FASTER SPECTRAL CLUSTERING WITH APPROXIMATION GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral clustering is a basic algorithm in machine learning with widespread applications. However, classical spectral clustering requires computing the bottom eigenvectors of the Laplacian of the input graph, which leads to high time complexity. To address this limitation, we propose a faster clustering algorithm based on coresets and the Nyström method, and prove that it achieves the same approximation guarantee as classical spectral clustering. We empirically evaluate our algorithm on both synthetic and real-world datasets, and the results confirm that our algorithm runs substantially faster while producing results comparable to those of classical spectral clustering.

1 INTRODUCTION

Spectral clustering (Ng et al., 2001; Von Luxburg, 2007) is a fundamental algorithm in machine learning and has widespread applications in different fields including data science, social network analysis, and biology. Given a graph $G = (V_G, E_G, w_G)$ and the number of clusters k as input, the classical spectral clustering algorithm consists of two steps:

1. Compute the bottom k eigenvectors of the normalized Laplacian matrix of G.
2. Apply the k eigenvectors to construct an embedding, and run k-means on the embedded points.

Due to its simplicity and excellent performance in practice, spectral clustering has been extensively studied over the past decade from two major directions. On one side, a sequence of works has analyzed the performance of spectral clustering under various assumptions on the input graph (Dey et al., 2019; Kolev & Mehlhorn, 2016; Mizutani, 2021; Peng et al., 2017; Macgregor & Sun, 2022). On the other side, since the time complexity of spectral clustering depends on computing the k eigenvectors of the graph Laplacian matrix, researchers have developed procedures for fast approximation of these k eigenvectors (Boutsidis & Magdon-Ismail, 2013; Boutsidis et al., 2015; Choromanska et al., 2013; Macgregor, 2023). These studies of two directions raise an important problem in the field:

Is it possible to improve the time complexity of spectral clustering in practice while preserving its theoretical guarantee?

We answer this question affirmatively in the paper. Our spectral clustering algorithm is built on importance sampling and the Nyström method. Specifically, it samples a subset of vertices according to a certain probability distribution, constructs an approximate normalized Laplacian for the sampled vertices, and uses the eigenvectors of this approximate matrix to form the spectral embedding for k-means. We prove that our algorithm achieves the same approximation guarantee as the state-of-the-art method (Macgregor & Sun, 2022), and has much lower time complexity than the classical spectral clustering while running in the same order as other fast spectral clustering methods (Macgregor, 2023) (see Theorem 3.1 for formal description). The approximation and the running-time guarantees of our algorithm are further demonstrated by our experiments on both synthetic and real-world datasets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Related Work

Our work relates to a number of studies on the analysis of spectral clustering under various conditions (Dey et al., 2019; Lee et al., 2014; Mizutani, 2021; Gharan & Trevisan, 2014; Peng et al., 2017; Macgregor & Sun, 2022). Peng et al. (2017) introduces the parameter $\Upsilon(k)$, and proves that the performance of spectral clustering can be theoretically analyzed when $\Upsilon(k) = \Omega(k^3)$. After a sequence of improvements, Macgregor & Sun (2022) shows that the performance of spectral clustering can be analyzed as long as $\Upsilon(k)$ is lower bounded by a universal constant; their analysis has remained the state-of-the-art on the analysis of spectral clustering. Our result holds under the same condition as Macgregor & Sun (2022), and achieves the same approximation guarantee as theirs.

Our work also relates to the extensive literature on speeding up the running time of spectral clustering. Choromanska et al. (2013) presents a spectral clustering algorithm based on the Nyström method, hence this work is closely related to ours. However, their algorithm samples vertices uniformly at random to construct a small matrix, and their work does not present the approximation guarantee with respect to the misclassified vertices; from this aspect, our result is substantially stronger than theirs. Boutsidis et al. (2015) designs an algorithm that speeds up the computation of the eigenvectors of graph Laplacians with the power method. Improving their result, Macgregor (2023) shows that, instead of k vectors, computing $O(\log(k))$ vectors with the power method suffices to construct the embedding for spectral clustering. Compared with (Macgregor, 2023; Choromanska et al., 2013), our algorithm achieves a better approximation guarantee, which is clearly witnessed by our experiments. Finally, Jourdan et al. (2025) designs a coreset spectral clustering algorithm, and shows that coresets can be used to speed up spectral clustering.

2 Preliminaries

2.1 Notation

Let $G = (V_G, E_G, w_G)$ be an undirected weighted graph with $|V_G| = n$ vertices, $|E_G| = m$ edges, and weight function $w_G : V_G \times V_G \to \mathbb{R}_{\geq 0}$. For an edge $e = \{u, v\} \in E_G$, we denote its weight by $w_G(u, v)$ or simply $w_G(e)$. The degree of a vertex $u \in V_G$ is $d_G(u) \triangleq \sum_{v \in V_G} w_G(u, v)$, and for a subset $S \subseteq V_G$ the volume is $\mathrm{vol}_G(S) \triangleq \sum_{u \in S} d_G(u)$. Given two subsets $S, T \subseteq V_G$, the cut value between them is $w_G(S, T) \triangleq \sum_{e \in E_G(S,T)} w_G(e)$, where $E_G(S, T) = \{\{u, v\} \in E_G : u \in S, v \in T\}$. The conductance of a set $S \subseteq V_G$ is defined as

$$\Phi_G(S) \triangleq \frac{w_G(S, V_G \setminus S)}{\min\{\mathrm{vol}_G(S), \mathrm{vol}_G(V_G \setminus S)\}}$$

if $S \neq \emptyset$, and $\Phi_G(S) = 1$ otherwise. For any $S \subset V_G$, let $G[S]$ be the subgraph induced by S. For an integer $k \geq 2$, a k-way partition of V_G is a collection of disjoint subsets $A_1, \ldots, A_k$ with $\bigcup_{i=1}^{k} A_i = V_G$, and the k-way expansion of G is defined as

$$\rho_G(k) \triangleq \min_{\text{partitions } A_1, \ldots, A_k} \max_{1 \leq i \leq k} \Phi_G(A_i).$$

Our analysis is based on the spectral properties of graphs, and we list the basics of spectral graph theory. For a graph $G = (V_G, E_G, w_G)$, let $D_G \in \mathbb{R}^{n \times n}$ be the degree matrix with $D_G(u, u) = d_G(u)$, and $A_G \in \mathbb{R}^{n \times n}$ the adjacency matrix with $A_G(u, v) = w_G(u, v)$. The normalized Laplacian of G is $\mathcal{L}_G \triangleq I - N_G$, where I is the identity matrix and $N_G \triangleq D_G^{-1/2} A_G D_G^{-1/2}$. The normalized Laplacian $\mathcal{L}_G$ is symmetric and positive semidefinite, with real eigenvalues

$$0 = \lambda_1 \leq \cdots \leq \lambda_n \leq 2,$$

and an associated orthonormal basis of eigenvectors $f_1, \ldots, f_n \in \mathbb{R}^n$ Chung (1997).

Lemma 2.1 (Higher-order Cheeger inequality, Lee et al. (2014))**.** *There is an absolute constant C such that it holds for any graph G and $k \geq 2$ that*

$$\frac{\lambda_k}{2} \leq \rho_G(k) \leq C \cdot k^3 \sqrt{\lambda_k}. \tag{1}$$

2.2 SPECTRAL CLUSTERING

Spectral clustering is a popular clustering algorithm used in practice Ng et al. (2001). To analyse its performance, we examine the scenario in which there is a large gap between $\lambda_{k+1}(\mathcal{L}_G)$ and $\rho_G(k)$. By the higher-order Cheeger inequality, a small $\rho_G(k)$ yields k clusters of low-conductance, while a large $\lambda_{k+1}(\mathcal{L}_G)$ forces any $(k+1)$-partition to contain a set A with $\Phi_G(A) \geq \rho_G(k+1) \geq \lambda_{k+1}(\mathcal{L}_G)/2$. To capture this structural gap, Peng et al. (2017) defines

$$\Upsilon(k) \triangleq \frac{\lambda_{k+1}(\mathcal{L}_G)}{\rho_G(k)},$$

and proves that a lower bound on $\Upsilon(k)$ suffices to analyse the performance of spectral clustering.

2.3 CORESET NOTATION

For the kernel k-means problem, let V be a set of n points with weights $w_G : V \to \mathbb{R}_{>0}$ and kernel $K_G : V \times V \to \mathbb{R}$. Let $\varphi : V \to \mathcal{H}$ be a feature map satisfying $K_G(x,y) = \langle \varphi(x), \varphi(y) \rangle$. Let $\mathrm{dist}(x,y) \triangleq \|\varphi(x) - \varphi(y)\|_2^2$, $\mathrm{dist}(x,C) \triangleq \min_{c \in C} \mathrm{dist}(x,c)$ for $C \subseteq V$, and define the weighted kernel k-means cost as $\mathsf{COST}(V,C) \triangleq \sum_{x \in V} w_G(x)\,\mathrm{dist}(x,C)$. Importance sampling is a standard coreset technique that reduces the effective sample size by drawing points with nonuniform probabilities. Given a set of coarse centroids C and an associated partition $\{P_1, \ldots, P_k\}$ of V, one forms sensitivity scores $\sigma_x > 0$ and the sampling distribution

$$q_x \;=\; \frac{\sigma_x}{\sum_{y \in V} \sigma_y}, \quad \sum_{x \in V} q_x = 1, \quad \sigma_x \;=\; w_G(x) \left(\frac{\mathrm{dist}(x,C)}{\mathsf{COST}(V,C)} + \frac{1}{w(P_i)} \right).$$

Here $x \in P_i$ and $w(P_i) \triangleq \sum_{y \in P_i} w_G(y)$. Following Jourdan et al. (2025), the distance term is omitted before normalizing the remaining block-weight terms if $\mathsf{COST}(V,C) = 0$. In the graph setting below we take $K = D^{-1}AD^{-1} + D^{-1}$ and $w_G(x) = d_G(x)$.

2.4 THE NYSTRÖM METHOD

The Nyström method Fowlkes et al. (2004); Gittens & Mahoney (2013) approximates a positive semidefinite matrix from a subset of its columns. Given a symmetric positive semidefinite matrix $S \in \mathbb{R}^{n \times n}$ and a list $\Omega = \{i_1, \ldots, i_\ell\}$ drawn independently with replacement from a distribution q, we write $S(:,\Omega) \in \mathbb{R}^{n \times \ell}$ for the columns of S indexed by Ω and $S(\Omega,\Omega) \in \mathbb{R}^{\ell \times \ell}$ for the $\ell \times \ell$ principal submatrix indexed by Ω. We define $D_q \triangleq \mathrm{diag}\big(1/\sqrt{\ell\, q_{i_s}}\big)_{s=1}^{\ell}$ and

$$C \;\triangleq\; S(:,\Omega)\, D_q \in \mathbb{R}^{n \times \ell}, \qquad W \;\triangleq\; D_q\, S(\Omega,\Omega)\, D_q \in \mathbb{R}^{\ell \times \ell}.$$

The corresponding Nyström approximation is defined by

$$\tilde{S} \;\triangleq\; C\, W^{\dagger}\, C^{\top},$$

where $W^{\dagger}$ denotes the Moore–Penrose pseudoinverse. In particular, the top eigenpairs of $\tilde{S}$ can be recovered from an $\ell \times \ell$ eigendecomposition of a core matrix built from C and W, at cost depending on n and ℓ rather than on a dense factorization of S.

3 ALGORITHM

This section presents our designed spectral clustering algorithm with the Nyström method, which we call the SCNM algorithm. Following Macgregor & Sun (2022), we assume that the k optimal clusters $\{S_i\}_{i=1}^k$ are almost-balanced, i.e., $\mathrm{vol}(S_i) \in [\frac{1}{2k}, \frac{2}{k}] \cdot \mathrm{vol}(G)$ for every $i \in [k]$. This assumption can be dropped if the lower bound on $\Upsilon(k)$ is strengthened by an extra factor of k. The guarantee of SCNM is as follows:

Theorem 3.1 (Main Result)**.** *Let G be a graph with n vertices, and m edges. Assume that G has k clusters $\{S_i\}_{i=1}^k$ of almost-balanced size, and satisfies $\Upsilon(k) \geq C$ for a universal constant $C \in \mathbb{R}^+$. Then, there is an algorithm that runs in $O(k^4 n + k^2 m \log n + k^6)$ time and returns a partition $\{A_i\}_{i=1}^k$ of V_G. Assuming without loss of generality that the optimal counterpart of A_i is S_i, it holds that*

$$\sum_{i=1}^{k} \mathrm{vol}(A_i \triangle S_i) = O(\mathrm{vol}(G)/\Upsilon(k)).$$

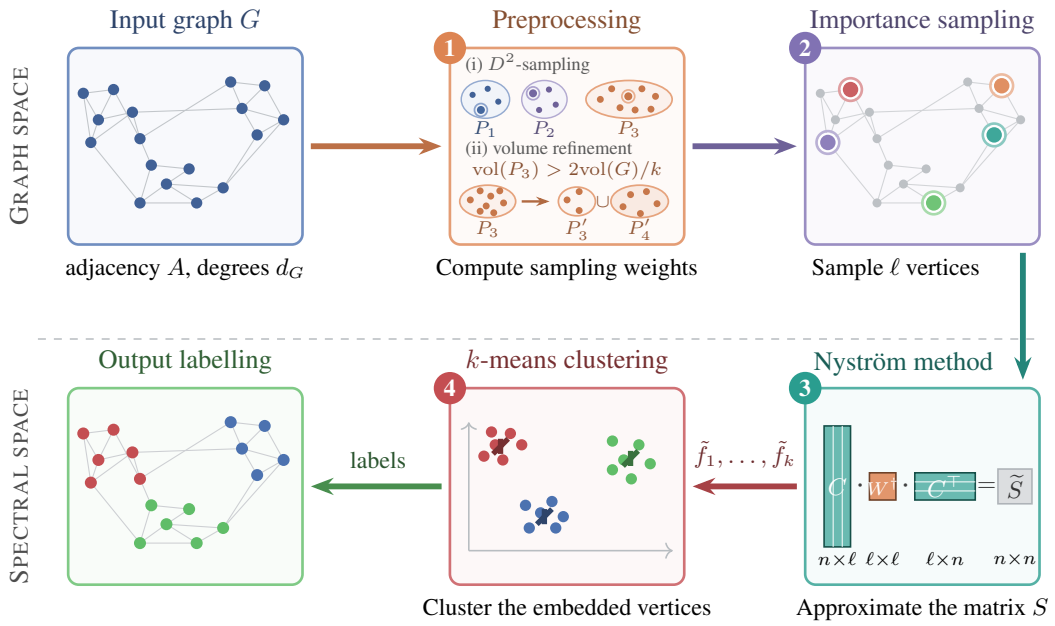

Figure 1: The pipeline of the SCNM algorithm: (1) compute the importance sampling distribution q; (2) sample ℓ vertices with respect to q; (3) apply the Nyström method to compute the approximate top k eigenvectors $\tilde{f}_1, \ldots, \tilde{f}_k$ of S; (4) run k-means on the embedded points, and return the constructed clustering as the final output.

