PARTITIONING WELL-CLUSTERED GRAPHS: SPECTRAL CLUSTERING WORKS!*

RICHARD PENG†, HE SUN‡, AND LUCA ZANETTI‡

Abstract. In this paper we study variants of the widely used *spectral clustering* that partitions a graph into k clusters by (1) embedding the vertices of a graph into a low-dimensional space using the bottom eigenvectors of the Laplacian matrix and (2) grouping the embedded points into k clusters via k-means algorithms. We show that, for a wide class of graphs, spectral clustering gives a good approximation of the optimal clustering. While this approach was proposed in the early 1990s and has comprehensive applications, prior to our work similar results were known only for graphs generated from stochastic models. We also give a nearly linear time algorithm for partitioning well-clustered graphs based on computing a matrix exponential and approximate nearest neighbor data structures.

Key words. graph partitioning, spectral clustering, k-means, heat kernel

AMS subject classifications. 68W20, 68W25

DOI. 10.1137/15M1047209

1. Introduction. Partitioning a graph into two or more pieces is one of the most fundamental problems in combinatorial optimization and has comprehensive applications in various disciplines of computer science.

One of the most studied graph partitioning problems is the *edge expansion problem*, i.e., finding a cut with few crossing edges normalized by the size of the smaller side of the cut. Formally, let $G = (V, E)$ be an undirected graph. For any set S, the conductance of set S is defined by

$$\phi_G(S) \triangleq \frac{|E(S, V \setminus S)|}{\mathrm{vol}(S)},$$

where $\mathrm{vol}(S)$ is the total weight of edges incident to vertices in S, and let the conductance of G be

$$\phi(G) \triangleq \min_{S: \mathrm{vol}(S) \leqslant \mathrm{vol}(G)/2} \phi_G(S).$$

The edge expansion problem asks for a set $S \subseteq V$ of $\mathrm{vol}(S) \leqslant \mathrm{vol}(V)/2$ such that $\phi_G(S) = \phi(G)$. This problem is known to be NP-hard [26], and the current best approximation algorithm achieves an approximation ratio of $O(\sqrt{\log n})$ [5].

The *k-way partitioning problem* is a natural generalization of the edge expansion problem. We call subsets of vertices (i.e., *clusters*) $A_1, \ldots, A_k$ a *k-way partition* of G if $A_i \cap A_j = \emptyset$ for different i and j, and $\bigcup_{i=1}^{k} A_i = V$. The k-way partitioning problem asks for a k-way partition of G such that the conductance of any A_i in the partition is at most the *k-way expansion constant*, defined by

$$(1) \qquad \rho(k) \triangleq \min_{\text{partition } A_1, \ldots, A_k} \max_{1 \leqslant i \leqslant k} \phi_G(A_i).$$

Clusters of low conductance in networks appearing in practice usually capture the

†Georgia Institute of Technology, Atlanta, GA 30332 (rpeng@cc.gatech.edu).
‡University of Bristol, Bristol, UK (h.sun@bristol.ac.uk, luca.zanetti@bristol.ac.uk).

notion of *community*, and algorithms for finding these subsets have applications in various domains such as community detection and network analysis. In computer vision, most image segmentation procedures are based on region-based merge and split [10], which in turn rely on partitioning graphs into multiple subsets [36]. On the theoretical side, decomposing vertex/edge sets into multiple disjoint subsets is used in designing approximation algorithms for unique games [39] and efficient algorithms for graph problems [18, 23, 38].

Despite widespread use of various graph partitioning schemes over the past decades, the quantitative relationship between the k-way expansion constant and the eigenvalues of the graph Laplacians were unknown until a sequence of very recent results [22, 24]. For instance, Lee, Oveis Gharan, and Trevisan [22] proved the following higher-order Cheeger inequality:

$$(2) \qquad \frac{\lambda_k}{2} \leqslant \rho(k) \leqslant O(k^2)\sqrt{\lambda_k},$$

where $0 = \lambda_1 \leqslant \cdots \leqslant \lambda_n \leqslant 2$ are the eigevalues of the normalized Laplacian matrix $\mathcal{L}$ of G. Informally, the higher-order Cheeger inequality shows that a graph G has a k-way partition with low $\rho(k)$ if and only if λ_k is small. Indeed, (2) implies that a large gap between λ_{k+1} and $\rho(k)$ *guarantees* (i) existence of a k-way partition $\{S_i\}_{i=1}^k$ with bounded $\phi_G(S_i) \leqslant \rho(k)$, and (ii) any $(k+1)$-way partition of G contains a subset with significantly higher conductance $\rho(k+1) \geqslant \lambda_{k+1}/2$ compared with $\rho(k)$. Hence, a suitable lower bound on the *gap* $\Upsilon(k)$ for some k, defined by

$$(3) \qquad \Upsilon(k) \triangleq \frac{\lambda_{k+1}}{\rho(k)},$$

implies the existence of a k-way partition for which every cluster has low conductance, and that G is a *well-clustered* graph.

We study well-clustered graphs which satisfy a gap assumption on $\Upsilon(k)$ in this paper. Our gap assumption on $\Upsilon(k)$ is slightly weaker than assuming gaps between the eigenvalues but is nonetheless related via Cheeger-type inequalities. Our assumption is also well-grounded in practical studies: clustering algorithms have been studied before under this assumption in machine learning, e.g., [1]. Sharp drop-offs between two consecutive eigenvalues have also been observed to give good indicators for the number of clusters, e.g., [40] and [14, section D].

1.1. Our results. We give structural results that show close connections between the eigenvectors and the indicator vectors of the clusters. This characterization allows us to show that many variants of spectral clustering, that are based on the spectral embedding and that work "in practice," can be rigorously analyzed "in theory." Moreover, exploiting our gap assumption, we can approximate this spectral embedding using the heat kernel of the graph. Combining this with approximate nearest neighbor data structures, we give a nearly linear time algorithm for the k-way partitioning problem.

Our structural results can be summarized as follows. Let $\{f_i\}_{i=1}^k$ be the eigenvectors corresponding to the k smallest eigenvalues of $\mathcal{L}$, and $\{S_i\}_{i=1}^k$ be a k-way partition of G achieving $\rho(k)$ defined in (1). We define $\{g_i\}_{i=1}^k$ to be the indicator vectors of the clusters $\{S_i\}_{i=1}^k$, where $g_i(u) = 1$ if $u \in S_i$, and $g_i(u) = 0$ otherwise. We further use $\{\bar{g}_i\}_{i=1}^k$ to express the normalized indicator vectors of the clusters $\{S_i\}_{i=1}^k$, defined by

$$\bar{g}_i = \frac{\mathbf{D}^{1/2}g_i}{\|D^{1/2}g_i\|}.$$

We show that, under the condition of $\Upsilon(k) = \Omega(k^2)$, the span of $\{\bar{g}_i\}_{i=1}^k$ and the span of $\{f_i\}_{i=1}^k$ are close to each other, which is stated formally in Theorem 1.1.

THEOREM 1.1 (the structure theorem). *Let $\{S_i\}_{i=1}^k$ be a k-way partition of G achieving $\rho(k)$, and let $\Upsilon(k) = \lambda_{k+1}/\rho(k) = \Omega(k^2)$. Let $\{f_i\}_{i=1}^k$ and $\{\bar{g}_i\}_{i=1}^k$ be defined as above. Then, the following statements hold:*
  1. *For every $\bar{g}_i$, there is a linear combination of $\{f_i\}_{i=1}^k$, called $\hat{f}_i$, such that $\|\bar{g}_i - \hat{f}_i\|^2 \leqslant 1/\Upsilon(k)$.*
  2. *For every f_i, there is a linear combination of $\{\bar{g}_i\}_{i=1}^k$, called $\hat{g}_i$, such that $\|f_i - \hat{g}_i\|^2 \leqslant 1.1k/\Upsilon(k)$.*

This theorem generalizes the result shown by Arora, Barak, and Steurer [2, Theorem 2.2], which proves the easier direction (the first statement, Theorem 1.1) and can be considered as a stronger version of the well-known Davis–Kahan theorem [12]. We remark that, despite our use of the higher-order Cheeger inequality (2) to motivate the definition of $\Upsilon(k)$, our proof of the structure theorem is self-contained. Specifically, it omits much of the machinery used in the proofs of higher-order and improved Cheeger inequalities [21, 22].

The structure theorem has several applications. For instance, we look at the well-known spectral embedding $F : V[G] \to \mathbb{R}^k$ defined by

$$(4) \qquad F(u) \triangleq \frac{1}{\mathsf{NormalizationFactor}(u)} \cdot (f_1(u), \dots, f_k(u))^{\mathsf{T}},$$

where $\mathsf{NormalizationFactor}(u) \in \mathbb{R}$ is a normalization factor for $u \in V[G]$. We use Theorem 1.1 to show that this well-known spectral embedding exhibits very nice geometric properties: (i) *all* points $F(u)$ from the same cluster are close to each other, and (ii) *most pairs of* points $F(u), F(v)$ from different clusters are far from each other; (iii) the bigger the value of $\Upsilon(k)$, the higher concentration the embedded points within the same cluster.

