Appendix: Spectral Clustering with Approximate Data

1 Appendix

In this Appendix we provide more detailed analyses and proofs that are omited in the main body of the paper due to space limitation.

Some heuristics on assumption B3 It is difficult to show that assumption B3 is valid under very general conditions. To gain insight, we here show its validity in the case where the original similarity matrix has the following block-diagonal structure (assuming there are two clusters with sizes p and q, respectively):

$$P_0 = \left[\begin{array}{cc} 1_{p\times p} & 0_{p\times q} \\ 0_{q\times p} & 1_{q\times q} \end{array} \right],$$

where $1_{p\times p}$, $1_{q\times q}$ denote matrices with all elements 1, and $0_{p\times q}$, $0_{q\times p}$ denote matrices with all elements 0. This is the case where points in the same cluster have perfect affinity and points from different clusters have no affinity.

We further assume the perturbed similarity matrix is given by

$$P_\epsilon = \begin{bmatrix} 1_{p\times p} & U(\tilde{0},\epsilon) \\ U(\tilde{0},\epsilon)^T & 1_{q\times q} \end{bmatrix},$$

where $U(\tilde{0},\epsilon)$ denotes a $p\times q$ matrix with elements generated i.i.d. according to some distribution $\mathbb{P}$ (e.g., uniform distribution) on interval $[0,\epsilon]$, with ϵ being a small constant. This model has been studied recently in [1], which obtains the following result for the unnormalized second eigenvector $\tilde{\mathbf{v}}_2$ of the Laplacian matrix of P_ϵ:

Proposition 5 *Assuming all elements ϵ_{ij} in matrix P_ϵ are i.i.d. uniform over interval $[0,\epsilon]$ for some constant $\epsilon = o(\frac{1}{p+q})$, and $\frac{p}{q} \to \alpha$ for some constant α when p and q grow. Then when p and q grow, the following holds*

$$\tilde{v}_{2k} = \left\{ \begin{array}{ll} 1-(1+\gamma)p^2[\frac{\epsilon_{k.}}{p+\epsilon_{k.}} - \mathbb{E}\frac{\epsilon_{1.}}{p+\epsilon_{1.}}] + R_{k,p,q} & k=1,...,p-1 \\ 1-(1+\gamma)(p-1)[\frac{\epsilon_{p.}}{p+\epsilon_{p.}} - \mathbb{E}\frac{\epsilon_{1.}}{p+\epsilon_{1.}}] + R_{p,p,q} & k=p \\ -\gamma+(1+\gamma)q^2[\frac{\epsilon_{.(k-p)}}{q+\epsilon_{.(k-p)}} - \mathbb{E}\frac{\epsilon_{.1}}{q+\epsilon_{.1}}] + R_{k,p,q} & k=p+1,...,p+q-1 \\ -\gamma+(1+\gamma)(q-1)[\frac{\epsilon_{.q}}{q+\epsilon_{.q}} - \mathbb{E}\frac{\epsilon_{.1}}{q+\epsilon_{.1}}] + R_{p+q,p,q} & k=p+q \end{array} \right.$$

where $\tilde{v}_{2k}$ denotes the k^{th} component of eigenvector $\tilde{\mathbf{v}}_2$, γ is a constant, $\epsilon_{i.} = \sum_{j=1}^{q}\epsilon_{ij}, i=1,...,p$, and $\epsilon_{.j} = \sum_{i=1}^{p}\epsilon_{ij}, j=1,...,q$, $R_{k,p,q}$'s are remainders with $\max_{1\le k\le p+q}|R_{k,p,q}| = o_p(1)$.

From Proposition 5, we can easily see that excluding the p^{th} and the $(p+q)^{th}$ components, all other components in $\tilde{\mathbf{v}}_2$ that belong to the same clusters follow the same distribution, up to a first order approximation (in the sense of the general matrix perturbation theory). Moreover, since the second eigenvector $\mathbf{v}_2$ of the Laplacian matrix of P_0 is piecewise constant, it immediately follows that the individual perturbations are uncorrelated with their initial values.