3.1 Description of the Algorithm

We now give the formal description of the SCNM algorithm. Let $c_1, c_2 \in \mathbb{R}^+$ be two universal constants. Given any graph $G = (V_G, E_G, w_G)$, $2 \le k < n$, parameters $t = c_1 \log(n)$, and $\ell = c_2 k^2$ as input, our algorithm consists of the following steps:

1. *Importance sampling and refinement:* We run weighted D^2-sampling, and obtain a partition $\{P_i\}_{i=1}^k$ with centroids C. For each P_i with $\mathrm{vol}(P_i) > \frac{2}{k}\mathrm{vol}(G)$, order its vertices arbitrarily and split it into consecutive blocks of controlled volume. We call the resulting partition $\mathcal{P}' = \{P_i'\}$, and the associated degree-weighted sensitivity distribution q.
2. *Construction of a sketch matrix:* We draw ℓ indices $i_1, \ldots, i_\ell$ i.i.d. from q. Let $\Omega \triangleq \{i_1, \ldots, i_\ell\}$, and matrix $D_q \in \mathbb{R}^{\ell\times\ell}$ be $D_q \triangleq \mathrm{diag}\big(1/\sqrt{\ell\, q_{i_1}}, \ldots, 1/\sqrt{\ell\, q_{i_\ell}}\big)$.
3. *Nyström approximation:* We define

$$S \triangleq \left(\frac{2I - \mathcal{L}_G}{2}\right)^t, \qquad C \triangleq S(:,\Omega)D_q, \qquad W \triangleq D_q S(\Omega,\Omega)D_q,$$

   and compute

$$\tilde{S} \triangleq C\, W^\dagger\, C^\top.$$

4. *Eigen-decomposition.* Compute the top k eigen-pairs $(\hat{y}_j, \tilde{f}_j)$ of $\tilde{S}$ for $j \in [k]$.
5. *Construction of the embedding.* Let

$$\tilde{F}(u) \triangleq \frac{1}{\sqrt{d_G(u)}} \cdot (\tilde{f}_1(u), \ldots, \tilde{f}_k(u))^\top, \qquad \forall u \in V_G$$

6. *Partition V_G into k clusters.* Run an APT-approximate k-means on $\{\tilde{F}(u)\}_{u\in V_G}$, and output the partition by k-means.

The pipeline of the algorithm is illustrated in Figure 1.

3.2 Further Discussion

The main difference between the SCNM algorithm and the classical spectral clustering is the construction of the spectral embedding. For classical spectral clustering, the embedding is constructed by the bottom k eigenvectors of $\mathcal{L}_G$. While this embedding can be approximately computed in nearly-linear time (Peng et al., 2017), in practice computing the eigenvectors of a matrix usually relies on the SVD of the matrix, which takes $\Omega(n^\omega)$ time. By contrast, the spectral embedding in the SCNM algorithm is constructed with importance sampling and the Nyström method, hence avoids direct computation of $\mathcal{L}_G$'s eigenvectors. We prove that the total running time of the SCNM algorithm is $O(k^4 n + k^2 m \log n + k^6)$, which is linear in n and m. Moreover, the algorithm is highly practical, and Theorem 3.1 shows that the algorithm achieves the same theoretical guarantee as the classical spectral clustering.

We remark that, among the fast spectral clustering algorithms discussed in Section 1.1, only Macgregor (2023) analyzes the approximation guarantee with respect to misclassified vertices – the same metric as we employ. However, with the same time complexity, our algorithm achieves much stronger theoretical guarantee due to the following reason: notice that the choice of ϵ in Theorem 3.1 of Macgregor (2023) controls both the running time and the approximation ratio. In order to achieve the same running time as ours, one would need to set ϵ as a constant in Macgregor's algorithm, resulting as the total volume of the misclassified vertices as $O(\mathrm{vol}(V_G))$. By contrast, the total volume of the misclassified vertices by our algorithm is $O(\mathrm{vol}(V_G)/\Upsilon(k))$; notice that $\Upsilon(k)$ is usually of the order of $\Omega(\mathrm{poly}\log(n))$ for graphs occurring in practice.

4 Approximation Guarantee of the Algorithm

In this section we analyze the algorithm, and prove Theorem 3.1. We write $B \triangleq I + N_G = 2I - \mathcal{L}_G$ with the eigenvalues $\rho_j \triangleq (2-\lambda_j)/2$, and define the eigenvalues of $S = (B/2)^t$ to be $\mu_j \triangleq \rho_j^t$ for any $j \in [n]$. We define

$$\Delta_t \triangleq \mu_k - \mu_{k+1} = \rho_k^t - \rho_{k+1}^t = \rho_k^t\big(1-\gamma^t\big), \qquad \gamma \triangleq \frac{\rho_{k+1}}{\rho_k} \in [0,1].$$

Notice that matrices $\mathcal{L}_G$, $B/2$, and S share the same set of eigenvectors $f_1, \ldots, f_n$. Let $F \triangleq [f_1, \ldots, f_k]$, and $F_\perp \triangleq [f_{k+1}, \ldots, f_n]$. For the top k eigen-pairs $(\hat{y}_j, \tilde{f}_j)$ of $\tilde{S}$, we define $\tilde{F} \triangleq [\tilde{f}_1, \ldots, \tilde{f}_k]$, $\hat{Y} \triangleq \mathrm{diag}(\hat{y}_1, \ldots, \hat{y}_k)$, and $E \triangleq S - \tilde{S}$. Define its sampling matrix by

$$\Pi_\ell = [e_{i_1}, \ldots, e_{i_\ell}] D_q = \left[\frac{e_{i_1}}{\sqrt{\ell q_{i_1}}}, \ldots, \frac{e_{i_\ell}}{\sqrt{\ell q_{i_\ell}}}\right], \tag{2}$$

where $e_i \in \mathbb{R}^n$ is defined by $e_i(j) = 1$ if $j = i$, and $e_i(j) = 0$ otherwise. Hence, matrices C, W and $\tilde{S}$ defined in Section 2.4 can be written as

$$C = S\Pi_\ell, \qquad W = \Pi_\ell^\top S \Pi_\ell, \qquad \tilde{S} = CW^\dagger C^\top. \tag{3}$$

For the tail eigenvalues of S, write

$$\Lambda_\perp = \mathrm{diag}(\mu_{k+1}, \ldots, \mu_n), \qquad SF_\perp = F_\perp \Lambda_\perp. \tag{4}$$

4.1 Analyzing $\tilde{F}$ with respect to F

We first analyze the spectral embedding $\tilde{F}$ with respect to F. Let $S_1, \ldots, S_k$ be the optimal clusters corresponding to $\rho_G(k)$, and χ_i the indicator vector of S_i for $i \in [k]$. We define

$$\bar{g}_i \triangleq \frac{D^{1/2}\chi_i}{\sqrt{\mathrm{vol}(S_i)}}, \qquad H = [\bar{g}_1, \ldots, \bar{g}_k] \in \mathbb{R}^{n\times k}. \tag{5}$$

The following lemma shows the property of importance sampling, and upper bounds $\|H_{x,:}\|_2$

Lemma 4.1 (Properties of importance sampling)**.** *It holds for every $x \in V_G$ that*

$$q_x \geq \frac{d_G(x)}{12\mathrm{vol}(G)}, \qquad \frac{\|H_{x,:}\|_2^2}{q_x} \leq 24k. \tag{6}$$

Next, we analyze the quality of matrix $\tilde{S}$ constructed by the Nyström method, and present the condition on the number of samples under which $\|S - \tilde{S}\|$ can be bounded with respect to $\Upsilon(k)$ and ρ_j^t for $k+1 \le j \le n$.

Lemma 4.2 (Nyström approximation by importance sampling)**.** *For $\varepsilon \in (0,1]$ and $\delta \in (0,1)$, set*

$$\ell_0 = \left\lceil \max\left\{1536k^2, \frac{3072k}{\varepsilon^2}\right\}\right\rceil, \qquad b = \left\lceil \frac{\log(1/\delta)}{\log 4}\right\rceil. \tag{7}$$

Assume that $\ell \ge b\ell_0 = O((k^2 + k/\varepsilon^2)\log(1/\delta))$ indices are drawn independently from the importance-sampling distribution q. Then, it holds with probability at least $1-\delta$ that the full approximation in Equation (3) *satisfies*

$$\|E\|_2 = \|S - \tilde{S}\|_2 \le (1 + \frac{32}{\Upsilon(k)})\|\Lambda_\perp\|_2 + \varepsilon^2 \mathrm{tr}(\Lambda_\perp) \le \eta_t, \tag{8}$$

where

$$\eta_t \triangleq \left(1 + \frac{32}{\Upsilon(k)}\right)\rho_{k+1}^t + \varepsilon^2 \sum_{j=k+1}^{n} \rho_j^t. \tag{9}$$

Based on this, we analyze the difference between the embedding constructed by $\tilde{F}(\cdot)$ and $F(\cdot)$, and prove the following result.

Lemma 4.3 (Rotated identity)**.** *For every orthogonal $O \in \mathbb{R}^{k\times k}$,*

$$\sum_u d_G(u)\left\|O^\top \tilde{F}(u) - F(u)\right\|_2^2 \;=\; \|\tilde{F}O - F\|_F^2,$$

and simultaneously rotating the embedding and any associated centers by $O^\top$ preserves the weighted k-means objective, the given partition, and the approximation ratio. In particular, the sets of optimal partitions of V_G remain the same.

We remark that the orthogonal operator is only used in the analysis. For the partition and centers returned on $\{\tilde{F}(u)\}$, rotating the embedding and the centers preserves the assignment costs, and the labels of the vertices. The following example explains that an appropriate O is needed in the analysis.

Remark 4.4. Assume that G consists of $k=2$ connected components, implying $\mu_1 = \mu_2$. We take $\tilde{S} = S$, so $E = 0$. Since eigenvectors within the degenerate subspace are arbitrary up to rotation, we take the valid choice $\tilde{f}_1 = (f_1 + f_2)/\sqrt{2}$, $\tilde{f}_2 = (f_1 - f_2)/\sqrt{2}$ implying that

$$\|\tilde{F} - F\|_F^2 \;=\; \left\|\tfrac{f_1+f_2}{\sqrt{2}} - f_1\right\|^2 + \left\|\tfrac{f_1-f_2}{\sqrt{2}} - f_2\right\|^2 \;=\; 4 > 0.$$

This implies that every rotation-free per-vector or Frobenius bound fails identically in this case, but $O^* = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$ gives $\tilde{F}O^* = F$ and zero error.

Inspired by Remark 4.4, we analyze

$$\mathcal{E}^2 \;\triangleq\; \min_O \|\tilde{F}O - F\|_F^2,$$

with respect to the spectral embedding $\tilde{F}(u)$ and $F(u)$ for any $u \in V_G$.

Lemma 4.5. *Let $S, \tilde{S}$ be real and symmetric with $\|E\|_2 \le \eta_t$, $\mu_k > \mu_{k+1}$, and $\eta_t < \Delta_t$. Then, it holds that*

$$\mathcal{E}^2 \;\le\; \frac{2\|E\tilde{F}\|_F^2}{(\Delta_t - \eta_t)^2} \;\le\; \frac{2k\,\eta_t^2}{(\Delta_t - \eta_t)^2}. \tag{10}$$

Finally, we develop the connection between the number of samples needed and the desired error $\mathcal{E}^2$.

Lemma 4.6. *Suppose $0 < \gamma < 1$. For any $c \in (0, 2)$, let*

$$r(c) \triangleq \frac{\sqrt{c/2}}{1+\sqrt{c/2}} \in (0, \tfrac{1}{2}), \qquad A_0 \triangleq 1 + \frac{32}{\Upsilon(k)} + \frac{\varepsilon^2 n}{2\rho_{k+1}}.$$

If

$$t \geq T(c) \triangleq \left\lceil \frac{\ln(1 + A_0/r(c))}{\ln(1/\gamma)} \right\rceil,$$

then $\eta_t \leq r(c)\Delta_t < \Delta_t$ and $\frac{2\eta_t^2}{(\Delta_t - \eta_t)^2} \leq c$. For $0 \leq \eta_t < \Delta_t$, the latter inequality is equivalent to $\eta_t \leq r(c)\Delta_t$. In particular, $c = 1/2$ gives $r(c) = 1/3$ and the sufficient condition $\eta_t \leq \Delta_t/3$.