Based on these facts, we analyze the performance of spectral clustering, aiming at answering the following longstanding open question: *Why does spectral clustering perform well in practice?* We show that the partition $\{A_i\}_{i=1}^k$ produced by spectral clustering gives a good approximation of any "optimal" partition $\{S_i\}_{i=1}^k$: every A_i has low conductance and has large overlap with its corresponding S_i. This algorithm has comprehensive applications and has been the subject of extensive experimental studies for more than 20 years, e.g., [28, 40]. Prior to this work, similar results on spectral clustering mainly focus on graphs generated from the stochastic block model. Instead, our gap assumption captures more general classes of graphs by replacing the input model with a structural condition. Our result represents the first rigorous analysis of spectral clustering for the general family of graphs that exhibit a multicut structure but are not captured by the stochastic block model. Our result is as follows.

THEOREM 1.2 (approximation guarantee of spectral clustering). *Let G be a graph satisfying the condition $\Upsilon(k) = \lambda_{k+1}/\rho(k) = \Omega(k^3)$, and $k \in \mathbb{N}$. Let $F : V[G] \to \mathbb{R}^k$ be the embedding defined in (4). Let $\{A_i\}_{i=1}^k$ be a k-way partition by any k-means algorithm running in $\mathbb{R}^k$ that achieves an approximation ratio APT. Then, the following statements hold:* (i) $\mathrm{vol}(A_i \triangle S_i) = O(\mathsf{APT} \cdot k^3/\Upsilon(k)) \, \mathrm{vol}(S_i)$, *and* (ii) $\phi_G(A_i) = 1.1 \cdot \phi_G(S_i) + O(\mathsf{APT} \cdot k^3/\Upsilon(k))$.

We further study fast algorithms for partitioning well-clustered graphs. Notice that, for moderately large values of k, e.g., $k = \omega(\log n)$, directly applying k-means algorithms and Theorem 1.2 does not give a nearly linear time algorithm, since (i) ob-

taining the spectral embedding (4) requires $\Omega(mk)$ time for computing k eigenvectors, and (ii) most k-means algorithms run in $\Omega(nk)$ time.

To overcome the first obstacle, we study the so-called heat kernel embedding $x_t : V[G] \to \mathbb{R}^n$, an embedding from V to $\mathbb{R}^n$ defined by

$$x_t(u) \triangleq \frac{1}{\mathsf{NormalizationFactor}(u)} \cdot \left(\mathrm{e}^{-t \cdot \lambda_1} f_1(u), \ldots, \mathrm{e}^{-t \cdot \lambda_n} f_n(u) \right)$$

for some $t \in \mathbb{R}_{\geqslant 0}$. The heat kernel of a graph is a well-studied mathematical concept and is related to, for example, the study of random walks [34]. We exploit the heat kernel embedding to approximate the squared-distance $\|F(u) - F(v)\|^2$ of the embedded points $F(u)$ and $F(v)$ via their *heat-kernel distance* $\|x_t(u) - x_t(v)\|^2$. Since the heat kernel distances between vertices can be approximated in nearly linear time [29], this approach avoids the computation of eigenvectors for a large value of k. For the second obstacle, instead of applying k-means algorithms as a black-box, we apply approximate nearest-neighbor data structures. This can be viewed as an ad-hoc version of a k-means algorithm and indicates that in many scenarios the standard Lloyd-type heuristic widely used in k-means algorithms can eventually be avoided. Our result is as follows.

THEOREM 1.3 (nearly linear time algorithm for partitioning graphs). *Let $G = (V, E)$ be a graph of n vertices and m edges, and $k = \omega(\log n)$ be the number of clusters. Assume that $\Upsilon(k) = \lambda_{k+1}/\rho(k) = \widetilde{\Omega}(k^5)$, and $\{S_i\}_{i=1}^k$ is a k-way partition such that $\phi_G(S_i) \leqslant \rho(k)$. Then there is an algorithm which runs in $\widetilde{O}(m)$ time and outputs a k-way partition $\{A_i\}_{i=1}^k$ such that (i) $\mathrm{vol}(A_i \triangle S_i) = \widetilde{O}(k^4/\Upsilon(k)) \, \mathrm{vol}(S_i)$, and (ii) $\phi_G(A_i) = 1.1 \cdot \phi_G(S_i) + \widetilde{O}(k^4/\Upsilon(k))$. The $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ terms here hide a factor of* $\mathrm{poly} \log n$.

We remark that bounds of other expansion parameters of k-way partitioning can be derived from our analysis as well. For instance, it is easy to see that $\rho(k)$ and the *normalized cut* [36] studied in machine learning, which is defined as the sum of the conductance of all returned clusters, differ by at most a factor of k, and the normalized cut value of a k-way partition from spectral clustering can be derived from our results.

1.2. Related work. In the broadest sense, our algorithms are clustering routines. Clustering can be formulated in many ways, and the study of algorithms in many such formulations are areas of active work [7, 8, 17, 25]. Among these, our work is most closely related to spectral clustering, which is closely related to normalized or low conductance cuts [36]. The k-way expansion that we study is always within a factor of k of k-way normalized cuts.

Theoretical studies of graph partitioning are often based on augmenting the fractional relaxation of these cut problems with additional constraints in the form of semidefinite programs or Lasserre hierarchy. The goal of our study is to obtain similar bounds using more practical tools such as k-means and heat-kernel embedding.

Oveis Gharan and Trevisan [32] formulate the notion of clusters with respect to the *inner* and *outer* conductance: a cluster S should have low outer conductance, and the conductance of the induced subgraph by S should be high. Under a gap assumption between λ_{k+1} and λ_k, they present a polynomial-time algorithm which finds a k-way partition $\{A_i\}_{i=1}^k$ that satisfies the inner- and outer-conductance condition. In order to ensure that every A_i has high inner conductance, they assume that $\lambda_{k+1} \geqslant \mathrm{poly}(k)\lambda_k^{1/4}$, which is much stronger than ours. Moreover, their algorithm runs in polynomial-time, in contrast to our nearly linear time algorithm.

Dey, Rossi, and Sidiropoulos [13] study the properties of the spectral embedding for graphs having a gap between λ_k and λ_{k+1} and present a k-way partition algorithm, which is based on k-center clustering and is similar in spirit to our work. Using combinatorial arguments, they are able to show that the clusters concentrate around k distant points in the spectral embedding. In contrast to our work, their result only holds for bounded-degree graphs and cannot provide an approximate guarantee for individual clusters. Moreover, their algorithm runs in nearly linear time only if $k = O(\mathrm{poly} \log n)$.

We also explore the separation between λ_k and λ_{k+1} from an algorithmic perspective and show that this assumption interacts well with heat-kernel embeddings. The heat kernel has been used in previous algorithms on local partitioning [9] and balanced separators [29]. It also plays a key role in current efficient approximation algorithms for finding low conductance cuts [30, 35]. However, most of these theoretical guarantees are through the matrix multiplicative weights update framework [3, 4]. Our algorithm instead directly uses the heat-kernel embedding to find low conductance cuts.

There is also a considerable amount of research on partitioning random graphs. For instance, in the stochastic block model (SBM) [27], the input graph with k clusters is generated according to probabilities p and q with $p > q$: an edge between any two vertices within the same cluster is placed with probability p, and an edge between any two vertices from different clusters is placed with probability q. It is proved that spectral algorithms give the correct clustering for certain ranges of p and q [27, 33, 41]. However, the analysis of these algorithms cannot be easily generalized into our setting: we consider graphs where edges are not necessarily chosen independently with certain probabilities but can be added in an "adversarial" way. For this reason, standard perturbation theorems used in the analysis of algorithms for SBMs, such as the Davis–Kahan theorem [12], cannot be always applied, and ad-hoc arguments specific for graphs, like our structure theorem (Theorem 1.1), become necessary.

2. Preliminaries. Let $G = (V, E)$ be an undirected and unweighted graph with n vertices and m edges. The set of neighbors of a vertex u is represented by $N(u)$, and its degree is $d_u = |N(u)|$. For any set $S \subseteq V$, let $\mathrm{vol}(S) \triangleq \sum_{u \in S} d_u$. For any set $S, T \subseteq V$, we define $E(S, T)$ to be the set of edges between S and T, that is, $E(S, T) \triangleq \{\{u, v\} | u \in S \text{ and } v \in T\}$. For simplicity, we write $\partial S = E(S, V \setminus S)$ for any set $S \subseteq V$. For two sets X and Y, the symmetric difference of X and Y is defined as $X \triangle Y \triangleq (X \setminus Y) \cup (Y \setminus X)$.

We work extensively with algebraic objects related to G. We use $\mathbf{D}$ to denote the $n \times n$ diagonal matrix with $\mathbf{D}_{uu} = d_u$ for $u \in V[G]$. The *Laplacian matrix* of G is defined by $\mathbf{L} \triangleq \mathbf{D} - \mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix of G defined by $\mathbf{A}_{u,v} = 1$ if $\{u, v\} \in E[G]$, and $\mathbf{A}_{u,v} = 0$ otherwise. The *normalized Laplacian matrix* of G is defined by $\mathcal{L} \triangleq \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$. For this matrix, we denote its n eigenvalues with $0 = \lambda_1 \leqslant \cdots \leqslant \lambda_n \leqslant 2$, with their corresponding orthonormal eigenvectors $f_1, \ldots, f_n$. Note that if G is connected, the first eigenvector is $f_1 = \mathbf{D}^{1/2} f$, where f is any nonzero constant vector.

