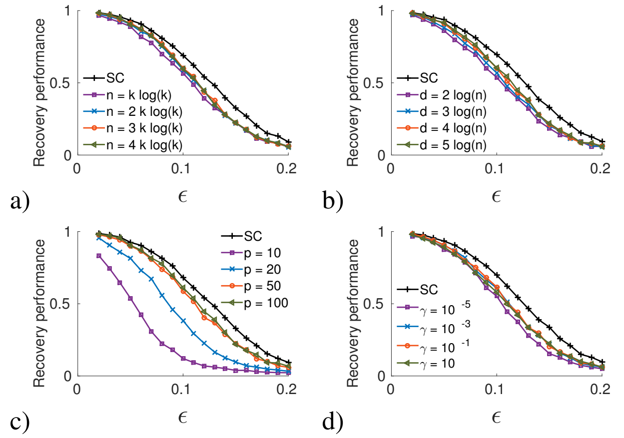

Figure 1. (a-d): recovery performance of CSC on a SBM with $N=10^3, k=20, s=16$ and hetereogeneous community sizes versus ϵ, for different n, d, p, γ. Default is $n=2k\log k$, $d=4\log n$, $p=50$ and $\gamma=10^{-3}$. All results are averaged over 20 graph realisations.

References

Achlioptas, D. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671 – 687, 2003.