Lemma 4.6 implies that, for constant value of ε, setting

$$t \geq T\big(1/\Upsilon(k)\big) = \left\lceil \frac{\ln\big(1 + A_0/r(1/\Upsilon(k))\big)}{\ln(1/\gamma)} \right\rceil = O(\log n) \tag{11}$$

guarantees that

$$\mathcal{E}^2 \leq \frac{2k}{\Upsilon(k)}. \tag{12}$$

4.2 Bounding the total volume of misclassified vertices

We next upper bound the total volume of misclassified vertices of the output partition. Recall the normalized indicators $\bar{g}_i$ of the clusters in the fixed optimal partition $\{S_i\}_{i=1}^k$. Define their projections onto $\mathrm{span}(F)$, the projections of f_j onto $\mathrm{span}(H)$, and the approximate centers by

$$\hat{f}_i \triangleq \sum_{j=1}^k \langle f_j, \bar{g}_i\rangle f_j, \quad \hat{g}_j \triangleq \sum_{i=1}^k \langle f_j, \bar{g}_i\rangle \bar{g}_i, \quad p^{(i)}(j) \triangleq \frac{\langle f_j, \bar{g}_i\rangle}{\sqrt{\mathrm{vol}(S_i)}}. \tag{13}$$

For any partition $A_1, \ldots, A_k$ of V_G, we define the COST function of $A_1, \ldots, A_k$ by

$$\mathsf{COST}(A_1, \ldots, A_k) \triangleq \min_{c_1,\ldots,c_k \in \mathbb{R}^k} \sum_{i=1}^k \sum_{u \in A_i} d_G(u) \cdot \|\tilde{F}(u) - c_i\|^2.$$

The following lemma bounds the weighted k-means cost of the rotated embedding $X(u) \triangleq (O^*)^\top \tilde{F}(u)$ from Lemma 4.3, and the total volume of the output. Here, $O^* \in \arg\min_O \|\tilde{F}O - F\|_F^2$ is only used in the analysis and does not change the partition returned by k-means (cf. Remark 4.4).

Lemma 4.7. *Let $X(u) \triangleq O^{*\top}\tilde{F}(u)$ be the rotated embedding from Lemma 4.3, and $\{A_i\}_{i=1}^k$ the partition returned by an* APT*-approximate k-means algorithm. Then, it holds that*

$$\sum_{i=1}^k \sum_{u \in S_i} d_G(u) \big\|X(u) - p^{(i)}\big\|^2 \;\leq\; \frac{2k}{\Upsilon(k)} + 2\mathcal{E}^2,$$

and

$$\mathsf{COST}(A_1, \ldots, A_k) \;\leq\; \mathrm{APT} \cdot \Big(\frac{2k}{\Upsilon(k)} + 2\mathcal{E}^2\Big).$$

Moreover, if the clusters S_i are almost-balanced, then there exists a permutation σ^* of $[k]$ such that

$$\sum_{i=1}^k \mathrm{vol}\big(A_i \triangle S_{\sigma^*(i)}\big) \;\leq\; 4352\,(1 + \mathrm{APT}) \cdot \mathrm{vol}(G) \cdot \Big(\frac{1}{\Upsilon(k)} + \frac{\mathcal{E}^2}{k}\Big). \tag{14}$$

Based on Lemma 4.7, Equation (12), and applying $O(1)$-approximate k-means gives us that

$$\sum_{i=1}^k \mathrm{vol}\big(A_i \triangle S_{\sigma^*(i)}\big) = O(\mathrm{vol}(G)/\Upsilon(k)).$$

4.3 Running Time Analysis

Now we analyze the running time of the algorithm. The algorithm takes $O(nk)$ time to sample k centers, and $O(\ell)$ time to sample ℓ indices. After this, the algorithm takes $O(\ell tm)$ time to compute $W \in \mathbb{R}^{\ell \times \ell}$, and by (Li et al., 2010) takes $O(\ell^3 + n\ell^2)$ to compute the spectral embedding $\tilde{F}(\cdot)$. Since an $O(1)$-approximate k-means clustering can be computed in linear time, the overall running time of the algorithm is

$$O(nk + \ell tm + \ell^3 + n\ell^2).$$

We can choose the values of ϵ, δ in Lemma 4.2 as a small constant implying that $\ell = O(k^2)$. Combining this with Equation (11) gives the total running time of $O(k^4 n + k^2 m \log n + k^6)$.

5 Experiments

We experimentally compare the performance of our algorithm against the spectral clustering (SC) (Ng et al., 2001), the Fast Spectral Clustering (FSC) (Macgregor, 2023), the Coreset Spectral Clustering (CSC) (Jourdan et al., 2025), and the Nyström Spectral Clustering (NSC) (Choromanska et al., 2013). We use the adjusted Rand index (ARI) (Rand, 1971), the misclassification rate (Abbe, 2018), the normalized cut value (Von Luxburg, 2007), and the running time as the evaluation metrics on both real-world data and synthetic data. All experiments are conducted on a laptop with an Intel Core Ultra 7 155H processor and 32 GB of RAM; all implementations are in Python.

5.1 Results on the Synthetic Datasets

We first compare the performance of the algorithms on the stochastic block model (SBM), in which every graph is randomly generated according to parameters n, k, p, q. Specifically, every generated graph has n vertices, and k clusters of the same size; every pair of vertices is connected with probability p if they belong to the same cluster, and connected with probability q otherwise. The quality of every algorithm's output is evaluated based on the misclassification rate which is the fraction of vertices whose assigned cluster differs from the ground truth, after the best permutation of the k cluster labels.

Our experiments consist of three sets of configurations. First of all, we set $n = 12,000$, $k = 15$, $p = 0.05$, and examine the performance of the algorithms with respect to different ratios of p/q. It is known that a higher value of p/q corresponds to an easier input instance for the clustering task. Secondly, we fix the values of $k = 20$, $p = 0.05$, and $q = 0.002$, and increase the values of n. Thirdly, we fix the values of $n = 12,000$, $p = 0.15$, and $q = 0.0001$, and increase the value of k. A larger value of n or k corresponds to a harder clustering task. As reported in Figure 2, the running time of SCNM is comparable with FSC and NSC, and is much lower than SC. On the other side, the misclassification rate of the output from SCNM is very close to the one from SC, and is much lower than the ones from FSC, CSC, and NSC. One can reach the same conclusion when applying the ARI and the normalized cut as the evaluation metrics, and we refer the reader to Section C.1 for additional experimental results.

5.2 Results on the Real-World Datasets

We further evaluate the performance of the algorithms on the following six real-world datasets:

- **HAR**: the dataset consists of 10,299 labelled smartphone sensor readings capturing six human activities, each represented as 561 numerical features (Anguita et al., 2013).
- **USPS**: the dataset consists of 9,298 16×16 pixel images of handwritten digits (Hull, 1994).
- **Pendigits**: the dataset consists of 10,992 labelled instances of 2D pen movements covering the digits 0–9, each represented with 16 numerical features (Alimoglu & Alpaydin, 1997).
- **Fashion-MNIST**: the dataset consists of 70,000 28×28 pixel images of fashion items from 10 categories (Xiao et al., 2017).
- **MNIST**: the dataset consists of 70,000 28×28 pixel images of handwritten digits (Deng, 2012).

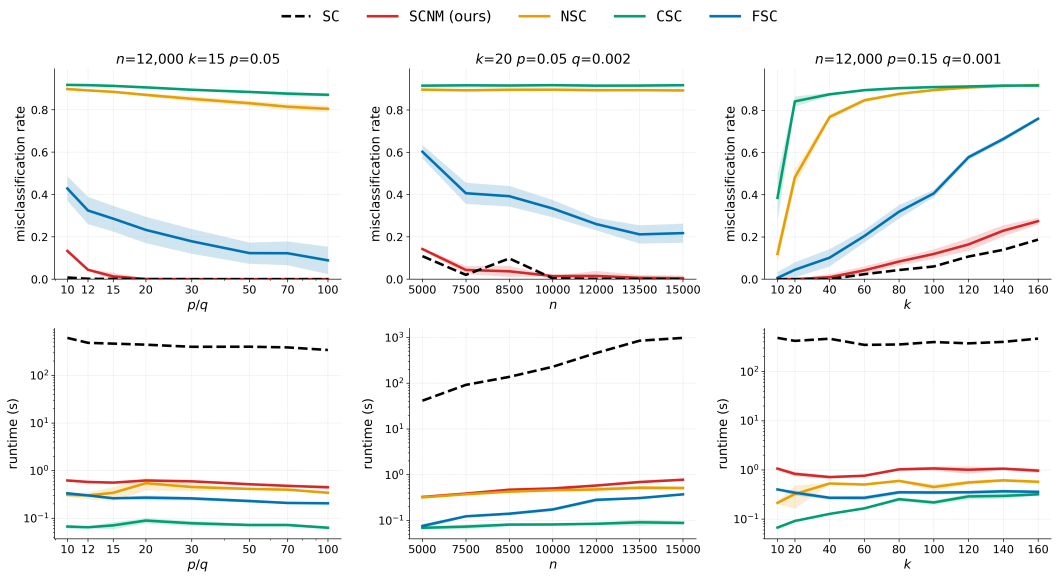

Figure 2: Misclassification rate (top) and running time (bottom) of the five algorithms on SBMs, while varying p/q, n, and k from left to right. Shaded areas indicate the standard deviation over 25 runs.

Table 1: The performance and running time of the five algorithms on the real-world datasets.

		Dataset					
	Algorithm	HAR	Pendigits	USPS	Letters	MNIST	Fashion-MNIST
ARI	SC	$\mathbf{0.52 \pm 0.00}$	$\mathbf{0.68 \pm 0.00}$	$\mathbf{0.65 \pm 0.00}$	$\mathbf{0.12 \pm 0.01}$	0.38 ± 0.00	$\mathbf{0.36 \pm 0.00}$
	SCNM	0.51 ± 0.03	0.47 ± 0.07	0.62 ± 0.02	0.11 ± 0.01	$\mathbf{0.40 \pm 0.03}$	$\mathbf{0.36 \pm 0.03}$
	NSC	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.01	$\mathbf{0.12 \pm 0.01}$	0.22 ± 0.04	0.25 ± 0.03
	FSC	0.49 ± 0.06	0.48 ± 0.07	0.53 ± 0.08	0.08 ± 0.01	0.33 ± 0.05	0.34 ± 0.03
	CSC	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.01	0.04 ± 0.01	0.08 ± 0.03	0.11 ± 0.04
Ncut	SC	$\mathbf{0.11 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.61 \pm 0.00}$	$\mathbf{7.47 \pm 0.03}$	$\mathbf{2.33 \pm 0.00}$	$\mathbf{1.22 \pm 0.00}$
	SCNM	0.23 ± 0.08	0.20 ± 0.06	0.66 ± 0.04	8.22 ± 0.26	2.35 ± 0.04	1.58 ± 0.08
	NSC	3.64 ± 0.14	5.24 ± 0.26	6.04 ± 0.18	13.03 ± 0.44	4.75 ± 0.33	3.49 ± 0.27
	FSC	0.64 ± 0.39	0.32 ± 0.07	1.40 ± 0.24	14.08 ± 0.75	4.33 ± 0.41	3.48 ± 0.47
	CSC	3.33 ± 0.09	4.90 ± 0.16	6.09 ± 0.29	13.67 ± 0.74	5.82 ± 0.39	4.54 ± 0.38
Time	SC	6.34 ± 0.34	3.23 ± 0.16	11.54 ± 0.72	213.30 ± 3.70	285.37 ± 48.17	49.11 ± 1.17
	SCNM	0.18 ± 0.01	0.16 ± 0.01	0.17 ± 0.01	0.67 ± 0.02	0.55 ± 0.04	0.67 ± 0.05
	NSC	0.08 ± 0.01	0.08 ± 0.01	0.09 ± 0.01	0.34 ± 0.03	0.26 ± 0.00	0.27 ± 0.00
	FSC	0.13 ± 0.01	0.11 ± 0.01	0.11 ± 0.01	1.23 ± 0.02	1.16 ± 0.02	1.25 ± 0.04
	CSC	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.05 \pm 0.00}$	$\mathbf{0.04 \pm 0.00}$	$\mathbf{0.09 \pm 0.01}$	$\mathbf{0.07 \pm 0.00}$	$\mathbf{0.08 \pm 0.01}$

- **Letters**: the dataset consists of 20,000 labelled instances of the 26 uppercase English letters, each represented with 16 numerical features (Frey & Slate, 1991).

All datasets can be downloaded with the `scikit-learn` library (Pedregosa et al., 2011). We evaluate every algorithm by the ARI, the misclassification rate, and the normalized cut defined by $\sum_{i=1}^{k} \frac{\mathrm{cut}(A_i, V_G \setminus A_i)}{\mathrm{vol}(A_i)}$. A higher ARI score and a lower normalized cut value represent a better clustering output.

We construct a nearest-neighbour graph from each dataset, and run the five clustering algorithms on the constructed graphs. As reported in Table 1, the performance of our algorithm is comparable with SC, and is better than FSC, CSC, and NSC. On the other side, the running time of our algorithm is much lower than SC, comparable with FSC, and higher than NSC and CSC.