For a vector $x \in \mathbb{R}^n$, the Euclidean norm of x is given by $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$. For any $f : V \to \mathbb{R}$ and $h \triangleq \mathbf{D}^{-1/2} f$, the *Rayleigh quotient* of f with respect to graph G is given by

$$\mathcal{R}(f) \triangleq \frac{f^{\intercal} \mathcal{L} f}{\|f\|^2} = \frac{h^{\intercal} \mathbf{L} h}{\|h\|_{\mathbf{D}}} = \frac{\sum_{\{u,v\} \in E(G)} (h(u) - h(v))^2}{\sum_u d_u h(u)^2},$$

where $\|h\|_{\mathbf{D}} \triangleq h^{\intercal} \mathbf{D} h$. Based on the Rayleigh quotient, the conductance of a set S_i

$$\hat{f}_i \quad \underset{\|\hat{f}_i - \overline{g}_i\|^2 \leqslant 1/\Upsilon}{\overline{\hspace{3cm}\text{Part 1}\hspace{3cm}}} \quad \overline{g}_i$$

$\hat{f}_i =$ a linear combination of $\{f_j\}$ with coefficients $\alpha_j^{(i)}$

$\hat{g}_i =$ a linear combination of $\{\overline{g}_j\}$ with coefficients $\beta_j^{(i)}$

$$f_i \quad \underset{\|f_i - \hat{g}_i\|^2 \leqslant 1.1k/\Upsilon}{\overline{\hspace{3cm}\text{Part 2}\hspace{3cm}}} \quad \hat{g}_i$$
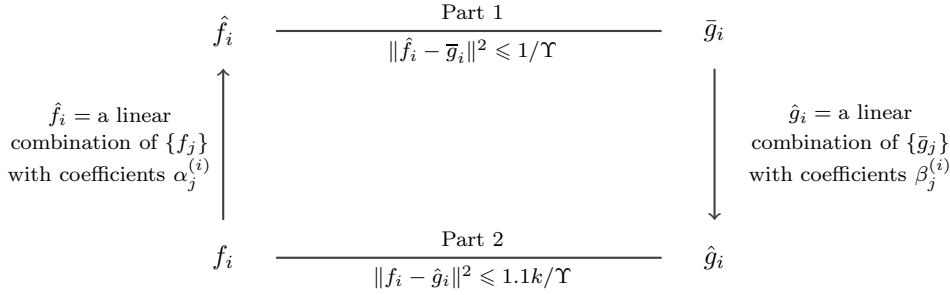
FIG. 1. *Relations among $\{\hat{f}_i\}$, $\{f_i\}$, $\{\overline{g}_i\}$, and $\{\hat{g}_i\}$ given in Theorem 3.1. Here Υ is the gap defined with respect to λ_{k+1} and $\rho(k)$.*

can be expressed as $\phi_G(S_i) = \mathcal{R}(\overline{g}_i)$, and the gap $\Upsilon(k)$ can be written as

$$(5) \qquad \Upsilon(k) = \frac{\lambda_{k+1}}{\rho(k)} = \min_{1 \leqslant i \leqslant k} \frac{\lambda_{k+1}}{\phi_G(S_i)} = \min_{1 \leqslant i \leqslant k} \frac{\lambda_{k+1}}{\mathcal{R}(\overline{g}_i)}.$$

Since k is always fixed as part of the algorithm's input, throughout the rest of the paper we always use Υ to express $\Upsilon(k)$ for simplicity. We will also use $S_1, \ldots, S_k$ to express a k-way partition of G achieving $\rho(k)$. Note that this partition may not be unique.

3. Connection between eigenvectors and indicator vectors of clusters. In this section we study the relations between the multiple cuts of a graph and the eigenvectors of the graph's normalized Laplacian matrix. Given clusters $S_1 \ldots S_k$, define the indicator vector of cluster S_i by

$$(6) \qquad g_i(u) = \begin{cases} 1 & \text{if } u \in S_i, \\ 0 & \text{if } u \notin S_i, \end{cases}$$

and define the corresponding normalized indicator vector by

$$(7) \qquad \overline{g}_i = \frac{\mathbf{D}^{1/2} g_i}{\|\mathbf{D}^{1/2} g_i\|}.$$

A basic result in spectral graph theory states that G has k connected components if and only if the k smallest eigenvalues are 0, implying that the spaces spanned by $f_1, \ldots, f_k$ and $\overline{g}_1, \ldots, \overline{g}_k$ are the same. Generalizing this result, we expect that these two spaces would be still similar if these k components of G are loosely connected, in the sense that (i) every eigenvector f_i can be approximately expressed by a linear combination of $\{\overline{g}_i\}_{i=1}^k$, and (ii) every indicator vector $\overline{g}_i$ can be approximately expressed by a linear combination of $\{f_i\}_{i=1}^k$. This leads to our structure theorem, which is illustrated in Figure 1.

THEOREM 3.1 (the structure theorem, formal statement). *Let $\Upsilon = \Omega(k^2)$, and $1 \leqslant i \leqslant k$. Then, the following statements hold:*
1. *There is a linear combination of the eigenvectors $f_1, \ldots, f_k$ with coefficients $\alpha_j^{(i)}$: $\hat{f}_i = \alpha_1^{(i)} f_1 + \cdots + \alpha_k^{(i)} f_k$, such that $\|\overline{g}_i - \hat{f}_i\|^2 \leqslant 1/\Upsilon$.*
2. *There is a linear combination of the vectors $\overline{g}_1, \ldots, \overline{g}_k$ with coefficients $\beta_j^{(i)}$: $\hat{g}_i = \beta_1^{(i)} \overline{g}_1 + \cdots + \beta_k^{(i)} \overline{g}_k$, such that $\|f_i - \hat{g}_i\|^2 \leqslant 1.1k/\Upsilon$.*

Part 1 of Theorem 3.1 shows that the normalized indicator vectors $\bar{g}_i$ of every cluster S_i can be approximated by a linear combination of *the first k eigenvectors*, with respect to the value of Υ. The proof follows from the fact that if $\bar{g}_i$ has small Rayleigh quotient, then the inner product between $\bar{g}_i$ and the eigenvectors corresponding to larger eigenvalues must be small. This statement was also shown implicitly in [2, Theorem 2.2].

Proof of Part 1 *of Theorem* 3.1. We write $\overline{g}_i$ as a linear combination of the eigenvectors of $\mathcal{L}$, i.e.,

$$\overline{g}_i = \alpha_1^{(i)} f_1 + \cdots + \alpha_n^{(i)} f_n$$

and let the vector $\hat{f}_i$ be the projection of vector $\bar{g}_i$ on the subspace spanned by $\{f_i\}_{i=1}^k$, i.e.,

$$\hat{f}_i = \alpha_1^{(i)} f_1 + \cdots + \alpha_k^{(i)} f_k.$$

By the definition of Rayleigh quotients, we have that

$$
\begin{aligned}
\mathcal{R}(\overline{g}_i) &= \left(\alpha_1^{(i)} f_1 + \cdots + \alpha_n^{(i)} f_n \right)^{\mathsf{T}} \mathcal{L} \left(\alpha_1^{(i)} f_1 + \cdots + \alpha_n^{(i)} f_n \right) \\
&= \left(\alpha_1^{(i)} \right)^2 \lambda_1 + \cdots + \left(\alpha_n^{(i)} \right)^2 \lambda_n \\
&\geqslant \left(\alpha_2^{(i)} \right)^2 \lambda_2 + \cdots + \left(\alpha_k^{(i)} \right)^2 \lambda_k + \left(1 - \alpha' - \left(\alpha_1^{(i)} \right)^2 \right) \lambda_{k+1},
\end{aligned}
$$

where $\alpha' \triangleq (\alpha_2^{(i)})^2 + \cdots + (\alpha_k^{(i)})^2$. Therefore, we have that

$$1 - \alpha' - \left(\alpha_1^{(i)} \right)^2 \leqslant \mathcal{R}(\overline{g}_i)/\lambda_{k+1} \leqslant 1/\Upsilon$$

and

$$\|\overline{g}_i - \hat{f}_i\|^2 = \left(\alpha_{k+1}^{(i)} \right)^2 + \cdots + \left(\alpha_n^{(i)} \right)^2 = 1 - \alpha' - \left(\alpha_1^{(i)} \right)^2 \leqslant 1/\Upsilon,$$

which finishes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Part 2 of Theorem 3.1 is more interesting and shows that the opposite direction holds as well, i.e., any f_i ($1 \leqslant i \leqslant k$) can be approximated by a linear combination of the normalized indicator vectors $\{\overline{g}_i\}_{i=1}^k$. To sketch the proof, note that if we could write every $\overline{g}_i$ *exactly* as a linear combination of $\{f_i\}_{i=1}^k$, then we could write every f_i ($1 \leqslant i \leqslant k$) as a linear combination of $\{\overline{g}_i\}_{i=1}^k$. This is because both $\{f_i\}_{i=1}^k$ and $\{\overline{g}_i\}_{i=1}^k$ are sets of linearly independent vectors of the same dimension and span $\{\overline{g}_1, \ldots, \overline{g}_k\} \subseteq \mathrm{span}\{f_1, \ldots, f_k\}$. However, the $\overline{g}_i$'s are only close to a linear combination of the first k eigenvectors, as shown in Part 1. We will denote this combination as $\hat{f}_i$ and use the fact that the errors of approximation are small to show that these $\{\hat{f}_i\}_{i=1}^k$ are almost orthogonal between each other. This allows us to show that span $\{\hat{f}_1, \ldots, \hat{f}_k\} = \mathrm{span}\{f_1, \ldots, f_k\}$, which implies Part 2.