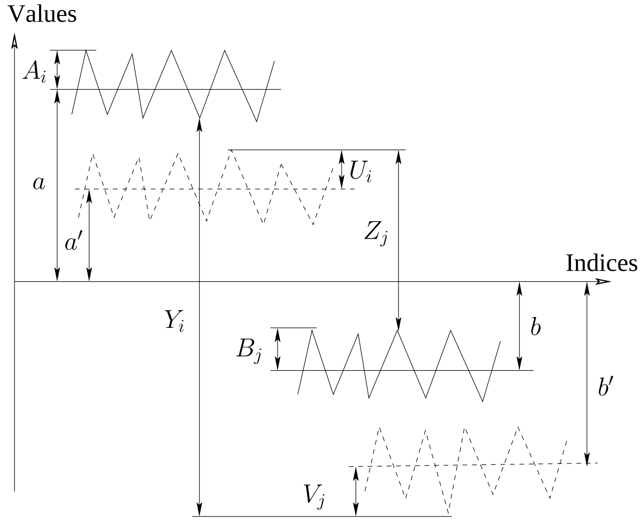

Figure 1: Notation in the Proof of Proposition 1.

maximize $\quad m = m_+ + m_-$

subject to:

$$\sum_{i=1}^{k_1}(a+A_i)^2 + \sum_{j=1}^{k_1}(b+B_j)^2 = 1 \tag{1}$$

$$\sum_{i=1}^{k_1-m_-}(a'+U_i)^2 + \sum_{j=1}^{k_2-m_+}(b'+V_j)^2 + \sum_{i=1}^{m_-}(a+A_i-Y_i)^2 + \sum_{j=1}^{m_+}(b+B_j+Z_j)^2 = 1 \tag{2}$$

$$\sum_{i=1}^{k_1-m_-}(a+A_i-a'-U_i)^2 + \sum_{j=1}^{k_2-m_+}(b+B_j-b'-V_j)^2 + \sum_{i=1}^{m_-}Y_i^2 + \sum_{j=1}^{m_+}Z_j^2 = \delta^2 \tag{3}$$

Figure 2: The optimization problem for mis-clustering rate.

Proof of Proposition 1 We use the notation introduced earlier in Section 3.1, and introduce additional notation in Fig. 1. Recall that a and b denote the sets of elements in $\mathbf{v}_2$ corresponding to two clusters under consideration, and similarly for a' and b' in $\tilde{\mathbf{v}}_2$. Let m denote the total number of missed clusterings: $m = m_- + m_+$, where m_- denotes the number of cluster flippings of $a \to b'$, and m_+ the number of flippings of $b \to a'$. As an abuse of notation, we also use a, b, a', b' to denote the mean of element values in the corresponding set in $\mathbf{v}_2$ and $\tilde{\mathbf{v}}_2$, respectively. Referring to the value δ^2 as the "energy," we aim to determine the maximum number of flippings m (i.e., missed clusterings) for any (randomly) given energy δ^2; this yields an upper bound for η.

Let $A_1, \ldots, A_{k_1}$ and $B_1, \ldots, B_{k_2}$ denote the zero mean fluctuations around a and b, respectively, and let $U_1, \ldots, U_{k_1-m_-}$ and $V_1, \ldots, V_{k_2-m_+}$ denote the zero mean fluctuations around a' and b', respectively. Let variables $Y_1, \ldots, Y_{m_-} \geq 0$ denote the long-range (random) jumps from a to b' for elements in set $a \to b'$, and let $Z_1, \ldots, Z_{m_+} \geq 0$ for elements in set $b \to a'$; Apparently, we have $Z_j = a' + U_j - b - B_j$. We have the optimization problem described in Fig. 2 for the maximum number of mis-clusterings m. In the constraints, Eqs. (1) and (2) refer to the unit length of $\mathbf{v}_2$ and $\tilde{\mathbf{v}}_2$, respectively; and Eq. (3) refers to the total energy constraint.

Without loss of generality we consider the case described in Fig. 1 with $a, a' > 0, b, b' \leq 0$, and $(a'-b)^2 < (a-b')^2$. Due to symmetry, the same argument follows when we consider the other cases. According to assumption B3, we have $\mathrm{E}Y_i = a - b'$ and $\mathrm{E}Z_j = a' - b$, for all i and j. By working with expectations, we have

$$\begin{aligned} \mathrm{E}Y_1^2 + ... + \mathrm{E}Y_{m_-}^2 &\geq \frac{1}{m_-}(\mathrm{E}Y_1 + ... + \mathrm{E}Y_{m_-})^2 = m_-(a-b')^2 \\ \mathrm{E}Z_1^2 + ... + \mathrm{E}Z_{m_+}^2 &\geq \frac{1}{m_+}(\mathrm{E}Z_1 + ... + \mathrm{E}Z_{m_Z Z+})^2 = m_+(a'-b)^2. \end{aligned}$$

Summing and substituting into Eq. (3), we get

$$m_+(a'-b)^2 + m_-(a-b')^2 \leq \mathrm{E}\sum_{i=1}^{m_-} Y_i^2 + \mathrm{E}\sum_{j=1}^{m_+} Z_j^2 \leq \delta^2 \tag{4}$$

To maximize the objective function $m = m_- + m_-$ given (4) and given $(a'-b)^2 < (a-b')^2$, we should set $m_- = 0$ and $m = m_+$. This turns (4) into $m(a'-b)^2 \leq \delta^2$, under which m achieves its maximum when $b = 0$.