In conclusion, the experiments confirm that our algorithm, while enjoying fast running time similar with the other algorithms, preserves the approximation guarantee of classical spectral clustering.

AI USE STATEMENT

In this work, we used generative AI tools to assist in improving the proof of Lemma 4.2 and to help implement the experimental code. We have not used generative AI tools to develop the research idea, to prove other lemmas or theorems, or to write the paper, and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work: the revised proof of Lemma 4.2 was checked for mathematical correctness by the authors, and the experimental code was verified and tested. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper studies a clustering algorithm and its approximation guarantee. The experiments use only publicly available benchmark datasets and synthetic graphs, and do not involve human subjects or private data. The method is intended for general-purpose clustering research and is not designed for high-stakes decision making.

REPRODUCIBILITY STATEMENT

The algorithm is specified in Section 3 and Algorithm 1. The main guarantee is stated in Theorem 3.1, and the complete proof is presented in Section 4 and the Section B. Experimental setups, datasets, and additional plots are described in the Section C. Anonymous implementation code is included in the supplementary material.

REFERENCES

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Fevzi Alimoglu and Ethem Alpaydin. Combining multiple representations and classifiers for pen-based handwritten digit recognitio. In *Proceedings of the 4th International Conference on Document Analysis and Recognition*, pp. 637–640, 1997.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pp. 3–4, 2013.

Christos Boutsidis and Malik Magdon-Ismail. Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*, 59(9):6099–6110, 2013.

Christos Boutsidis, Prabhanjan Kambadur, and Alex Gittens. Spectral clustering via the power method - provably. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 40–48, 2015.

Anna Choromanska, Tony Jebara, Hyungtae Kim, Mahesh Mohan, and Claire Monteleoni. Fast spectral clustering via the nyström method. In *Algorithmic Learning Theory - 24th International Conference*, pp. 367–381, 2013.

Fan RK Chung. *Spectral graph theory*, volume 92. 1997.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Tamal K. Dey, Pan Peng, Alfred Rossi, and Anastasios Sidiropoulos. Spectral concentration and greedy k-clustering. *Comput. Geom.*, 76:19–32, 2019.

Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nyström method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):214–225, January 2004.

Peter W. Frey and David J. Slate. Letter recognition using holland-style adaptive classifiers. *Mach. Learn.*, 6(2):161–182, March 1991.

Shayan Oveis Gharan and Luca Trevisan. Partitioning into expanders. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1256–1266, 2014.

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 567–575, 2013.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Ben Jourdan, Gregory Schwartzman, Peter Macgregor, and He Sun. Coreset spectral clustering. In *International Conference on Learning Representations*, pp. 53311–53326, 2025.

Pavel Kolev and Kurt Mehlhorn. A note on spectral clustering. In *24th Annual European Symposium on Algorithms*, volume 57, pp. 57:1–57:14, 2016.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(34):981–1006, 2012.

James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Mu Li, James Tin-Yau Kwok, and Bao-Liang Lu. Making large-scale nyström approximation possible. In *International Conference on Machine Learning*, 2010.

Peter Macgregor. Fast and simple spectral clustering in theory and practice. *Advances in Neural Information Processing Systems*, 36:34410–34425, 2023.

Peter Macgregor and He Sun. A tighter analysis of spectral clustering, and beyond. In *International conference on machine learning*, pp. 14717–14742, 2022.

Tomohiko Mizutani. Improved analysis of spectral algorithm for clustering. *Optimization Letters*, 15:1303 – 1325, 2021.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011. ISSN 1532-4435.

Richard Peng, He Sun, and Luca Zanetti. Partitioning Well-clustered graphs: Spectral clustering works! *SIAM Journal on Computing*, 46(2):710–743, 2017.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A Omitted Details from Section 3

This appendix specifies the implementation of the algorithm in Section 3.

Algorithm 1 SCNM: Spectral Clustering with Nyström Method

Require: Graph $G = (V_G, E_G, w)$ given by adjacency lists; number of clusters k with $2 \le k < n$; sketch size ℓ; filtering depth $t \ge 1$; an APT-approximate weighted k-means subroutine.
Ensure: A partition $\{A_i\}_{i=1}^{k}$ of V_G.

1: **Preprocessing:** $d_G(u) \leftarrow \sum_v w(u,v)$ $(\forall u)$, $\mathrm{vol}(G) \leftarrow \sum_u d_G(u)$; cache $d_G^{-1/2}(u)$.
2: $(C, \{P_i\}, \mathrm{dist}(x,C), \mathsf{COST}(V_G, C)) \leftarrow$ WEIGHTED-D2-SAMPLING(G,k) ▷ Algorithm 2: a coarse partition and its centroids
3: $\mathcal{P}' \leftarrow$ VOLUMEREFINE$(G, \{P_i\})$ ▷ Algorithm 3: splitting parts of large volume (Lemma B.1)
4: **Sensitivities:** for every $x \in V_G$, set $\sigma_x \leftarrow d_G(x)\big(\mathrm{dist}(x,C)/\mathsf{COST}(V_G,C) + 1/\mathrm{vol}(P'(x))\big)$ if $\mathsf{COST}(V_G,C) > 0$, and $\sigma_x \leftarrow d_G(x)/\mathrm{vol}(P'(x))$ if $\mathsf{COST}(V_G,C) = 0$; then set $Z \leftarrow \sum_{x \in V_G} \sigma_x$ and $q_x \leftarrow \sigma_x / Z$. ▷ $P'(x)$ is the refined block containing x
5: Build an alias table for q; draw $\Omega = \{i_1, \dots, i_\ell\}$ independently with replacement from q; $D_q \leftarrow \mathrm{diag}\big((\ell q_{i_s})^{-1/2}\big)_{s=1}^{\ell}$.
6: **Filtering: for** $s = 1, \dots, \ell$: $v_s \leftarrow e_{i_s}$; repeat t times $v_s \leftarrow \frac{1}{2}\big(v_s + D^{-1/2} A D^{-1/2} v_s\big)$ ▷ $v_s = S e_{i_s}$ with $S = (B/2)^t$
7: $C \leftarrow [v_1 \ \cdots \ v_\ell]\, D_q \in \mathbb{R}^{n \times \ell}$; $W = D_q S(\Omega,\Omega) D_q$; $W \leftarrow (W + W^\top)/2$
8: $(\hat{y}_j, \tilde{f}_j)_{j=1}^{k} \leftarrow$ NYSTROMEIGEN(C, W, k) ▷ Algorithm 4: top-k eigenpairs of $\tilde{S} = C W^{\dagger} C^\top$
9: $\tilde{F} \leftarrow [\tilde{f}_1 \ \cdots \ \tilde{f}_k]$; **for** $u \in V_G$: $\tilde{F}(u) \leftarrow \big(\tilde{f}_1(u), \dots, \tilde{f}_k(u)\big)^\top / \sqrt{d_G(u)}$.
10: $\{A_i\}_{i=1}^{k} \leftarrow$ APT-approximate weighted k-means on $\{\tilde{F}(u)\}_{u \in V_G}$ with weights $d_G(u)$. ▷ The stated approximation guarantee is required.
11: **return** $\{A_i\}_{i=1}^{k}$.

Algorithm 2 Procedure A: WEIGHTED-D2-SAMPLING(G,k) — a coarse partition and its centroids

1: Draw c_1 with $\Pr[x] = d_G(x)/\mathrm{vol}(G)$; $C \leftarrow \{c_1\}$; $\mathrm{dist}(x,C) \leftarrow \mathrm{dist}(x,c_1)$, $a(x) \leftarrow 1$ $(\forall x)$.
2: **for** $j = 2, \dots, k$ **do**
3: **if** $\sum_{x \in V_G} d_G(x)\mathrm{dist}(x,C) = 0$ **then**
4: **break**
5: **end if**
6: draw c_j with $\Pr[x] \propto d_G(x)\,\mathrm{dist}(x,C)$; $C \leftarrow C \cup \{c_j\}$;
7: **for all** $x \in V_G$ **do**
8: **if** $\mathrm{dist}(x,c_j) < \mathrm{dist}(x,C)$ **then** $\mathrm{dist}(x,C) \leftarrow \mathrm{dist}(x,c_j)$, $a(x) \leftarrow j$
9: **end if**
10: **end for**
11: **end for**
12: Set $P_i \leftarrow \{x \in V_G : a(x) = i\}$ for the selected centers; discard empty parts, and set $\mathsf{COST}(V_G,C) \leftarrow \sum_{x \in V_G} d_G(x)\mathrm{dist}(x,C)$.
13: **return** $(C, \{P_i\}, D, \mathsf{COST}(V_G,C))$.

Remark on Algorithm 2. After each of at most k center selections, the procedure updates $\mathrm{dist}(x,C)$ for every vertex. This uses $O(nk)$ distance evaluations and runs in $O(nk)$ time if each evaluation takes constant time.

Algorithm 3 Procedure B: VOLUMEREFINE$(G, \{P_i\})$ — splitting parts of large volume (implementation of Lemma B.1)

Discard empty parts. Set $\tau_0 \leftarrow \mathrm{vol}(G)/(2k)$ and $T \leftarrow 2\mathrm{vol}(G)/k$; compute $\mathrm{vol}(P_i)$ for all retained input parts; $\mathcal{P}' \leftarrow \emptyset$.
for each input part P_i **do**
  if $\mathrm{vol}(P_i) \le T$ **then** $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{P_i\}$
  else
    order the vertices of P_i in an arbitrary fixed order $(v_1, \dots, v_{n_i})$ (no sorting needed); $B \leftarrow \emptyset$, $s \leftarrow 0$, $R \leftarrow \mathrm{vol}(P_i)$;
    for $j = 1, \dots, n_i$ **do**
      $B \leftarrow B \cup \{v_j\}$; $s \leftarrow s + d_G(v_j)$; $R \leftarrow R - d_G(v_j)$;
      if $s \ge \tau_0$ **and** ($R = 0$ **or** $R \ge \tau_0$) **then**
        $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{B\}$; $B \leftarrow \emptyset$; $s \leftarrow 0$;
      end if
    end for
  end if
end for
return $\mathcal{P}'$.

Properties of Algorithm 3. Parts of volume at most $T = 2\mathrm{vol}(G)/k$ are kept intact; each larger part is split by a single pass over its vertices in an arbitrary order. Under the almost-balanced assumption, Lemma B.1 yields $\mathrm{vol}(P') \le 3\mathrm{vol}(G)/k$ for every $P' \in \mathcal{P}'$ and $\#\mathcal{P}' \le 3k$. The lower bound $\mathrm{vol}(P') \ge \tau_0$ holds only for blocks created by a split. No sorting is required, and the running time is $O(n)$.

Algorithm 4 Procedure C: NYSTROMEIGEN(C, W, k) — top-k eigenpairs of $\tilde{S} = CW^{\dagger}C^{\top}$

$K \leftarrow C^{\top}C \in \mathbb{R}^{\ell \times \ell}$.
Symmetric eigendecomposition $W = U\Lambda U^{\top}$ (descending order); $r \leftarrow \#\{i : \Lambda_{ii} > 0\}$; $R \leftarrow U_{[:,1:r]}\,\Lambda_{[1:r]}^{-1/2} \in \mathbb{R}^{\ell \times r}$ ▷ $RR^{\top} = W^{\dagger}$
$M \leftarrow R^{\top}KR \in \mathbb{R}^{r \times r}$; eigendecompose $M = \hat{U}\,\mathrm{diag}(\sigma_1^2, \dots, \sigma_r^2)\,\hat{U}^{\top}$ (descending order).
for $j = 1, \dots, k$ **do**
  if $j \le r$ and $\sigma_j > 0$ **then**
    $\hat{y}_j \leftarrow \sigma_j^2$; $\tilde{f}_j \leftarrow \sigma_j^{-1}CR\hat{u}_j \in \mathbb{R}^n$.
  else
    $\hat{y}_j \leftarrow 0$; choose a unit vector $\tilde{f}_j$ orthogonal to all eigenvectors already returned.
  end if
end for
return $(\hat{y}_j, \tilde{f}_j)_{j=1}^{k}$.

Correctness of Algorithm 4. Algorithm 4 uses the classical Nyström method Fowlkes et al. (2004); Kumar et al. (2012) to compute the top-k eigenvalues and eigenvectors of $\tilde{S} = CW^{\dagger}C^{\top}$ without forming this $n \times n$ matrix. Writing $RR^{\top} = W^{\dagger}$ as in the procedure, we have $\tilde{S} = (CR)(CR)^{\top}$, so the nonzero eigenvalues of $\tilde{S}$ are the eigenvalues σ_j^2 of $M = (CR)^{\top}(CR)$, with corresponding eigenvectors $\sigma_j^{-1}CR\hat{u}_j$. The procedure therefore always returns k eigenpairs, and runs in $O(n\ell^2 + \ell^3)$ time.

B OMITTED DETAILS FROM SECTION 4

This section presents the detail omitted from Section 4. We first present the properties of our refined partitioning $\mathcal{P}'$.

B.1 VOLUME REFINEMENT AND SAMPLING BOUNDS

This subsection presents the results related to volume refinement.

Lemma B.1 (Volume refinement)**.** *Let $\mathcal{P}$ be any coarse partition of V_G into at most k nonempty parts, and let $\mathcal{P}'$ be its refinement under Algorithm 3. Then, it holds for every block $P' \in \mathcal{P}'$ that*

$$\mathrm{vol}(P') \le \frac{3\mathrm{vol}(G)}{k}.$$

Moreover, the total number of blocks h in $\mathcal{P}'$ is at most $3k$.