We will use the following two classical results in our proof.

THEOREM 3.2 (Geršgorin circle theorem).   *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix , and let* $R_i(\mathbf{A}) = \sum_{j \neq i} |\mathbf{A}_{i,j}|$ *for* $1 \leqslant i \leqslant n$. *Then, all eigenvalues of* $\mathbf{A}$ *are in the union of Geršgorin Discs defined by*

$$\bigcup_{i=1}^{n} \left\{ z \in \mathbb{C} : |z - \mathbf{A}_{i,i}| \leqslant R_i(\mathbf{A}) \right\}.$$

THEOREM 3.3 (see [15, Corollary 6.3.4]). *Let $\mathbf{A}$ be an $n \times n$ real and symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, and $\mathbf{E}$ be an $n \times n$ matrix. If $\hat{\lambda}$ is an eigenvalue of $\mathbf{A} + \mathbf{E}$, then there is some eigenvalue λ_i of $\mathbf{A}$ for which $|\hat{\lambda} - \lambda_i| \leqslant \|\mathbf{E}\|$.*

Proof of Part 2 of Theorem 3.1. By Part 1, every $\overline{g}_i$ is approximated by a vector $\hat{f}_i$ defined by

$$\hat{f}_i = \alpha_1^{(i)} f_1 + \cdots \alpha_k^{(i)} f_k.$$

Define a k by k matrix $\mathbf{A}$ such that $\mathbf{A}_{i,j} = \alpha_i^{(j)}$, i.e., the jth column of matrix $\mathbf{A}$ consists of values $\{\alpha_i^{(j)}\}_{i=1}^k$ representing $\hat{f}_j$. We express the jth column of $\mathbf{A}$ by a vector $\alpha^{(j)}$, defined as

$$\alpha^{(j)} = \left(\alpha_1^{(j)}, \ldots, \alpha_k^{(j)} \right)^{\mathsf{T}}.$$

We will show that the vectors $\{\alpha^{(j)}\}_{j=1}^k$ are linearly independent, which implies that $\{\hat{f}^{(j)}\}_{j=1}^k$ are linearly independent as well. To prove this, we will show that $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ has no zero eigenvalue, and hence $\mathbf{A}$ is invertible.

First, notice that it holds by the orthonormality of $\{f_i\}_{i=1}^k$ that

$$
\begin{aligned}
\left| \left\langle \alpha^{(i)}, \alpha^{(j)} \right\rangle \right| &= \left| \left\langle \hat{f}_i, \hat{f}_j \right\rangle \right| = \left| \left\langle \overline{g}_i - (\overline{g}_i - \hat{f}_i), \overline{g}_j - (\overline{g}_j - \hat{f}_j) \right\rangle \right| \\
&= \left| \langle \overline{g}_i, \overline{g}_j \rangle - \left\langle \overline{g}_i - \hat{f}_i, \overline{g}_j \right\rangle - \left\langle \overline{g}_i, \overline{g}_j - \hat{f}_j \right\rangle + \left\langle \overline{g}_i - \hat{f}_i, \overline{g}_j - \hat{f}_j \right\rangle \right| \\
&\leqslant \left\| \overline{g}_i - \hat{f}_i \right\| + \left\| \overline{g}_j - \hat{f}_j \right\| + \left\| \overline{g}_i - \hat{f}_i \right\| \left\| \overline{g}_j - \hat{f}_j \right\| \\
&\leqslant 2\sqrt{1/\Upsilon} + 1/\Upsilon,
\end{aligned}
$$

where the first inequality follows from the orthonormality of $\overline{g}_i$ and $\overline{g}_j$, and the second inequality follows by Part 1 of Theorem 3.1. So it holds for any $i \neq j$ that

$$|(\mathbf{A}^{\mathsf{T}}\mathbf{A})_{i,j}| = \left| \sum_{\ell=1}^k \mathbf{A}_{\ell,i} \mathbf{A}_{\ell,j} \right| = \left| \sum_{\ell=1}^k \alpha_\ell^{(i)} \alpha_\ell^{(j)} \right| = \left| \left\langle \alpha^{(i)}, \alpha^{(j)} \right\rangle \right| \leqslant 3\sqrt{1/\Upsilon}$$

while

$$(\mathbf{A}^{\mathsf{T}}\mathbf{A})_{i,i} = \sum_{\ell=1}^k \left(\alpha_\ell^{(i)} \right)^2 \geqslant 1 - 1/\Upsilon.$$

Then, by the Geršgorin circle theorem (cf. Theorem 3.2), it holds that all the eigenvalues of $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ are at least

$$1 - 1/\Upsilon - (k-1) \cdot 3\sqrt{1/\Upsilon}.$$

Therefore, $\mathbf{A}$ has no eigenvalue with value 0 as long as $\Upsilon > 10k^2$, proving that the vectors $\{\alpha^{(j)}\}_{j=1}^k$ are linearly independent. Combining this with the fact that $\mathrm{span}\{\hat{f}_1, \ldots, \hat{f}_k\} \subseteq \mathrm{span}\{f_1, \ldots, f_k\}$ and $\dim(\mathrm{span}(\{f_1, \ldots, f_k\})) = k$, it holds that $\mathrm{span}\{\hat{f}_1, \ldots, \hat{f}_k\} = \mathrm{span}\{f_1, \ldots, f_k\}$. Hence, we can write every f_i $(1 \leqslant i \leqslant k)$ as a linear combination of $\{\hat{f}_i\}_{i=1}^k$, i.e.,

$$(8) \qquad f_i = \beta_1^{(i)} \hat{f}_1 + \beta_2^{(i)} \hat{f}_2 + \cdots + \beta_k^{(i)} \hat{f}_k.$$

Now define the value of $\hat{g}_i$ as

$$(9) \qquad \hat{g}_i = \beta_1^{(i)} \overline{g}_1 + \beta_2^{(i)} \overline{g}_2 + \cdots + \beta_k^{(i)} \overline{g}_k$$

and define $\|\beta\|^2 = \sum_{j=1}^k (\beta_j^{(i)})^2$. Then, it holds that

$$
\begin{aligned}
1 = \|f_i\|^2 &= \sum_{\ell=1}^k \left(\beta_\ell^{(i)}\right)^2 \left\|\hat{f}_\ell\right\|^2 + \sum_{\ell \neq \ell'} \beta_\ell^{(i)} \beta_{\ell'}^{(i)} \left\langle \hat{f}_\ell, \hat{f}_{\ell'} \right\rangle \\
&\geqslant \|\beta\|^2 (1 - 1/\Upsilon) - \sum_\ell \left|\beta_\ell^{(i)}\right| \sum_{\ell' \neq \ell} \left|\beta_{\ell'}^{(i)}\right| \left\langle \hat{f}_\ell, \hat{f}_{\ell'} \right\rangle \\
&\geqslant \|\beta\|^2 (1 - 1/\Upsilon) - \left(\sqrt{k} \cdot \|\beta\|\right) \cdot \left(\sqrt{k} \cdot \|\beta\|\right) \cdot \left(3 \cdot \sqrt{1/\Upsilon}\right) \\
&\geqslant \left(1 - 1/\Upsilon - 3k/\sqrt{\Upsilon}\right) \|\beta\|^2,
\end{aligned}
$$

where the second inequality holds by the Cauchy–Schwarz inequality. Since $\Upsilon = \Omega(k^2)$, we have that

$$
\|\beta\|^2 \leqslant \left(1 - \frac{1}{\Upsilon} - \frac{3k}{\sqrt{\Upsilon}}\right)^{-1} \leqslant 1.1.
$$

Combining this with Part 1 of Theorem 3.1 and the Cauchy–Schwarz inequality, we have that

$$
\|f_i - \hat{g}_i\| \leqslant \sum_{j=1}^k \left|\beta_j^{(i)}\right| \left\|\hat{f}_j - \overline{g}_j\right\| \leqslant \left(1/\sqrt{\Upsilon}\right) \sum_{j=1}^k \left|\beta_j^{(i)}\right| \leqslant \sqrt{1.1k/\Upsilon},
$$

which proves Part 2 of the theorem. $\qquad\square$

Theorem 3.1 shows a close connection between the first k eigenvectors and the indicator vectors of the clusters. We leverage this and the fact that the $\{\hat{g}_i\}$'s are almost orthogonal between each other to show that, for any two different clusters S_i and S_j, there exists an eigenvector having reasonably different values on the coordinates which correspond to S_i and S_j.