Let σ_Z^2 denote the variance of Z_i, which may depend on m or n. The following arguments allow us to assume that $a'^2 = b'^2 + \mathrm{E}V_1^2 - \sigma_Z^2 - \mathrm{E}B_1^2$. The results $b = 0$ and $m = m_+$ simplify Eq. (2) into

$$\sum_{i=1}^{k_1}(a' + U_i)^2 + \sum_{j=1}^{k_2-m}(b' + V_j)^2 + \sum_{j=1}^{m}(B_j + Z_j)^2 = 1. \tag{5}$$

Taking an expectation on both sides and using the assumptions that the U_i are identically distributed with zero mean, as are the V_j, we get

$$k_1 a'^2 + k_1\mathrm{E}U_1^2 + (k_2 - m)b'^2 + (k_2 - m)\mathrm{E}V_1^2 + m(a'^2 + \sigma_Z^2) = 1, \tag{6}$$

which implies that $\sigma_Z^2 \leq \frac{1}{m}$. Hence we can assume that $\sigma_Z^2 = \frac{\beta}{m}$ for some $0 \leq \beta \leq 1$. Setting $b'^2 = a'^2 - \mathrm{E}V_1^2 + \sigma_Z^2 + \mathrm{E}B_1^2$ in Eq. (6) effectively disables this constraint for m, and Eq. (6) becomes:

$$k_1 a'^2 + k_1\mathrm{E}U_1^2 + k_2 b'^2 + k_2\mathrm{E}V_1^2 = 1.$$

Substituting $b'^2 = a'^2 - \mathrm{E}V_1^2 + \sigma_Z^2 + \mathrm{E}B_1^2$ into this equation, we obtain

$$k_1 a'^2 + k_1\mathrm{E}U_1^2 + k_2(a'^2 - \mathrm{E}V_1^2 + \sigma_Z^2) + k_2\mathrm{E}V_1^2 = na'^2 + k_2\sigma_Z^2 + k_1\mathrm{E}U_1^2 = 1,$$

from which we obtain

$$na'^2 = 1 - k_1\mathrm{E}U_1^2 - \frac{k_2\beta}{m} - k_2\mathrm{E}B_1^2. \tag{7}$$

Using $b = 0$ and $m = m_+$ also simplifies Eq. (3), which becomes:

$$\sum_{i=1}^{k_1}(a + A_i - a' - U_i)^2 + \sum_{j=1}^{k_2-m}(B_j - b' - V_j)^2 + \sum_{j=1}^{m}(a' - B_j + U_j)^2 = \delta^2.$$

Taking an expectation on both sides yields

$$k_1((a-a')^2 + \mathrm{E}A_1^2) + k_1\mathrm{E}U_1^2 + k_2\mathrm{E}B_1^2 + (k_2 - m)(b'^2 + \mathrm{E}V_1^2) + m(a'^2 + \mathrm{E}U_1^2) = \delta^2,$$

implying

$$\begin{aligned} k_1\mathrm{E}U_1^2 + k_2\mathrm{E}B_1^2 &\leq \delta^2 - \beta - ma'^2 &(8) \\ (n-m)a'^2 &\leq \frac{(n-m)(\delta^2 - \beta)}{m}. &(9) \end{aligned}$$

Substituting (8) into (7) and combining with (9), we get

$$1 - \delta^2 + \beta - \frac{k_2}{m}\beta \quad \leq \quad (n-m)a'^2 \leq \frac{(n-m)(\delta^2 - \beta)}{m}$$

Rearranging terms we have

$$\begin{aligned} m &\leq n\delta^2 - n\beta + k_2\beta \leq n\delta^2 \\ \eta = \frac{m}{n} &\leq \delta^2 = \|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2. \end{aligned}$$

When $\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$ is small, and if we further assume that all components of $\tilde{\mathbf{v}}_2 - \mathbf{v}_2$ are independent, then $\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$ is highly concentrated around its mean asymptotically in the number of data points. In this case we obtain $\eta \leq (1 + o_p(1))\mathrm{E}\|\tilde{\mathbf{v}}_2 - \mathbf{v}_2\|^2$.