Proof. For any optimal cluster S_i and $u \in S_i$, we have

$$d_G(u) \le \mathrm{vol}(S_i) \le \frac{2\mathrm{vol}(G)}{k},$$

and therefore $d_{\max}(G) \le 2\mathrm{vol}(G)/k$. By the rule of constructing $\mathcal{P}'$, we know that the volume of every block is at most

$$\frac{\mathrm{vol}(G)}{2k} + \max_{u \in V_G} d_G(u) + \frac{\mathrm{vol}(G)}{2k} \le \frac{3\mathrm{vol}(G)}{k}.$$

We consider the case in which a coarse cluster is split. By the algorithm description, every newly started block has available volume at least τ_0. Any positive remainder of volume less than τ_0 is absorbed into the current block. Thus every finally constructed block has volume at least τ_0.

Immediately before the current block first reaches the threshold, its volume is less than τ_0. Adding the next vertex increases its volume by at most $d_{\max}$. If the block is not closed at this point, the remaining volume is less than τ_0 and is added to the same block. Therefore, its final volume is less than

$$2\tau_0 + d_{\max} \le \frac{3\mathrm{vol}(G)}{k}.$$

A retained part already has volume at most $2\mathrm{vol}(G)/k$, so the same upper bound holds for all blocks in $\mathcal{P}'$.

The blocks formed from split parts are disjoint and each has volume at least τ_0. There are consequently at most $\mathrm{vol}(G)/\tau_0 = 2k$ blocks. Adding the at most k retained parts gives $h \le 3k$. □

Based on the properties of $\mathcal{P}'$, we prove Lemma 4.1.

Proof of Lemma 4.1. We first study the case of $\mathsf{COST}(V_G, C) > 0$. By definition,

$$\sum_{x \in V_G} \sigma_x = \frac{\sum_{x \in V_G} d_G(x)\mathrm{dist}(x, C)}{\mathsf{COST}(V_G, C)} + \sum_{P' \in \mathcal{P}'} \frac{\sum_{x \in P'} d_G(x)}{\mathrm{vol}(P')} = 1 + h.$$

Combining this with Lemma B.1 gives

$$q_x \ge \frac{d_G(x) \cdot k}{(1+h) \cdot 3\mathrm{vol}(G)} \ge \frac{k\, d_G(x)}{3(1+3k)\mathrm{vol}(G)} \ge \frac{d_G(x)}{12\mathrm{vol}(G)}.$$

Secondly, suppose that $\mathsf{COST}(V_G, C) = 0$. Then,

$$q_x = \frac{d_G(x)}{h\mathrm{vol}(P'(x))} \ge \frac{d_G(x)}{9\mathrm{vol}(G)}.$$

Combining the two cases proves the first inequality in Equation (6). By the definition of H and the almost-balanced assumption of the optimal clusters, we have

$$\frac{\|H_{x,:}\|_2^2}{q_x} \le \frac{d_G(x)/\mathrm{vol}(S_i)}{d_G(x)/(12\mathrm{vol}(G))} = \frac{12\mathrm{vol}(G)}{\mathrm{vol}(S_i)} \le 24k,$$

which proves the second inequality in Equation (6). □

In the following we use $\sigma_{\min}(A)$ to express the minimum singular value of matrix A. The next lemma bounds the Laplacian energy of H and $\|F_\perp^\top H\|_2$.

Lemma B.2. *Suppose $\rho_G(k) > 0$ and $\lambda_{k+1} > 0$. Then the following holds:*

1. $\mathrm{tr}(H^\top \mathcal{L}_G H) \le k\rho_G(k)$;
2. $H^\top \mathcal{L}_G H \preceq 2\rho_G(k) I_k$;
3. $\|F_\perp^\top H\|_F^2 \le \frac{k}{\Upsilon(k)}$, *and* $\|F_\perp^\top H\|_2^2 \le \frac{2}{\Upsilon(k)}$;
4. *If $\Upsilon(k) \ge 4$, then the minimum singular value $\sigma_{\min}(F^\top H)$ of $F^\top H$ satisfies that*

$$\sigma_{\min}(F^\top H) \ge \sqrt{1 - \frac{2}{\Upsilon(k)}} \ge \frac{1}{\sqrt{2}}. \tag{15}$$

Proof. For every $u \in \mathbb{R}^n$, the quadratic form of the normalized Laplacian is

$$u^\top \mathcal{L}_G u = \sum_{\{x,y\} \in E_G} w_G(x,y) \left(\frac{u(x)}{\sqrt{d_G(x)}} - \frac{u(y)}{\sqrt{d_G(y)}} \right)^2, \tag{16}$$

where each unordered edge is counted once. For $u = \bar{g}_i$, only edges between S_i and $V_G \setminus S_i$ contribute. Consequently,

$$\bar{g}_i^\top \mathcal{L}_G \bar{g}_i = \frac{\mathrm{cut}(S_i, V_G \setminus S_i)}{\mathrm{vol}(S_i)} = \Phi_G(S_i) \le \rho_G(k).$$

Summing over i proves the first statement.

Given $z \in \mathbb{R}^k$, the value of $(Hz)(x)/\sqrt{d_G(x)}$ is $z_i/\sqrt{\mathrm{vol}(S_i)}$ on S_i. Equation (16) and $(a-b)^2 \le 2a^2 + 2b^2$ therefore imply

$$\begin{aligned}
z^\top H^\top \mathcal{L}_G H z &= \sum_{i<j} \sum_{\substack{x \in S_i,\, y \in S_j \\ \{x,y\} \in E_G}} w_G(x,y) \left(\frac{z_i}{\sqrt{\mathrm{vol}(S_i)}} - \frac{z_j}{\sqrt{\mathrm{vol}(S_j)}} \right)^2 \\
&\le 2 \sum_{i<j} \sum_{\substack{x \in S_i,\, y \in S_j \\ \{x,y\} \in E_G}} w_G(x,y) \left(\frac{z_i^2}{\mathrm{vol}(S_i)} + \frac{z_j^2}{\mathrm{vol}(S_j)} \right) \\
&= 2 \sum_{i=1}^{k} \frac{\mathrm{cut}(S_i, V_G \setminus S_i)}{\mathrm{vol}(S_i)} z_i^2 \\
&\le 2\rho_G(k) \|z\|_2^2,
\end{aligned}$$

which proves the second statement.

Since the eigenvalues of $\mathcal{L}_G$ are nonnegative,

$$\mathcal{L}_G - \lambda_{k+1} F_\perp F_\perp^\top = \sum_{j=1}^{k} \lambda_j f_j f_j^\top + \sum_{j=k+1}^{n} (\lambda_j - \lambda_{k+1}) f_j f_j^\top \succeq 0.$$

Multiplying on the left by $H^\top$ and on the right by H yields

$$\lambda_{k+1} H^\top F_\perp F_\perp^\top H \preceq H^\top \mathcal{L}_G H. \tag{17}$$

Taking traces and applying the first statement gives

$$\lambda_{k+1} \|F_\perp^\top H\|_F^2 \le k\rho_G(k).$$

Taking the largest eigenvalue and applying the second statement similarly gives

$$\lambda_{k+1} \|F_\perp^\top H\|_2^2 \le 2\rho_G(k),$$

which proves the third statement.

Finally, we prove the last statement. If $\Upsilon(k) \geq 4$, then $2\rho_G(k) \leq \lambda_{k+1}/2 < \lambda_{k+1}$. By the fact of $FF^\top + F_\perp F_\perp^\top = I_n$, we have

$$\begin{aligned}(F^\top H)^\top (F^\top H) &= I_k - H^\top F_\perp F_\perp^\top H \\ &\succeq \left(1 - \frac{2}{\Upsilon(k)}\right) I_k \\ &\succeq \frac{1}{2} I_k.\end{aligned}$$

Taking the square root of the smallest eigenvalue proves the last statement. □

B.2 Importance sampling and bounds related to the Nyström methods

In this subsection we analyze the properties of importance sampling, and present the bounds related to the Nyström methods.

Lemma B.3 (Second-moment bound for importance sampling)**.** *Let q be a probability distribution with $q_x > 0$ and $\|H_{u,:}\|_2^2/q_u \leq 24k$ for all $u \in V_G$. For an integer $\ell_0 \geq 1$, draw ℓ_0 independent samples from q and define*

$$\Pi_0 = \left[\frac{e_{i_1}}{\sqrt{\ell_0 q_{i_1}}}, \ldots, \frac{e_{i_{\ell_0}}}{\sqrt{\ell_0 q_{i_{\ell_0}}}}\right].$$

Then, the estimate is unbiased for any fixed $Q \in \mathbb{R}^{n\times r}$; i.e.,

$$\mathbb{E}[Q^\top \Pi_0 \Pi_0^\top H] = Q^\top H.$$

Moreover, its mean-square error satisfies

$$\begin{aligned}&\mathbb{E}\left\|Q^\top \Pi_0 \Pi_0^\top H - Q^\top H\right\|_F^2 \\ &\quad = \frac{1}{\ell_0}\left(\sum_{x\in V_G} \frac{\|Q_{x,:}\|_2^2 \|H_{x,:}\|_2^2}{q_x} - \|Q^\top H\|_F^2\right) \leq \frac{24k}{\ell_0}\|Q\|_F^2. \end{aligned} \tag{18}$$

Proof. The definition of Π_0 gives

$$Q^\top \Pi_0 \Pi_0^\top H = \frac{1}{\ell_0}\sum_{a=1}^{\ell_0} \frac{Q^\top e_{i_a} e_{i_a}^\top H}{q_{i_a}}.$$

Hence, we have that

$$\mathbb{E}[Q^\top \Pi_0 \Pi_0^\top H] = \mathbb{E}_{i\sim q}\left[\frac{Q^\top e_i e_i^\top H}{q_i}\right] = \sum_{u\in V_G} Q^\top e_u e_u^\top H = Q^\top H.$$

Next, we prove the second statement. By definition, it holds that

$$\begin{aligned}&\mathbb{E}\left\|Q^\top \Pi_0 \Pi_0^\top H - Q^\top H\right\|_F^2 \\ &\quad = \frac{1}{\ell_0}\mathbb{E}_{i\sim q}\left\|\frac{Q^\top e_i e_i^\top H}{q_i} - Q^\top H\right\|_F^2 \\ &\quad = \frac{1}{\ell_0}\left(\mathbb{E}_{i\sim q}\left\|\frac{Q^\top e_i e_i^\top H}{q_i}\right\|_F^2 - \|Q^\top H\|_F^2\right).\end{aligned}$$

Hence,

$$\mathbb{E}_{i\sim q}\left\|\frac{Q^\top e_i e_i^\top H}{q_i}\right\|_F^2 = \sum_{u\in V_G} \frac{\|Q_{u,:}\|_2^2 \|H_{u,:}\|_2^2}{q_u}.$$

Combining the two equalities above proves the the equality in Equation (18). Discarding the second term and applying the condition on $\|H_{u,:}\|_2^2/q_u$ gives

$$\frac{1}{\ell_0}\sum_{u\in V_G}\frac{\|Q_{u,:}\|_2^2\|H_{u,:}\|_2^2}{q_u} \le \frac{24k}{\ell_0}\sum_{u\in V_G}\|Q_{u,:}\|_2^2 = \frac{24k}{\ell_0}\|Q\|_F^2,$$

and this proves the second statement. □

In the following lemma, we use I_n and I_k to denote the identity matrix of size n and k, respectively.

Lemma B.4 (A deterministic Nyström matrix bound)**.** *For any $\Pi \in \mathbb{R}^{n\times \ell_0}$, let*

$$\tilde{S}_\Pi = S\Pi(\Pi^\top S\Pi)^\dagger \Pi^\top S$$

and P_Π denote the orthogonal projection onto $\mathrm{range}(S^{1/2}\Pi)$*. Then, it holds that*

$$\tilde{S}_\Pi = S^{1/2}\mathsf{P}_\Pi S^{1/2}, \qquad S - \tilde{S}_\Pi = S^{1/2}(I_n - \mathsf{P}_\Pi)S^{1/2} \succeq 0. \tag{19}$$

If $J \in \mathbb{R}^{\ell_0\times k}$ satisfies $F^\top \Pi J = I_k$, then

$$\|S - \tilde{S}_\Pi\|_2 \le \|\Lambda_\perp\|_2 + \|\Lambda_\perp^{1/2}F_\perp^\top \Pi J\|_2^2. \tag{20}$$

Proof. Let $S^{1/2}\Pi = U_\Pi \Sigma_\Pi V_\Pi^\top$ be a compact singular value decomposition containing only the positive singular values. If the matrix is nonzero, Σ_Π is positive diagonal and

$$(\Pi^\top S\Pi)^\dagger = V_\Pi \Sigma_\Pi^{-2} V_\Pi^\top.$$

It follows that

$$(S^{1/2}\Pi)(\Pi^\top S\Pi)^\dagger (S^{1/2}\Pi)^\top = U_\Pi U_\Pi^\top = \mathsf{P}_\Pi.$$

Multiplying by $S^{1/2}$ on both sides proves Equation (19). Since P_Π is an orthogonal projection,

$$\|S - \tilde{S}_\Pi\|_2 = \|(I_n - \mathsf{P}_\Pi)S^{1/2}\|_2^2. \tag{21}$$

Indeed, the positive semidefinite matrix in Equation (19) is the product of the transpose of $(I_n - \mathsf{P}_\Pi)S^{1/2}$ with that matrix itself.