LEMMA 3.4. *Let* $\Upsilon = \Omega(k^3)$. *For any* $1 \leqslant i \leqslant k$, *let* $\hat{g}_i = \beta_1^{(i)} \overline{g}_1 + \cdots + \beta_k^{(i)} \overline{g}_k$ *be such that* $\|f_i - \hat{g}_i\| \leqslant 1.1k/\Upsilon$. *Then, for any* $\ell \neq j$, *there exists* $i \in \{1, \ldots, k\}$ *such that*

$$
(10) \qquad\qquad \left|\beta_\ell^{(i)} - \beta_j^{(i)}\right| \geqslant \zeta \triangleq \frac{1}{10\sqrt{k}}.
$$

Proof. Let $\beta^{(i)} = (\beta_1^{(i)}, \ldots, \beta_k^{(i)})^\intercal$ for $1 \leqslant i \leqslant k$. Since $\overline{g}_i \perp \overline{g}_j$ for any $i \neq j$, we have by the orthonormality of $\overline{g}_1, \ldots, \overline{g}_k$ that

$$
\begin{aligned}
\langle \hat{g}_i, \hat{g}_j \rangle &= \left\langle \beta_1^{(i)} \overline{g}_1 + \cdots + \beta_k^{(i)} \overline{g}_k, \beta_1^{(j)} \overline{g}_1 + \cdots + \beta_k^{(j)} \overline{g}_k \right\rangle \\
&= \sum_{\ell=1}^k \beta_\ell^{(i)} \beta_\ell^{(j)} \|\overline{g}_\ell\|^2 = \left\langle \beta^{(i)}, \beta^{(j)} \right\rangle
\end{aligned}
$$

and

$$
\begin{aligned}
\left|\left\langle \beta^{(i)}, \beta^{(j)} \right\rangle\right| &= |\langle \hat{g}_i, \hat{g}_j \rangle| = |\langle f_i - (f_i - \hat{g}_i), f_j - (f_j - \hat{g}_j) \rangle| \\
&= |\langle f_i, f_j \rangle - \langle f_i - \hat{g}_i, f_j \rangle - \langle f_j - \hat{g}_j, f_i \rangle + \langle f_i - \hat{g}_i, f_j - \hat{g}_j \rangle| \\
&\leqslant \|f_i - \hat{g}_i\| + \|f_j - \hat{g}_j\| + \|f_i - \hat{g}_i\| \|f_j - \hat{g}_j\| \\
&\leqslant 2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon.
\end{aligned}
$$

Moreover, it holds that

$$\left\|\beta^{(i)}\right\| = \|\hat{g}_i\| = \|f_i + \hat{g}_i - f_i\| \leqslant 1 + \|\hat{g}_i - f_i\| \leqslant 1 + \sqrt{1.1k/\Upsilon}$$

and

$$\left\|\beta^{(i)}\right\| = \|\hat{g}_i\| = \|f_i + \hat{g}_i - f_i\| \geqslant 1 - \|\hat{g}_i - f_i\| \geqslant 1 - \sqrt{1.1k/\Upsilon},$$

which implies that

$$(11) \qquad \left\|\beta^{(i)}\right\|^2 \in \left(1 - (2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon), 1 + 2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right).$$

In other words, we showed that $\beta^{(i)}$'s are almost orthonormal.

Now we construct a k by k matrix $\mathbf{B}$, where the jth column of $\mathbf{B}$ is $\beta^{(j)}$. By the Geršgorin circle theorem (Theorem 3.2), all eigenvalues λ of $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ satisfy

$$(12) \qquad |\lambda - (\mathbf{B}^{\mathsf{T}}\mathbf{B})_{i,i}| \leqslant (k-1) \cdot (2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon)$$

for any i. Combing this with (11), we have that the eigenvalues of $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ are close to 1.

Now we show that $\beta_\ell^{(i)}$ and $\beta_j^{(i)}$ are far from each other by contradiction. Suppose there exist $\ell \neq j$ such that

$$\zeta' \triangleq \max_{1 \leqslant i \leqslant k} \left|\beta_\ell^{(i)} - \beta_j^{(i)}\right| < \frac{1}{10\sqrt{k}}.$$

This implies that the jth row and ℓth row of matrix $\mathbf{B}$ are somewhat close to each other. Let us now define matrix $\mathbf{E} \in \mathbb{R}^{k \times k}$, where

$$\mathbf{E}_{\ell,i} \triangleq \beta_j^{(i)} - \beta_\ell^{(i)},$$

and $\mathbf{E}_{t,i} = 0$ for any $t \neq \ell$ and $1 \leqslant i \leqslant k$. Moreover, let $\mathbf{Q} = \mathbf{B} + \mathbf{E}$. Notice that $\mathbf{Q}$ has two identical rows, and rank at most $k-1$. Therefore, $\mathbf{Q}$ has an eigenvalue with value 0, and the spectral norm $\|\mathbf{E}\|$ of $\mathbf{E}$, the largest singular value of $\mathbf{E}$, is at most $\sqrt{k}\zeta'$. By definition of matrix $\mathbf{Q}$ we have that

$$\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{B}^{\mathsf{T}}\mathbf{B} + \mathbf{B}^{\mathsf{T}}\mathbf{E} + \mathbf{E}^{\mathsf{T}}\mathbf{B} + \mathbf{E}^{\mathsf{T}}\mathbf{E}.$$

Since $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ is symmetric and 0 is an eigenvalue of $\mathbf{Q}^{\mathsf{T}}\mathbf{Q}$, by Theorem 3.3 we know that if $\hat{\lambda}$ is an eigenvalue of $\mathbf{Q}^{\mathsf{T}}\mathbf{Q}$, then there is an eigenvalue λ of $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ such that

$$\begin{aligned}
|\hat{\lambda} - \lambda| &\leqslant \|\mathbf{B}^{\mathsf{T}}\mathbf{E} + \mathbf{E}^{\mathsf{T}}\mathbf{B} + \mathbf{E}^{\mathsf{T}}\mathbf{E}\| \\
&\leqslant \|\mathbf{B}^{\mathsf{T}}\mathbf{E}\| + \|\mathbf{E}^{\mathsf{T}}\mathbf{B}\| + \|\mathbf{E}^{\mathsf{T}}\mathbf{E}\| \\
&\leqslant 4\sqrt{k}\zeta' + k\zeta'^2,
\end{aligned}$$

which implies that

$$\hat{\lambda} \geqslant \lambda - 4\sqrt{k}\zeta' - k\zeta'^2 \geqslant 1 - k\left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right) - 4\sqrt{k}\zeta' - k\zeta'^2,$$

due to (11) and (12). By setting $\hat{\lambda} = 0$, we have that

$$1 - k\left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right) - 4\sqrt{k}\zeta' - k\zeta'^2 \leqslant 0.$$

By the condition of $\Upsilon = \Omega(k^3)$, the inequality above implies that $\zeta' \geqslant \frac{1}{10\sqrt{k}}$, which leads to a contradiction. $\qquad\square$

We point out that it was shown in [21] that the first k eigenvectors can be approximated by a $(2k+1)$-step function. The quality of the approximation is the same as the one given by our structure theorem. However, a $(2k+1)$-step approximation is not enough to show that most vertices belonging to the same cluster are mapped close to each other in the spectral embedding.

We further point out that standard matrix perturbation theorems cannot be applied in our setting. For instance, we look at a well-clustered graph G that contains a subset C of a cluster S_i such that most neighbors of vertices in C are outside S_i. In this case, the adjacency matrix representing crossing edges of G has a high spectral norm, and hence standard matrix perturbation arguments could not give us a meaningful result. However, our structure theorem takes the fact that $\mathrm{vol}(C)$ has to be small into account, and that is why the structure theorem is needed to analyze the cut structure of a graph.

4. Analysis of spectral clustering. In this section we analyze an algorithm based on the classical spectral clustering paradigm and give an approximation guarantee of this method on well-clustered graphs. We will show that any k-means algorithm $\mathsf{AlgoMean}(\mathcal{X}, k)$ with certain approximation guarantee can be used for the k-way partitioning problem. Furthermore, it suffices to call $\mathsf{AlgoMean}$ in a black-box manner with a point set $\mathcal{X} \subseteq \mathbb{R}^k$.

This section is structured as follows. We first give a quick overview of spectral and k-means clustering in section 4.1. In section 4.2, we use the structure theorem to analyze the spectral embedding. Section 4.3 gives a general result about k-means when applied to this embedding and gives the proof of Theorem 1.2.

4.1. k-means clustering. Given a set of points $\mathcal{X} \subseteq \mathbb{R}^d$, a *k-means algorithm* $\mathsf{AlgoMean}(\mathcal{X}, k)$ seeks to find a set $\mathcal{K}$ of k centers $c_1, \ldots, c_k$ to minimize the sum of the ℓ_2^2-distance between $x \in \mathcal{X}$ and the center to which it is assigned. Formally, for any partition $\mathcal{X}_1, \ldots, \mathcal{X}_k$ of the set $\mathcal{X} \subseteq \mathbb{R}^d$, we define the cost function by

$$\mathsf{COST}(\mathcal{X}_1, \ldots, \mathcal{X}_k) \triangleq \min_{c_1, \ldots, c_k \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{x \in \mathcal{X}_i} \|x - c_i\|^2,$$

i.e., the COST function minimizes the total ℓ_2^2-distance between the points x's and their individually closest center c_i, where $c_1, \ldots, c_k$ are chosen arbitrarily in $\mathbb{R}^d$. We further define the optimal clustering cost by

$$(13) \qquad \Delta_k^2(\mathcal{X}) \triangleq \min_{\text{partition } \mathcal{X}_1, \ldots, \mathcal{X}_k} \mathsf{COST}(\mathcal{X}_1, \ldots, \mathcal{X}_k).$$

Spectral clustering can be described as follows: (i) Compute the bottom k eigenvectors $f_1, \ldots, f_k$ of the normalized Laplacian matrix[1] of graph G. (ii) Map every vertex $u \in V[G]$ to a point $F(u) \in \mathbb{R}^k$ according to

$$(14) \qquad F(u) = \frac{1}{\mathsf{NormalizationFactor}(u)} \cdot (f_1(u), \ldots, f_k(u))^{\mathsf{T}}$$

with a proper normalization factor $\mathsf{NormalizationFactor}(u) \in \mathbb{R}$ for each $u \in V$. (iii) Let $\mathcal{X} \triangleq \{F(u) : u \in V\}$ be the set of the embedded points from vertices in G. Run

---

[1]Other graph matrices (e.g., the adjacency matrix and the Laplacian matrix) are also widely used in practice. Notice that, with proper normalization, the choice of these matrices does not substantially influence the performance of k-means algorithms.