Proof of Lemma 2 For the perturbation on the Laplacian matrix, we have

$$dL = I - (D+\Delta)^{-1}(K+dK) - I + D^{-1}K \tag{10}$$

Because the perturbation dK is small comparing to K, so is Δ comparing to D, and ΔD^{-1} is small. Using Taylor expansion for function $G(X) = (I+X)^{-1}$ around $X = 0_{n\times n}$, we have

$$(I + \Delta D^{-1})^{-1} = I - \Delta D^{-1} + O((\Delta D^{-1})^2).$$

Substituting it into Eq. (10), we get

$$\begin{aligned} dL &= -[I - \Delta D^{-1} + O((\Delta D)^{-2})]D^{-1}(K+dK) + D^{-1}K \\ &= -D^{-1}(K+dK) + \Delta D^{-2}(K+dK) - O((\Delta D)^{-2})D^{-1}(K+dK) + D^{-1}K \\ &= (1+o(1))\,\Delta D^{-2}K - D^{-1}dK. \end{aligned}$$

Proof of Lemma 3 Let $S_{ij} := ||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2$. For $i = j$, the result holds trivially. For $i \neq j$ and given X, $\left(S_{ij}/2\sigma^2\right)$ follows a non-central chi-square distribution with parameter (d, λ_{ij}), where $\lambda_{ij} = \left(||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2\right)$. The mean is $d + \lambda_{ij}$, the variance is $2(d + 2\lambda_{ij})$, and the moment generating function is $M_{ij}(t) = \left[\exp\left(\frac{\lambda_{ij}t}{1-2t}\right)/(1-2t)^{d/2}\right]$. $\tilde{K}_{ij}$ is an exponential function of the non-central chi-square random variable $\left(S_{ij}/2\sigma^2\right)$, so the first two moments of $\tilde{K}_{ij}$ can be computed using the moment generating function $M_{ij}(t)$, which gives the results in Eq.(15) in the main paper.

Proof of Lemma 4 Let $S_{ij} := ||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2$. For $i = j$, the result holds trivially. For $i \neq j$, S_{ij} is the sum of d (the dimension of X) independent variables, and approximately follows a Gaussian distribution for large d. We only need to work out its mean and variance. Let $\lambda_{ij} := ||\mathbf{x}_i - \mathbf{x}_j||^2$. Given input data X, we have

$$\begin{aligned} \mathrm{E}S_{ij} &= \mathrm{E}||\mathbf{x}_i - \mathbf{x}_j + \epsilon_i - \epsilon_j||^2 = \sum_{p=1}^{d}\left[(X_i^{(p)} - X_j^{(p)})^2 + 2\sigma^2\right] = \lambda_{ij} + 2d\sigma^2 \\ \mathrm{E}S_{ij}^2 &= \sum_{p=1}^{d}\sum_{q=1}^{d}\mathrm{E}\left[X_i^{(p)} - X_j^{(p)} + \epsilon_i^{(p)} - \epsilon_j^{(p)}\right]^2 \cdot \left[X_i^{(q)} - X_j^{(q)} + \epsilon_i^{(q)} - \epsilon_j^{(q)}\right]^2 \\ &= \lambda_{ij}^4 + 4(d^2 - d)\sigma^4 + d(2\mu^4 + 6\sigma^4) + 8\sigma^2\lambda_{ij}^2 + 4d\sigma^2\lambda_{ij}^2 \\ \mathrm{Var}(S_{ij}) &= \mathrm{E}S_{ij}^2 - (\mathrm{E}S_{ij})^2 = 2d\mu^4 + 2d\sigma^4 + 8\sigma^2\lambda_{ij}^2. \end{aligned}$$

So using the moment-generating function, we obtain the following asymptotic results [2]:

$$\begin{aligned} \mathrm{E}\left(\tilde{K}_{ij}\right) &= \mathrm{E}\left(\exp\left(-\frac{S_{ij}}{2\sigma_k^2}\right)\right) = M_{ij}\left(-\frac{1}{2\sigma_k^2}\right) \\ \mathrm{E}\left(\tilde{K}_{ij}^2\right) &= \mathrm{E}\left(\exp\left(-\frac{S_{ij}}{\sigma_k^2}\right)\right) = M_{ij}\left(-\frac{1}{\sigma_k^2}\right), \end{aligned}$$

which completes the proof.

References

[1] D. Yan, “Some issues with dimensionality in statistical inference,” Ph.D. dissertation, University of California, Berkeley, 2008.

[2] J. A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.