All columns of $S^{1/2}\Pi J F^\top$ belong to $\mathrm{range}(S^{1/2}\Pi)$. Thus,

$$\begin{aligned}(I_n - \mathsf{P}_\Pi)S^{1/2} &= (I_n - \mathsf{P}_\Pi)(S^{1/2} - S^{1/2}\Pi J F^\top),\\ \|(I_n - \mathsf{P}_\Pi)S^{1/2}\|_2 &\le \|S^{1/2} - S^{1/2}\Pi J F^\top\|_2.\end{aligned}$$

Representing the last matrix in the orthonormal eigenbasis $[F, F_\perp]$ and using $F^\top \Pi J = I_k$ gives

$$\begin{aligned}&[F, F_\perp]^\top (S^{1/2} - S^{1/2}\Pi J F^\top)[F, F_\perp]\\ &= \begin{pmatrix} 0 & 0 \\ -\Lambda_\perp^{1/2}F_\perp^\top \Pi J & \Lambda_\perp^{1/2}\end{pmatrix}.\end{aligned}$$

To see the vanishing upper-left block explicitly, write it as

$$\mathrm{diag}(\sqrt{\mu_1}, \ldots, \sqrt{\mu_k})(I_k - F^\top \Pi J) = 0.$$

The upper-right block vanishes because the two eigenspaces are orthogonal. Taking the product of the displayed block matrix with its transpose therefore shows that its squared spectral norm equals

$$\left\|\Lambda_\perp^{1/2}F_\perp^\top \Pi J J^\top \Pi^\top F_\perp \Lambda_\perp^{1/2} + \Lambda_\perp\right\|_2 \le \|\Lambda_\perp^{1/2}F_\perp^\top \Pi J\|_2^2 + \|\Lambda_\perp\|_2.$$

Combining this inequality with Equation (21) proves the second statement. □

Now we combine the importance sampling estimate in Lemma B.3 with the deterministic bound in Lemma B.4 to prove Lemma 4.2. The bound on $\|H_{x,:}\|_2^2/q_x$ follows from Lemma 4.1, and Lemma B.2 bounds $\|F_\perp^\top H\|_2$ and the smallest singular value of $F^\top H$.

Proof of Lemma 4.2. We first fix the probability distribution q. By Lemma 4.1, every q_x is positive. By Lemma B.2 and $\Upsilon(k) \geq 4$, we have

$$\vartheta \triangleq \|F_\perp^\top H\|_F^2 \leq \frac{2}{\Upsilon(k)} \leq \frac{1}{2}, \qquad \sigma_{\min}(F^\top H) \geq \frac{1}{\sqrt{2}}.$$

We will prove the the success probability for every batch of ℓ_0 samples is at least $3/4$, and then use monotonicity under the addition of sampled columns to improve the success probability to $1-\delta$. The overall proof consists of the following four steps.

1. Invertibility of the sampled matrix. For one batch, let Π_0 be as in Lemma B.3 and define

$$\mathcal{M} = F^\top \Pi_0 \Pi_0^\top H, \qquad \mathcal{M}_0 = F^\top H.$$

By the orthogonality of $[F, F_\perp]$ and Lemma B.2, we have

$$\begin{aligned}\mathcal{M}_0^\top \mathcal{M}_0 &= I_k - H^\top F_\perp F_\perp^\top H \\ &\succeq (1-\vartheta) I_k \succeq \frac{1}{2} I_k,\end{aligned}$$

implying that $\sigma_{\min}(\mathcal{M}_0) \geq 1/\sqrt{2}$. Combining Equation (18) with the fact of $Q = F$ and $\|F\|_F^2 = k$ gives us

$$\mathbb{E}\|\mathcal{M} - \mathcal{M}_0\|_F^2 \leq \frac{24k^2}{\ell_0}.$$

By $\|\cdot\|_2 \leq \|\cdot\|_F$ and Markov's inequality,

$$\mathbb{P}\left(\|\mathcal{M} - \mathcal{M}_0\|_2 > \frac{1}{2\sqrt{2}}\right) \leq \frac{24k^2/\ell_0}{(1/(2\sqrt{2}))^2} = \frac{192k^2}{\ell_0} \leq \frac{1}{8},$$

where the last inequality holds by the fact that $\ell_0 \geq 1536k^2$. Let $\mathcal{E}_1$ be the event that

$$\|\mathcal{M} - \mathcal{M}_0\|_2 \leq \frac{1}{2\sqrt{2}}.$$

Conditioned on $\mathcal{E}_1$, it holds for every unit vector $z \in \mathbb{R}^k$ that

$$\|\mathcal{M} z\|_2 \geq \|\mathcal{M}_0 z\|_2 - \|(\mathcal{M} - \mathcal{M}_0) z\|_2 \geq \frac{1}{\sqrt{2}} - \frac{1}{2\sqrt{2}} = \frac{1}{2\sqrt{2}},$$

implying that $\mathcal{M}$ is invertible and

$$\|\mathcal{M}^{-1}\|_2^2 \leq 8. \tag{22}$$

2. Error for the remaining eigenvectors. Define

$$\mathcal{N} = \Lambda_\perp^{1/2} F_\perp^\top \Pi_0 \Pi_0^\top H, \qquad \mathcal{N}_0 = \Lambda_\perp^{1/2} F_\perp^\top H.$$

Since

$$\|F_\perp \Lambda_\perp^{1/2}\|_F^2 = \mathrm{tr}(\Lambda_\perp^{1/2} F_\perp^\top F_\perp \Lambda_\perp^{1/2}) = \mathrm{tr}(\Lambda_\perp),$$

We apply Lemma B.3 with $Q = F_\perp \Lambda_\perp^{1/2}$ and have

$$\mathbb{E}\|\mathcal{N} - \mathcal{N}_0\|_F^2 \leq \frac{24k}{\ell_0} \mathrm{tr}(\Lambda_\perp).$$

If $\mathrm{tr}(\Lambda_\perp) > 0$, another application of Markov's inequality gives

$$\mathbb{P}\left(\|\mathcal{N} - \mathcal{N}_0\|_F^2 > \frac{\varepsilon^2}{16}\mathrm{tr}(\Lambda_\perp)\right) \leq \frac{(24k/\ell_0)\mathrm{tr}(\Lambda_\perp)}{(\varepsilon^2/16)\mathrm{tr}(\Lambda_\perp)} = \frac{384k}{\ell_0 \varepsilon^2} \leq \frac{1}{8},$$

where the last inequality holds by the fact that $\ell_0 \geq 3072k/\varepsilon^2$. If $\mathrm{tr}(\Lambda_\perp) = 0$, positive semidefiniteness implies $\Lambda_\perp = 0$, so $\mathcal{N} = \mathcal{N}_0 = 0$ and the event $\|\mathcal{N} - \mathcal{N}_0\|_F^2 \leq \frac{\varepsilon^2}{16}\mathrm{tr}(\Lambda_\perp)$ always holds. Hence, we define the event $\mathcal{E}_2$ as

$$\|\mathcal{N} - \mathcal{N}_0\|_F^2 \leq \frac{\varepsilon^2}{16}\mathrm{tr}(\Lambda_\perp), \tag{23}$$

and $\mathcal{E}_2$ holds with probability at least $7/8$. By Lemma B.2, we obtain

$$\|\mathcal{N}_0\|_2^2 \le \|\Lambda_\perp^{1/2}\|_2^2 \|F_\perp^\top H\|_2^2 = \vartheta \|\Lambda_\perp\|_2. \tag{24}$$

By the union bound,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - \frac{1}{8} - \frac{1}{8} = \frac{3}{4}.$$

3. Nyström error for one batch. Now we assume that both $\mathcal{E}_1$ and $\mathcal{E}_2$ occur, and set

$$J = \Pi_0^\top H \mathcal{M}^{-1}.$$

Then, we have

$$F^\top \Pi_0 J = F^\top \Pi_0 \Pi_0^\top H \mathcal{M}^{-1} = I_k,$$
$$\Lambda_\perp^{1/2} F_\perp^\top \Pi_0 J = \mathcal{N}\mathcal{M}^{-1}.$$

Moreover, Equations (22) to (24) imply

$$\begin{aligned}
\|\mathcal{N}\mathcal{M}^{-1}\|_2^2 &\le \|\mathcal{N}\|_2^2 \|\mathcal{M}^{-1}\|_2^2 \\
&\le 8\|\mathcal{N}\|_2^2 \\
&\le 16\big(\|\mathcal{N}_0\|_2^2 + \|\mathcal{N} - \mathcal{N}_0\|_2^2\big) \\
&\le 16\vartheta\|\Lambda_\perp\|_2 + 16\|\mathcal{N} - \mathcal{N}_0\|_F^2 \\
&\le 16\vartheta\|\Lambda_\perp\|_2 + \varepsilon^2 \mathrm{tr}(\Lambda_\perp),
\end{aligned}$$

where the third inequality followed by $(a+b)^2 \le 2a^2 + 2b^2$. Hence, by Lemma B.4 we have

$$\|S - \tilde{S}_{\Pi_0}\|_2 \le (1 + 16\vartheta)\|\Lambda_\perp\|_2 + \varepsilon^2 \mathrm{tr}(\Lambda_\perp) \tag{25}$$

with probability at least $3/4$.

4. Success probability for all samples. Partition the first $b\ell_0$ samples, for the purpose of analysis, into b consecutive batches of size ℓ_0. Denote the sampling matrix of batch a by

$$\Pi^{(a)} = \left[\frac{e_{i_{(a-1)\ell_0+1}}}{\sqrt{\ell_0 q_{i_{(a-1)\ell_0+1}}}}, \ldots, \frac{e_{i_{a\ell_0}}}{\sqrt{\ell_0 q_{i_{a\ell_0}}}} \right], \quad a \in [b].$$

Each column of $\Pi^{(a)}$ is $\sqrt{\ell/\ell_0}$ times the corresponding column of Π_ℓ. Since this scalar is nonzero,

$$\mathrm{range}(S^{1/2}\Pi^{(a)}) \subseteq \mathrm{range}(S^{1/2}\Pi_\ell).$$

For nested subspaces, the difference between their orthogonal projections is the orthogonal projection onto the part of the larger subspace orthogonal to the smaller one. In particular,

$$\mathsf{P}_{\Pi^{(a)}} \preceq \mathsf{P}_{\Pi_\ell}.$$

Using Equation (19) and $\tilde{S}_{\Pi_\ell} = \tilde{S}$, we obtain

$$0 \preceq S - \tilde{S} \preceq S - \tilde{S}_{\Pi^{(a)}}, \quad a \in [b]. \tag{26}$$

The spectral norm of a PSD matrix is monotone under the Loewner order. Therefore, whenever any one batch satisfies Equation (25), the approximation constructed from all ℓ samples satisfies the same bound. Since the batches consist of independent samples and each batch violates Equation (25) with probability at most $1/4$, we have

$$\begin{aligned}
&\mathbb{P}\left(\|S - \tilde{S}\|_2 > (1 + 16\vartheta)\|\Lambda_\perp\|_2 + \varepsilon^2 \mathrm{tr}(\Lambda_\perp)\right) \\
&\quad \le (1/4)^b \le \delta,
\end{aligned}$$

where the last inequality follows by the choice of b. Also, $\mu_j = \rho_j^t$ and Equation (4) give

$$\|\Lambda_\perp\|_2 = \rho_{k+1}^t, \qquad \mathrm{tr}(\Lambda_\perp) = \sum_{j=k+1}^{n} \rho_j^t.$$

Therefore, Lemma B.2 yields

$$\begin{aligned}(1+16\vartheta)\|\Lambda_\perp\|_2 + \varepsilon^2 \mathrm{tr}(\Lambda_\perp) &\le \left(1 + \frac{32}{\Upsilon(k)}\right)\rho_{k+1}^t + \varepsilon^2 \sum_{j=k+1}^{n} \rho_j^t \\ &= \eta_t.\end{aligned}$$

This proves Equation (8) conditionally on the preprocessing.

□

B.3 Proof of Lemma 4.3

We establish the weighted embedding identity in Lemma 4.3 and show that an orthogonal transformation preserves the weighted k-means objective and the set of optimal partitions of V_G.

Proof of Lemma 4.3. We first record the unrotated identity. By definition of the spectral embeddings,

$$\tilde{F}(u) - F(u) \;=\; \frac{1}{\sqrt{d_G(u)}}\big(\tilde{f}_1(u) - f_1(u), \ldots, \tilde{f}_k(u) - f_k(u)\big)^\top,$$

and therefore

$$\begin{aligned}\sum_{u \in V_G} d_G(u)\|\tilde{F}(u) - F(u)\|_2^2 &= \sum_{u \in V_G}\sum_{j=1}^{k}\big(\tilde{f}_j(u) - f_j(u)\big)^2 \\ &= \sum_{j=1}^{k}\|\tilde{f}_j - f_j\|_2^2 \\ &= \|\tilde{F} - F\|_F^2.\end{aligned}$$

For the rotated claim, observe that column j of $\tilde{F}O$ has entry $\sum_{\ell=1}^{k}\tilde{f}_\ell(u)\,O_{\ell j}$ at vertex u, so the induced embedding is $u \mapsto O^\top\tilde{F}(u)$. Substituting this embedding for $u \mapsto \tilde{F}(u)$ in the identity above yields

$$\sum_{u \in V_G} d_G(u)\big\|O^\top\tilde{F}(u) - F(u)\big\|_2^2 \;=\; \|\tilde{F}O - F\|_F^2.$$

Finally, since O is orthogonal, distances are preserved:

$$\big\|O^\top\tilde{F}(u) - O^\top c\big\|_2^2 \;=\; \big(\tilde{F}(u) - c\big)^\top OO^\top\big(\tilde{F}(u) - c\big) \;=\; \big\|\tilde{F}(u) - c\big\|_2^2.$$

For any fixed partition of V_G, rotating its centers with the embedding therefore preserves every assignment cost. The map $c \mapsto O^\top c$ is bijective, so minimizing over centers preserves the cost of each partition. Thus the sets of optimal partitions of V_G coincide, and any approximation ratio for a fixed output partition remains the same. □

B.4 Proof of Lemma 4.5

We prove Lemma 4.5 by bounding the eigenvalue separation, the eigenvector residual, and the principal angles between the two leading eigenspaces. Throughout this subsection, we work under its assumptions $\|E\|_2 \le \eta_t$, $\mu_k > \mu_{k+1}$, and $\eta_t < \Delta_t$. We write the positive separation as $\Delta_t - \eta_t$ throughout, reserving δ for the failure probability in Lemma 4.2.