AlgoMean$(\mathcal{X}, k)$, and group the vertices of G into k clusters according to the output of AlgoMean$(\mathcal{X}, k)$. This approach that combines a k-means algorithm with a spectral embedding has been widely used in practice for a long time, although there was a lack of rigorous analysis of its performance prior to our result.

4.2. Analysis of the spectral embedding. The first step of spectral clustering is to map vertices of a graph into points in Euclidean space, through the spectral embedding (14). This subsection analyzes the properties of this embedding. Let us define the normalization factor to be

$$\mathsf{NormalizationFactor}(u) \triangleq \sqrt{d_u}.$$

We will show that the embedding (14) with the normalization factor above has very nice properties: embedded points from the same cluster S_i are concentrated around their center $c_i \in \mathbb{R}^k$, and embedded points from different clusters of G are far from each other. These properties imply that a simple k-means algorithm is able to produce a good clustering.[2]

We first define k points $p^{(i)} \in \mathbb{R}^k$ $(1 \leqslant i \leqslant k)$, where

$$(15) \qquad p^{(i)} \triangleq \frac{1}{\sqrt{\mathrm{vol}\,(S_i)}} \left(\beta_i^{(1)}, \ldots, \beta_i^{(k)} \right)^{\mathsf{T}}$$

and the parameters $\{\beta_i^{(j)}\}_{j=1}^k$ are defined in Theorem 3.1. We will show in Lemma 4.1 that all embedded points $\mathcal{X}_i \triangleq \{F(u) : u \in S_i\}$ $(1 \leqslant i \leqslant k)$ are concentrated around $p^{(i)}$. Moreover, we bound the total ℓ_2^2-distance between points in $\mathcal{X}_i$ and $p^{(i)}$, which is proportional to $1/\Upsilon$: the bigger the value of Υ, the higher concentration the points within the same cluster have. Notice that we *do not* claim that $p^{(i)}$ is the actual center of $\mathcal{X}_i$. However, these approximated points $p^{(i)}$'s suffice for our analysis.

LEMMA 4.1. *It holds that*

$$\sum_{i=1}^k \sum_{u \in S_i} d_u \|F(u) - p^{(i)}\|^2 \leqslant 1.1 k^2/\Upsilon.$$

Proof. Since $\hat{g}_j(u) = \sqrt{\frac{d_u}{\mathrm{vol}(S_i)}} \beta^{(j)}$ and $p_j^{(i)} = \frac{1}{\sqrt{\mathrm{vol}(S_i)}} \beta_i^{(j)}$ hold for any $1 \leqslant j \leqslant k$ and $u \in S_i$ by definition, we have that

$$
\begin{aligned}
\sum_{i=1}^k \sum_{u \in S_i} d_u \left(F(u)_j - p_j^{(i)} \right)^2 &= \sum_{i=1}^k \sum_{u \in S_i} d_u \left(\frac{1}{\sqrt{d_u}} f_j(u) - \frac{1}{\sqrt{\mathrm{vol}(S_i)}} \beta_i^{(j)} \right)^2 \\
&= \sum_{i=1}^k \sum_{u \in S_i} \left(f_j(u) - \sqrt{\frac{d_u}{\mathrm{vol}(S_i)}} \beta_i^{(j)} \right)^2 \\
&= \sum_{i=1}^k \sum_{u \in S_i} (f_j(u) - \hat{g}_j(u))^2 \\
&= \|f_j - \hat{g}_j\|^2 \\
&\leqslant 1.1 k/\Upsilon,
\end{aligned}
$$

---

[2] Notice that this embedding is similar with the one used in [22], with the only difference that $F(u)$ is not normalized and so it is not necessarily a unit vector. This difference, though, is crucial for our analysis.

where the last inequality follows from Theorem 3.1. Summing over all j for $1 \leqslant j \leqslant k$ implies that

$$\sum_{i=1}^{k} \sum_{u \in S_i} d_u \left\| F(u) - p^{(i)} \right\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{u \in S_i} d_u \left(F(u)_j - p_j^{(i)} \right)^2 \leqslant 1.1 k^2 / \Upsilon. \qquad \square$$

The next lemma shows that the ℓ_2^2-norm of $p^{(i)}$ is inversely proportional to the volume of S_i. This implies that embedded points from a big cluster are close to the origin, while embedded points from a small cluster are far from the origin.

LEMMA 4.2. *It holds for every $1 \leqslant i \leqslant k$ that*

$$\frac{99}{100 \operatorname{vol}(S_i)} \leqslant \left\| p^{(i)} \right\|^2 \leqslant \frac{101}{100 \operatorname{vol}(S_i)}.$$

Proof. By (15), we have that

$$\left\| p^{(i)} \right\|^2 = \frac{1}{\operatorname{vol}(S_i)} \left\| \left(\beta_i^{(1)}, \ldots, \beta_i^{(k)} \right)^{\mathsf{T}} \right\|^2.$$

Notice that $p^{(i)}$ is just the ith row of the matrix $\mathbf{B}$ defined in the proof of Lemma 3.4, normalized by $\sqrt{\operatorname{vol}(S_i)}$. Since $\mathbf{B}$ and $\mathbf{B}^{\mathsf{T}}$ share the same singular values (this follows from the SVD decomposition), by (12) the eigenvalues of $\mathbf{B}\mathbf{B}^{\mathsf{T}}$ are close to 1. But since $(\mathbf{B}\mathbf{B}^{\mathsf{T}})_{i,i}$ is equal to the ℓ_2^2-norm of the ith row of $\mathbf{B}$, we have that

$$(16) \quad \left\| \left(\beta_i^{(1)}, \ldots, \beta_i^{(k)} \right)^{\mathsf{T}} \right\|^2 \in \left(1 - \left(2.2 \sqrt{k/\Upsilon} + 1.1 k/\Upsilon \right), 1 + 2.2 \sqrt{k/\Upsilon} + 1.1 k/\Upsilon \right),$$

which implies the statement. $\qquad \square$

We will further show in Lemma 4.3 that these points $p^{(i)} (1 \leqslant i \leqslant k)$ exhibit another excellent property: the distance between $p^{(i)}$ and $p^{(j)}$ is inversely proportional to the volume of the *smaller* cluster between S_i and S_j. Therefore, points in S_i of smaller $\operatorname{vol}(S_i)$ are far from points in S_j of bigger $\operatorname{vol}(S_j)$. Notice that, if this were not the case, a misclassification of a small fraction of points in S_j could introduce a large error to S_i.

LEMMA 4.3. *For every $i \neq j$, it holds that*

$$\left\| p^{(i)} - p^{(j)} \right\|^2 \geqslant \frac{\zeta^2}{10 \min \{ \operatorname{vol}(S_i), \operatorname{vol}(S_j) \}},$$

where ζ is defined in (10).

Proof. Let S_i and S_j be two arbitrary clusters. By Lemma 3.4, there exists $1 \leqslant \ell \leqslant k$ such that

$$\left| \beta_i^{(\ell)} - \beta_j^{(\ell)} \right| \geqslant \zeta.$$

By the definition of $p^{(i)}$ and $p^{(j)}$ it follows that

$$\left\| \frac{p^{(i)}}{\| p^{(i)} \|} - \frac{p^{(j)}}{\| p^{(j)} \|} \right\|^2 \geqslant \left(\frac{\beta_i^{(\ell)}}{\sqrt{\sum_{t=1}^{k} \left(\beta_i^{(t)} \right)^2}} - \frac{\beta_j^{(\ell)}}{\sqrt{\sum_{t=1}^{k} \left(\beta_j^{(t)} \right)^2}} \right)^2.$$

By (16), we know that

$$\sqrt{\sum_{\ell=1}^{k} \left(\beta_j^{(\ell)}\right)^2} = \left\|\left(\beta_j^{(1)}, \ldots, \beta_j^{(k)}\right)^{\mathsf{T}}\right\| \in \left(1 - \frac{\zeta}{10}, 1 + \frac{\zeta}{10}\right).$$

Therefore, we have that

$$\left\|\frac{p^{(i)}}{\|p^{(i)}\|} - \frac{p^{(j)}}{\|p^{(j)}\|}\right\|^2 \geqslant \frac{1}{2} \cdot \left(\beta_i^{(\ell)} - \beta_j^{(\ell)}\right)^2 \geqslant \frac{1}{2} \cdot \zeta^2$$

and

$$\left\langle \frac{p^{(i)}}{\|p^{(i)}\|}, \frac{p^{(j)}}{\|p^{(j)}\|} \right\rangle \leqslant 1 - \zeta^2/4.$$

Without loss of generality, we assume that $\|p^{(i)}\|^2 \geqslant \|p^{(j)}\|^2$. By Lemma 4.2, it holds that

$$\left\|p^{(i)}\right\|^2 \geqslant \frac{9}{10 \cdot \mathrm{vol}(S_i)}$$

and

$$\left\|p^{(i)}\right\|^2 \geqslant \left\|p^{(j)}\right\|^2 \geqslant \frac{9}{10 \cdot \mathrm{vol}(S_j)}.$$

Hence, it holds that

$$\left\|p^{(i)}\right\|^2 \geqslant \frac{9}{10 \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}.$$

We can now finish the proof by considering two cases based on $\|p^{(i)}\|$.