Lemma B.5 (Weyl's inequality Weyl (1912)). *For any symmetric matrices $A, B \in \mathbb{R}^{n\times n}$ whose eigenvalues are indexed in the same order, it holds for any $j \in [n]$ that $|\lambda_j(A)-\lambda_j(B)| \le \|A-B\|_2$.*

Lemma B.6. *Assume $\|E\|_2 \le \eta_t$, $\mu_k > \mu_{k+1}$, and $\eta_t < \Delta_t$. Then, it holds for any $i \in [k]$ that $|\hat{y}_i - \mu_i| \le \eta_t$. Consequently, it holds for any $1 \le i \le k < j \le n$ that*

$$\hat{y}_i - \mu_j \ge \Delta_t - \eta_t > 0.$$

Proof. By Lemma B.5 and the condition on E, it holds for any $i \in [k]$ that

$$|\hat{y}_i - \mu_i| \le \|S - \tilde{S}\|_2 = \|E\|_2 \le \eta_t.$$

Both sequences of eigenvalues are ordered decreasingly. Thus, for $1 \le i \le k < j \le n$,

$$\hat{y}_i - \mu_j \ge \hat{y}_k - \mu_{k+1} \ge (\mu_k - \eta_t) - \mu_{k+1} = \Delta_t - \eta_t > 0.$$

The strict positivity is precisely the assumption $\eta_t < \Delta_t$. □

Lemma B.7. *The matrix $R \triangleq S\tilde{F} - \tilde{F}\hat{Y} = E\tilde{F}$ satisfies that $\|R\|_F \le \eta_t\sqrt{k}$.*

Proof. By spectral decomposition we have $\tilde{S}\tilde{F} = \tilde{F}\hat{Y}$, so $R = (S - \tilde{S})\tilde{F} = E\tilde{F}$. Hence,

$$\|R\|_F^2 = \|E\tilde{F}\|_F^2 \le \|E\|_2^2\|\tilde{F}\|_F^2 \le \eta_t^2 \cdot k,$$

which proves the lemma. □

Lemma B.8. *Let $Z \triangleq F_\perp^\top \tilde{F} \in \mathbb{R}^{(n-k)\times k}$, and $R \triangleq E\tilde{F}$. Then, it holds that*

$$\|\sin\Theta(\tilde{F}, F)\|_F = \|Z\|_F \le \frac{\|R\|_F}{\Delta_t - \eta_t}.$$

Proof. We first prove $\|\sin\Theta\|_F = \|Z\|_F$. Since $[F\ F_\perp]$ is square and orthogonal, it holds that $FF^\top + F_\perp F_\perp^\top = I$ and

$$\|\tilde{F}\|_F^2 = \mathrm{tr}(\tilde{F}^\top\tilde{F}) = \mathrm{tr}\big(\tilde{F}^\top(FF^\top + F_\perp F_\perp^\top)\tilde{F}\big) = \|F^\top\tilde{F}\|_F^2 + \|F_\perp^\top\tilde{F}\|_F^2.$$

Since $\|\tilde{F}\|_F^2 = k$ and $Z = F_\perp^\top\tilde{F}$, we have $\|Z\|_F^2 = k - \|F^\top\tilde{F}\|_F^2$. Moreover, the singular values of $F^\top\tilde{F}$ are the cosines of the principal angles between $\mathrm{span}(F)$ and $\mathrm{span}(\tilde{F})$, so $\|F^\top\tilde{F}\|_F^2 = \sum_i \cos^2\theta_i$ and hence $\|Z\|_F = \|\sin\Theta\|_F$.

Secondly, we prove $\|Z\|_F \le \frac{\|R\|_F}{\Delta_t - \eta_t}$. By Equation (4) and the symmetry of S we have $F_\perp^\top S = \Lambda_\perp F_\perp^\top$. Left-multiplying the residual identity in Lemma B.7 by $F_\perp^\top$, we obtain

$$F_\perp^\top R = F_\perp^\top S\tilde{F} - F_\perp^\top\tilde{F}\hat{Y} = \Lambda_\perp Z - Z\hat{Y}, \qquad \text{i.e.} \qquad (\mu_{k+j} - \hat{y}_i)\, Z_{ji} = (F_\perp^\top R)_{ji}.$$

By Lemma B.6, $|\mu_{k+j} - \hat{y}_i| \ge \Delta_t - \eta_t > 0$, hence $|Z_{ji}| \le |(F_\perp^\top R)_{ji}|/(\Delta_t - \eta_t)$. Squaring and summing over all the indices i, j gives us that

$$\|Z\|_F^2 \le \frac{\|F_\perp^\top R\|_F^2}{(\Delta_t - \eta_t)^2}.$$

For each column of R, we have

$$\big\|F_\perp^\top R_{:,j}\big\|_2^2 = R_{:,j}^\top\big(F_\perp F_\perp^\top\big)R_{:,j} \le \big\|R_{:,j}\big\|_2^2, \qquad j \in [k],$$

and summing over the k columns gives $\|F_\perp^\top R\|_F^2 \le \|R\|_F^2$. □

Lemma B.9. *It holds that $\min_O \|\tilde{F}O - F\|_F^2 \le 2\|\sin\Theta(\tilde{F}, F)\|_F^2$.*

Proof. By the property of Frobenius norm, we have

$$\|\tilde{F}O - F\|_F^2 = \|\tilde{F}\|_F^2 + \|F\|_F^2 - 2\langle\tilde{F}O, F\rangle = 2k - 2\mathrm{tr}(O^\top\tilde{F}^\top F).$$

Let $\tilde{F}^\top F = U\Sigma V^\top$ be the SVD with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, $\sigma_i = \cos\theta_i \in [0,1]$. For every orthogonal O, it holds that

$$\operatorname{tr}(O^\top U\Sigma V^\top) \;=\; \operatorname{tr}(\Sigma V^\top O^\top U) \;=\; \sum_i \sigma_i \cdot (V^\top O^\top U)_{ii} \;\le\; \sum_i \sigma_i,$$

since the diagonal entries of an orthogonal matrix are at most 1 and $\sigma_i \ge 0$. With $O^* = UV^\top$, we have

$$\operatorname{tr}(O^{*\top} U\Sigma V^\top) \;=\; \operatorname{tr}(VU^\top U\Sigma V^\top) \;=\; \operatorname{tr}(\Sigma).$$

Hence, $\min_O \|\tilde{F}O - F\|_F^2 = 2k - 2\sum_i \sigma_i$. Since $\sigma_i \le 1$, we have $\sigma_i \ge \sigma_i^2$. Moreover, by the first part of the proof of Lemma B.8 we have

$$\sum_i \sigma_i \;\ge\; \sum_i \sigma_i^2 \;=\; \|\tilde{F}^\top F\|_F^2 \;=\; k - \|\sin\Theta(\tilde{F}, F)\|_F^2.$$

Combining the equalities above proves $\min_O \|\tilde{F}O - F\|_F^2 \le 2\|\sin\Theta(\tilde{F}, F)\|_F^2$. □

Proof of Lemma 4.5. It holds that

$$\begin{aligned}
\mathcal{E}^2 &= \min_O \|\tilde{F}O - F\|_F^2 \\
&\le 2\|\sin\Theta(\tilde{F}, F)\|_F^2 \\
&\le \frac{2\|R\|_F^2}{(\Delta_t - \eta_t)^2} \\
&\le \frac{2k\eta_t^2}{(\Delta_t - \eta_t)^2},
\end{aligned}$$

where the first inequality holds by Lemma B.9, the second one holds by Lemma B.8, and the third one holds by Lemma B.7. □

B.5 Proof of Lemma 4.6

Proof of Lemma 4.6. We assume that $0 < \rho_{k+1} < \rho_k$, so $0 < \gamma < 1$. For every integer $t \ge 1$, monotonicity of the eigenvalues gives

$$\sum_{j=k+1}^{n} \rho_j^t = \sum_{j=k+1}^{n} \rho_j \rho_j^{t-1} \le \rho_{k+1}^{t-1} \sum_{j=k+1}^{n} \rho_j \le \frac{n}{2}\rho_{k+1}^{t-1}.$$

Combining it with Equation (9) yields

$$\begin{aligned}
\eta_t &= \left(1 + \frac{32}{\Upsilon(k)}\right)\rho_{k+1}^t + \varepsilon^2 \sum_{j=k+1}^{n} \rho_j^t \\
&\le \left(1 + \frac{32}{\Upsilon(k)} + \frac{\varepsilon^2 n}{2\rho_{k+1}}\right)\rho_{k+1}^t = A_0\rho_{k+1}^t. \qquad (27)
\end{aligned}$$

Since $\Delta_t = \rho_k^t(1 - \gamma^t) > 0$, we have

$$\frac{\eta_t}{\Delta_t} \le \frac{A_0\rho_{k+1}^t}{\rho_k^t(1-\gamma^t)} = \frac{A_0\gamma^t}{1-\gamma^t}.$$

The condition $t \ge T(c)$ of the lemma gives

$$t\ln(1/\gamma) \ge \ln(1 + A_0/r(c)),$$

which implies

$$\gamma^t \le \frac{1}{1 + A_0/r(c)} = \frac{r(c)}{A_0 + r(c)}.$$

Because $1-\gamma^t > 0$, this is equivalent to $A_0\gamma^t \le r(c)(1-\gamma^t)$. Consequently,

$$\frac{\eta_t}{\Delta_t} \le \frac{A_0\gamma^t}{1-\gamma^t} \le r(c) < 1,$$

and $\eta_t \le r(c)\Delta_t < \Delta_t$, which proves the first inequality. In particular, the condition $\eta_t < \Delta_t$ in Lemma 4.5 holds. To prove the second inequality, notice that

$$\begin{aligned} \frac{2\eta_t^2}{(\Delta_t-\eta_t)^2} \le c \quad &\iff \quad \frac{\eta_t}{\Delta_t-\eta_t} \le \sqrt{c/2} \\ &\iff \quad (1+\sqrt{c/2})\eta_t \le \sqrt{c/2}\,\Delta_t \\ &\iff \quad \eta_t \le r(c)\Delta_t. \end{aligned}$$

For $c = 1/2$, we have $r(c) = 1/3$. □

B.6 Proof of Lemma 4.7

The following result will be used in our analysis.

Lemma B.10 (The Stronger Structure Theorem, Macgregor & Sun (2022))**.** *For every $i \in [k]$, the following bounds hold:*

$$\left\|\bar{g}_i - \hat{f}_i\right\|^2 \;\le\; \frac{1}{\Upsilon(k)}, \qquad \sum_{j=1}^{k}\left\|f_j - \hat{g}_j\right\|^2 \;\le\; \frac{k}{\Upsilon(k)}.$$

Proof of Lemma 4.7. For $u \in S_i$, the definitions in Equation (13) and the disjoint supports of the indicators imply

$$\hat{g}_j(u) = \langle f_j, \bar{g}_i\rangle \frac{\sqrt{d_G(u)}}{\sqrt{\mathrm{vol}(S_i)}} = \sqrt{d_G(u)}\, p^{(i)}(j).$$

Hence,

$$\begin{aligned} \sum_{i=1}^{k}\sum_{u\in S_i} d_G(u)\|F(u)-p^{(i)}\|^2 &= \sum_{i=1}^{k}\sum_{u\in S_i}\sum_{j=1}^{k}\big(f_j(u)-\hat{g}_j(u)\big)^2 \\ &= \sum_{j=1}^{k}\|f_j-\hat{g}_j\|_2^2 \le \frac{k}{\Upsilon(k)}, \end{aligned}$$

where the last inequality follows from Lemma B.10. By Lemma 4.3, it holds for every orthogonal $O \in \mathbb{R}^{k\times k}$ that

$$\sum_u d_G(u)\left\|O^\top \tilde{F}(u) - F(u)\right\|^2 \;=\; \|\tilde{F}O - F\|_F^2.$$

By the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ we have

$$\begin{aligned} &\sum_{i=1}^{k}\sum_{u\in S_i} d_G(u)\|X(u)-p^{(i)}\|^2 \\ &\quad\le 2\sum_{u\in V_G} d_G(u)\|X(u)-F(u)\|^2 + 2\sum_{i=1}^{k}\sum_{u\in S_i} d_G(u)\|F(u)-p^{(i)}\|^2 \\ &\quad\le 2\mathcal{E}^2 + \frac{2k}{\Upsilon(k)}. \end{aligned}$$

Rotating the embedding and the output centers together preserves the approximation guarantee by Lemma 4.3, without changing the output partition. Hence,

$$\mathsf{COST}(A_1,\ldots,A_k) \le \mathrm{APT}\left(2\mathcal{E}^2 + \frac{2k}{\Upsilon(k)}\right).$$

Then, we follow the same analysis as Macgregor & Sun (2022) and obtain the claimed statement. □

C EXPERIMENTAL RESULTS

This appendix reports the remaining plots from Subsections 5.1 and 5.2. Subsection C.1 gives the remaining ARI and normalized cut values on the SBM as p/q, n, and k vary. Subsection C.2 presents the additional results on HAR, Letters, MNIST, Fashion-MNIST, Pendigits, and USPS, as well as the running times of the algorithms on all datasets.