Case 1: Suppose that $\|p^{(i)}\| \geqslant 4\|p^{(j)}\|$. We have that

$$\left\|p^{(i)} - p^{(j)}\right\| \geqslant \left\|p^{(i)}\right\| - \left\|p^{(j)}\right\| \geqslant \frac{3}{4}\left\|p^{(i)}\right\|,$$

which implies that

$$\left\|p^{(i)} - p^{(j)}\right\|^2 \geqslant \frac{9}{16}\left\|p^{(i)}\right\|^2 \geqslant \frac{1}{2 \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}.$$

Case 2: Suppose $\|p^{(j)}\| = \alpha\|p^{(i)}\|$ for $\alpha \in (\frac{1}{4}, 1]$. In this case, we have that

$$\begin{aligned}
\left\|p^{(i)} - p^{(j)}\right\|^2 &= \left\|p^{(i)}\right\|^2 + \left\|p^{(j)}\right\|^2 - 2\left\langle \frac{p^{(i)}}{\|p^{(i)}\|}, \frac{p^{(j)}}{\|p^{(j)}\|} \right\rangle \left\|p^{(i)}\right\| \left\|p^{(j)}\right\| \\
&\geqslant \left\|p^{(i)}\right\|^2 + \left\|p^{(j)}\right\|^2 - 2(1 - \zeta^2/4) \cdot \left\|p^{(i)}\right\| \left\|p^{(j)}\right\| \\
&= (1 + \alpha^2)\left\|p^{(i)}\right\|^2 - 2(1 - \zeta^2/4)\alpha \cdot \left\|p^{(i)}\right\|^2 \\
&= (1 + \alpha^2 - 2\alpha + \alpha\zeta^2/2) \cdot \left\|p^{(i)}\right\|^2 \\
&\geqslant \frac{\alpha\zeta^2}{2} \cdot \left\|p^{(i)}\right\|^2 \geqslant \zeta^2 \cdot \frac{1}{10 \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}},
\end{aligned}$$

and the lemma follows. $\qquad\square$

4.3. Approximation guarantees of spectral clustering. Now we analyze why spectral clustering performs well for solving the k-way partitioning problem. We assume that $A_1, \ldots, A_k$ is any k-way partition returned by a k-means algorithm with an approximation ratio of APT.

We map every vertex u to d_u identical points in $\mathbb{R}^k$. This "trick" allows us to bound the volume of the overlap between the clusters retrieved by a k-means algorithm and the optimal ones. For this reason we define the cost function of partition $A_1, \ldots, A_k$ of $V[G]$ by

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

and the optimal clustering cost is defined by

$$\Delta_k^2 \triangleq \min_{\text{partition } A_1, \ldots, A_k} \mathsf{COST}(A_1, \ldots, A_k),$$

i.e., we define the optimal clustering cost in the same way as in (13), except that we look at the embedded points from vertices of G in the definition. From now on, we always refer COST and Δ_k^2 as the COST and optimal COST values of points $\{F(u)\}_{u \in V}$, and for technical reasons every point is counted d_u times. The next lemma gives an upper bound to the cost of the optimal k-means clustering which depends on the gap Υ

LEMMA 4.4. *It holds that* $\Delta_k^2 \leqslant 1.1 k^2 / \Upsilon$.

Proof. Since Δ_k^2 is obtained by minimizing over all partitions $A_1, \ldots, A_k$ and $c_1, \ldots, c_k$, we have

$$(17) \qquad \Delta_k^2 \leqslant \sum_{i=1}^{k} \sum_{u \in S_i} d_u \left\| F(u) - p^{(i)} \right\|^2.$$

Hence the statement follows by applying Lemma 4.1. $\qquad \square$

Since $A_1, \ldots, A_k$ is the output of a k-means algorithm with approximation ratio APT, by Lemma 4.4 we have that $\mathsf{COST}(A_1, \ldots, A_k) \leqslant \mathsf{APT} \cdot 1.1 k^2 / \Upsilon$. We will show that this upper bound of $\mathsf{APT} \cdot 1.1 k^2 / \Upsilon$ suffices to show that this approximate clustering $A_1, \ldots, A_k$ is close to the "actual" clustering $S_1, \ldots, S_k$, in the sense that (i) every A_i has low conductance, and (ii) under a proper permutation $\sigma : \{1, \ldots, k\} \to \{1, \ldots, k\}$, the symmetric difference between A_i and $S_{\sigma(i)}$ is small. The fact is proved by contradiction: If we could always find a set A_i with high symmetric difference with its correspondence $S_{\sigma(i)}$, regardless of how we map $\{A_i\}$ to their corresponding $\{S_{\sigma(i)}\}$, then the COST value will be high, which contradicts the fact that $\mathsf{COST}(A_1, \ldots, A_k) \leqslant \mathsf{APT} \cdot 1.1 k^2 / \Upsilon$. The core of of the whole contradiction arguments is the following technical lemma, whose proof will be presented in the next subsection.

LEMMA 4.5. *Let* $A_1, \ldots, A_k$ *be a partition of* V. *Suppose that, for every permutation of the indices* $\sigma : \{1, \ldots, k\} \to \{1, \ldots, k\}$, *there exists* i *such that* $\mathrm{vol}(A_i \triangle S_{\sigma(i)}) \geqslant 2\varepsilon \, \mathrm{vol}(S_{\sigma(i)})$ *for* $\varepsilon \geqslant 10^5 \cdot k^3 / \Upsilon$; *then* $\mathsf{COST}(A_1, \ldots, A_k) \geqslant 10^{-4} \cdot \varepsilon / k$.

Proof of Theorem 1.2. Let $A_1, \ldots, A_k$ be a k-way partition that achieves an approximation ratio of APT, and let

$$\varepsilon = \frac{2 \cdot 10^5 \cdot k^3 \cdot \mathsf{APT}}{\Upsilon}.$$

We first show that there exists a permutation σ of the indices such that

(18) $$\text{vol}\left(A_i \triangle S_{\sigma(i)}\right) \leqslant \varepsilon \, \text{vol}(S_{\sigma(i)}) \qquad \text{for any } 1 \leqslant i \leqslant k.$$

Assume for contradiction that for all permutation σ there is $1 \leqslant i \leqslant k$ such that

$$\text{vol}(A_i \triangle S_{\sigma(i)}) > \varepsilon \, \text{vol}\left(S_{\sigma(i)}\right).$$

This implies by Lemma 4.5 that

$$\mathsf{COST}(A_1, \ldots, A_k) \geqslant 10 \cdot \mathsf{APT} \cdot k^2 / \Upsilon,$$

which contradicts the fact that $A_1, \ldots, A_k$ is an APT-approximation to a k-way partition, whose corresponding k-means cost is at most $1.1 \cdot \mathsf{APT} \cdot k^2 / \Upsilon$.

Now we assume that $\sigma : \{1, \ldots, k\} \to \{1, \ldots, k\}$ is the permutation satisfying (18), and bound the conductance of every cluster A_i. For any $1 \leqslant i \leqslant k$, the number of leaving edges of A_i is upper bounded by

$$
\begin{aligned}
|\partial(A_i)| &\leqslant \left|\partial\left(A_i \setminus S_{\sigma(i)}\right)\right| + \left|\partial\left(A_i \cap S_{\sigma(i)}\right)\right| \\
&\leqslant \left|\partial\left(A_i \triangle S_{\sigma(i)}\right)\right| + \left|\partial\left(A_i \cap S_{\sigma(i)}\right)\right|.
\end{aligned}
$$

Notice that $|\partial(A_i \triangle S_{\sigma(i)})| \leqslant \varepsilon \, \text{vol}(S_{\sigma(i)})$ by our assumption on σ, and every node in $|\partial(A_i \cap S_{\sigma(i)})|$ either belongs to $\partial S_{\sigma(i)} \setminus S_{\sigma(i)}$ or $\partial(A_i \triangle S_{\sigma(i)})$, and hence

$$
\begin{aligned}
|\partial(A_i)| &\leqslant \varepsilon \, \text{vol}\left(S_{\sigma(i)}\right) + \phi_G\left(S_{\sigma(i)}\right) \text{vol}\left(S_{\sigma(i)}\right) + \varepsilon \, \text{vol}\left(S_{\sigma(i)}\right) \\
&= \left(2\varepsilon + \phi_G\left(S_{\sigma(i)}\right)\right) \text{vol}(S_{\sigma(i)}).
\end{aligned}
$$

On the other hand, we have that

$$\text{vol}(A_i) \geqslant \text{vol}\left(A_i \cap S_{\sigma(i)}\right) \geqslant (1 - 2\varepsilon) \text{vol}(S_{\sigma(i)}).$$

Hence,

$$
\begin{aligned}
\phi_G(A_i) &\leqslant \frac{(2\varepsilon + \phi_G(S_{\sigma(i)})) \text{vol}(S_{\sigma(i)})}{(1 - 2\varepsilon) \text{vol}(S_{\sigma(i)})} \\
&= \frac{2\varepsilon + \phi_G(S_{\sigma(i)})}{1 - 2\varepsilon} \\
&\leqslant 1.1 \cdot \phi_G(S_{\sigma(i)}) + O(\mathsf{APT} \cdot k^3 / \Upsilon). \qquad \square
\end{aligned}
$$

4.4. Proof of Lemma 4.5. It remains to show Lemma 4.5. Our proof is based on the following high-level idea: suppose by contradiction that there is a cluster S_j which is very different from every cluster A_ℓ. Then there is a cluster A_i with significant overlap with two different clusters S_j and $S_{j'}$ (Lemma 4.6). However, we already proved in Lemma 4.3 that any two clusters are far from each other. This implies that the COST value of $A_1, \ldots, A_k$ is high, which leads to a contradiction.