C.1 ADDITIONAL RESULTS ON THE SYNTHETIC DATASETS

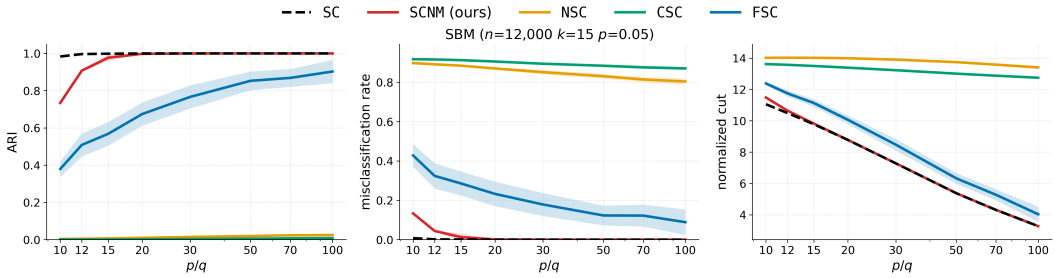

Figure 3: Comparison of the ARI, misclassification rate, and normalized cut versus p/q on an SBM with $n = 12{,}000$, $k = 15$, and $p = 0.05$. Shaded areas indicate the standard deviation over 25 runs.

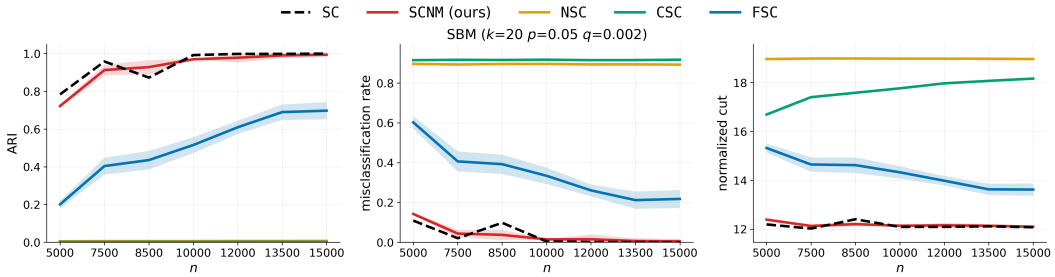

Figure 4: Comparison of the ARI, misclassification rate, and normalized cut versus n on an SBM with $k = 20$, $p = 0.05$, and $q = 0.002$. Shaded areas indicate the standard deviation over 25 runs.

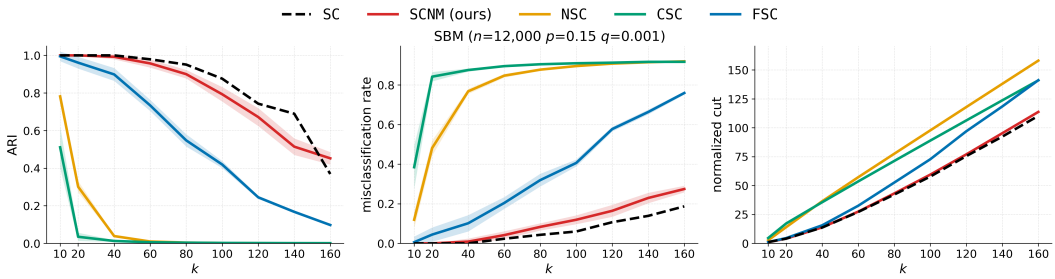

Figure 5: Comparison of the ARI, misclassification rate, and normalized cut versus k on an SBM with $n = 12{,}000$, $p = 0.15$, and $q = 0.0001$. Shaded areas indicate the standard deviation over 25 runs.

C.2 Additional Results on the Real-World Datasets

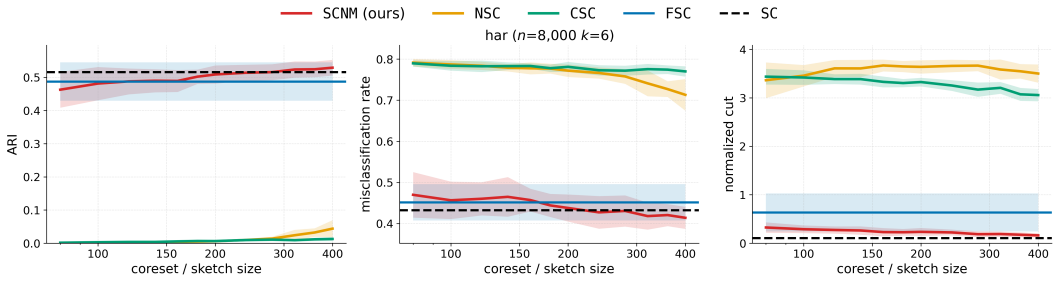

Figure 6: Comparison of the ARI, misclassification rate, and normalized cut for the five algorithms on a graph with 20 neighbors and 8,000 points from the HAR dataset. Shaded areas indicate the standard deviation of the metric over 25 runs.

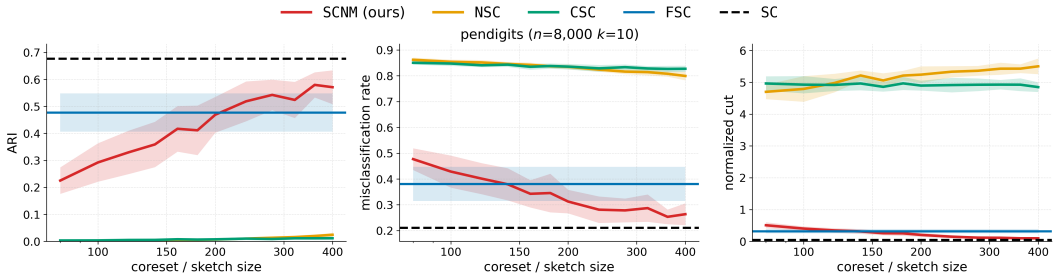

Figure 7: Comparison of the ARI, misclassification rate, and normalized cut for the five algorithms on a graph with 20 neighbors and 8,000 points from the Pendigits dataset. Shaded areas indicate the standard deviation of the metric over 25 runs.

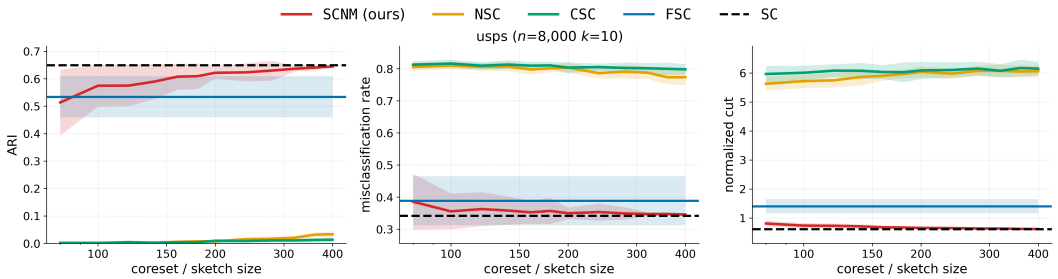

Figure 8: Comparison of the ARI, misclassification rate, and normalized cut for the five algorithms on a graph with 20 neighbors and 8,000 points from the USPS dataset. Shaded areas indicate the standard deviation of the metric over 25 runs.

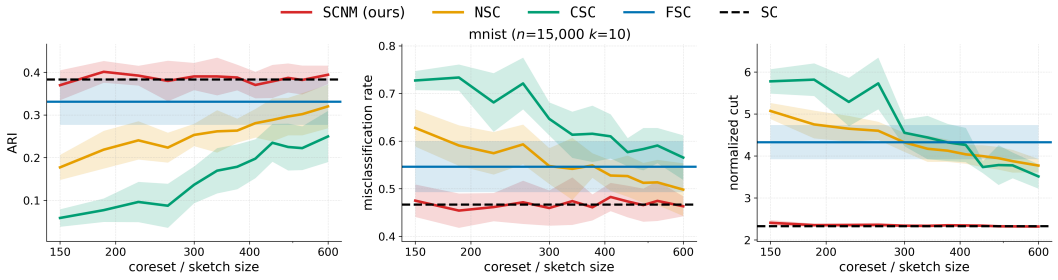

Figure 9: Comparison of the ARI, misclassification rate, and normalized cut for the five algorithms on a graph with 200 neighbors and 15,000 points from the MNIST dataset. Shaded areas indicate the standard deviation of the metric over 10 runs.

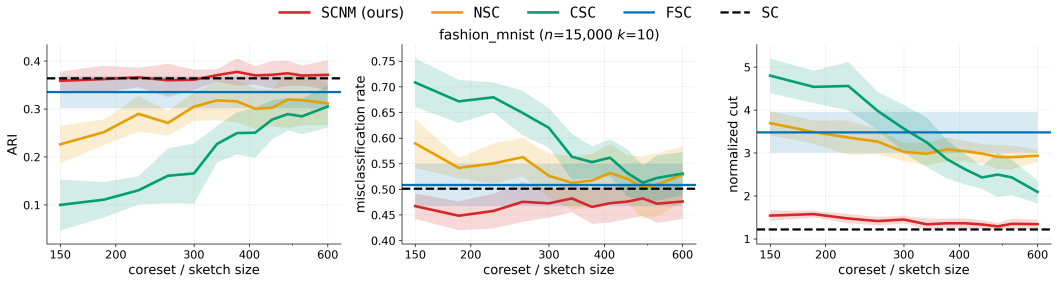

Figure 10: Comparison of the ARI, misclassification rate, and normalized cut for the five algorithms on a graph with 200 neighbors and 15,000 points from the Fashion-MNIST dataset. Shaded areas indicate the standard deviation of the metric over 10 runs.

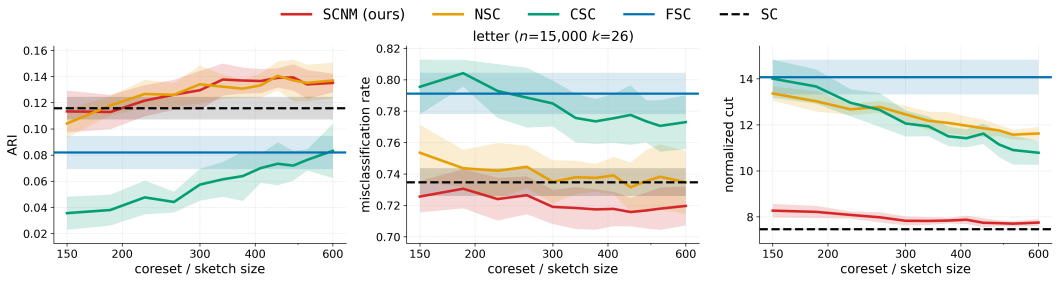

Figure 11: Comparison of the ARI, misclassification rate, and normalized cut for the five algorithms on a graph with 200 neighbors and 15,000 points from the Letters dataset. Shaded areas indicate the standard deviation of the metric over 10 runs.

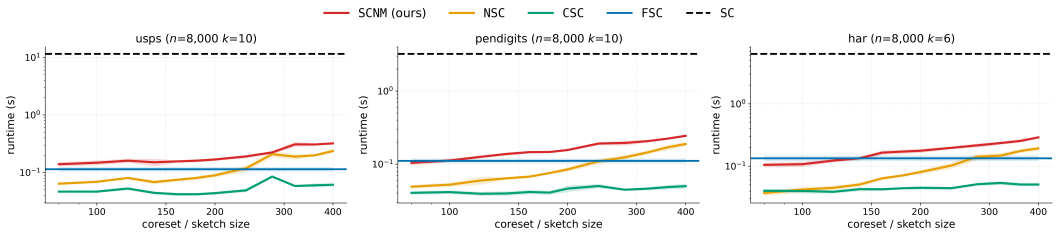

Figure 12: Comparison of the running times for the five algorithms on sparse graphs from USPS ($n = 8{,}000$, $k = 10$), Pendigits ($n = 8{,}000$, $k = 10$), and HAR ($n = 8{,}000$, $k = 6$). Each setting is repeated 25 times.

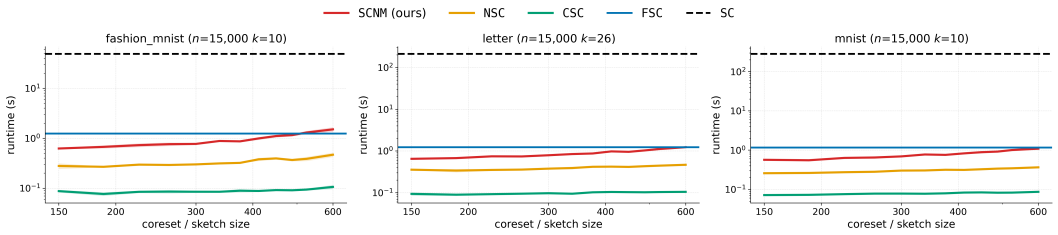

Figure 13: Comparison of the running times for the five algorithms on dense graphs from Fashion-MNIST ($n = 15{,}000$, $k = 10$), Letters ($n = 15{,}000$, $k = 26$), and MNIST ($n = 15{,}000$, $k = 10$). Each setting is repeated 10 times.