LEMMA 4.6. *Suppose for every permutation $\pi : \{1, \ldots, k\} \to \{1, \ldots, k\}$ there exists an index i such that $\text{vol}(A_i \triangle S_{\pi(i)}) \geqslant 2\varepsilon \, \text{vol}(S_{\pi(i)})$. Then, at least one of the following two cases holds:*

1. for any index i there are indices $i_1 \neq i_2$ and $\varepsilon_i \geqslant 0$ such that

$$\text{vol}(A_i \cap S_{i_1}) \geqslant \text{vol}(A_i \cap S_{i_2}) \geqslant \varepsilon_i \min\{\text{vol}(S_{i_1}), \text{vol}(S_{i_2})\},$$

and $\sum_{i=1}^{k} \varepsilon_i \geqslant \varepsilon$;

2. *there exist indices i, ℓ and $\varepsilon_j \geqslant 0$ such that, for $j \neq \ell$,*

$$\mathrm{vol}(A_i \cap S_\ell) \geqslant \varepsilon_i \, \mathrm{vol}(S_\ell), \quad \mathrm{vol}(A_i \cap S_j) \geqslant \varepsilon_i \, \mathrm{vol}(S_\ell)$$

and $\sum_{i=1}^k \varepsilon_i \geqslant \varepsilon$.

Proof. Let $\sigma : \{1, \ldots, k\} \rightarrow \{1, \ldots, k\}$ be the function defined by

$$\sigma(i) = \underset{1 \leqslant j \leqslant k}{\mathrm{argmax}} \, \frac{\mathrm{vol}(A_i \cap S_j)}{\mathrm{vol}(S_j)}.$$

We first assume that σ is one-to-one, i.e. σ is a permutation. By the hypothesis of the lemma, there exists an index i such that $\mathrm{vol}(A_i \triangle S_{\sigma(i)}) \geqslant 2\varepsilon \, \mathrm{vol}(S_{\sigma(i)})$. Without loss of generality, we assume that $i = 1$ and $\sigma(j) = j$ for $j = 1, \ldots, k$. Notice that

$$(19) \qquad \mathrm{vol}\,(A_1 \triangle S_1) = \sum_{j \neq 1} \mathrm{vol}\,(A_j \cap S_1) + \sum_{j \neq 1} \mathrm{vol}\,(A_1 \cap S_j).$$

Hence, one of the summations on the right-hand side of (19) is at least $\varepsilon \, \mathrm{vol}(S_1)$. Now the proof is based on the case distinction.

Case 1: Assume that $\sum_{j \neq 1} \mathrm{vol}(A_j \cap S_1) \geqslant \varepsilon \, \mathrm{vol}(S_1)$. We define τ_j for $2 \leqslant j \leqslant k$ to be

$$\tau_j = \frac{\mathrm{vol}\,(A_j \cap S_1)}{\mathrm{vol}\,(S_1)}.$$

We have that

$$\sum_{j \neq 1} \tau_j \geqslant \varepsilon,$$

and by the definition of σ it holds that

$$\frac{\mathrm{vol}\,(A_j \cap S_j)}{\mathrm{vol}\,(S_j)} \geqslant \frac{\mathrm{vol}\,(A_j \cap S_1)}{\mathrm{vol}\,(S_1)} = \tau_j$$

for $2 \leqslant j \leqslant k$. Setting $\varepsilon_j = \tau_j$ for $2 \leqslant j \leqslant k$ and $\varepsilon_1 = 0$ finishes the proof of Case 1.

Case 2: Assume that

$$(20) \qquad \sum_{j \neq 1} \mathrm{vol}(A_1 \cap S_j) \geqslant \varepsilon \, \mathrm{vol}(S_1).$$

Let us define τ'_j for $1 \leqslant j \leqslant k, j \neq 1$, to be

$$\tau'_j = \frac{\mathrm{vol}(A_1 \cap S_j)}{\mathrm{vol}\,(S_1)}.$$

Then, (20) implies that

$$\sum_{j \neq 1} \tau'_j \geqslant \varepsilon.$$

The statement in this case holds by assuming $\mathrm{vol}(A_1 \cap S_1) \geqslant \varepsilon \, \mathrm{vol}(S_1)$, since otherwise we have

$$\mathrm{vol}\,(S_1) - \mathrm{vol}\,(A_1 \cap S_1) = \sum_{j \neq 1} \mathrm{vol}\,(A_j \cap S_1) \geqslant (1 - \varepsilon) \, \mathrm{vol}\,(S_1) \geqslant \varepsilon \, \mathrm{vol}\,(S_1),$$

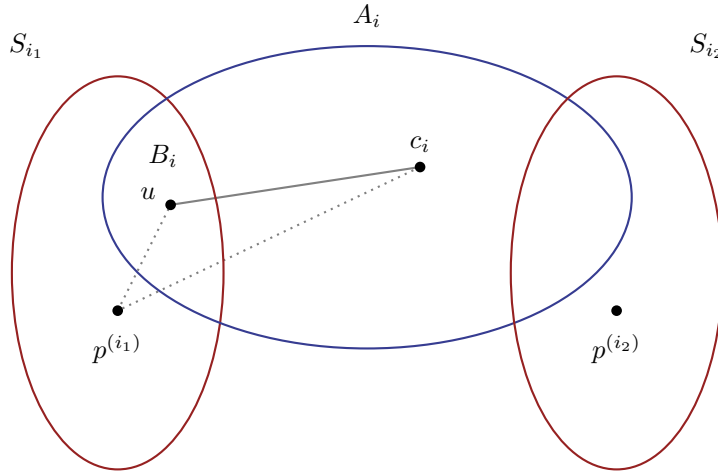and this case was proved in Case 1.

FIG. 2. *We use the fact that $\|p^{(i_1)} - c_i\| \geqslant \|p^{(i_2)} - c_i\|$ and lower bound the value of COST function by only looking at the contribution of points $u \in B_i$ for all $1 \leqslant i \leqslant k$.*

So it suffices to study the case in which σ defined earlier is not one-to-one. Then, there is j $(1 \leqslant j \leqslant k)$ such that $j \notin \{\sigma(1), \ldots, \sigma(k)\}$. For any $1 \leqslant \ell \leqslant k$, let

$$\tau_\ell'' = \frac{\mathrm{vol}(A_\ell \cap S_j)}{\mathrm{vol}(S_j)}.$$

Then, $\sum_{\ell=1}^k \tau_\ell'' = 1 \geqslant \varepsilon$ and it holds for any $1 \leqslant \ell \leqslant k$ that

$$\frac{\mathrm{vol}\left(A_\ell \cap S_{\sigma(\ell)}\right)}{\mathrm{vol}\left(S_{\sigma(\ell)}\right)} \geqslant \frac{\mathrm{vol}(A_\ell \cap S_j)}{\mathrm{vol}(S_j)} = \tau_\ell''. \qquad \qquad \square$$

Proof of Lemma 4.5. We first consider the case when part 1 of Lemma 4.6 holds, i.e., for every i there exist $i_1 \neq i_2$ such that

(21)
$$\mathrm{vol}(A_i \cap S_{i_1}) \geqslant \varepsilon_i \min\left\{\mathrm{vol}(S_{i_1}), \mathrm{vol}(S_{i_2})\right\},$$
$$\mathrm{vol}(A_i \cap S_{i_2}) \geqslant \varepsilon_i \min\left\{\mathrm{vol}(S_{i_1}), \mathrm{vol}(S_{i_2})\right\}$$

for some $\varepsilon \geqslant 0$, and $\sum_{i=1}^k \varepsilon_i \geqslant \varepsilon$.

Let c_i be the center of A_i. Let us assume without loss of generality that $\|c_i - p^{(i_1)}\| \geqslant \|c_i - p^{(i_2)}\|$, which implies $\|p^{(i_1)} - c_i\| \geqslant \|p^{(i_1)} - p^{(i_2)}\|/2$. However, points in $B_i = A_i \cap S_{i_1}$ are far away from c_i; see Figure 2. We lower bound the value of $\mathsf{COST}(A_1, \ldots, A_k)$ by only looking at the contribution of points in the B_is . Notice that by Lemma 4.1 the sum of the squared-distances between points in B_i and $p^{(i_1)}$ is at most k^2/Υ, while the distance between $p^{(i_1)}$ and $p^{(i_2)}$ is large (Lemma 4.3). Therefore, we have that

$$\mathsf{COST}(A_1, \ldots, A_k) = \sum_{i=1}^k \sum_{u \in A_i} d_u \|F(u) - c_i\|^2 \geqslant \sum_{i=1}^k \sum_{u \in B_i} d_u \|F(u) - c_i\|^2.$$

By applying the inequality $a^2 + b^2 \geqslant (a - b)^2/2$, we have that

$$\mathsf{COST}(A_1, \dots, A_k) \geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \left(\frac{\left\| p^{(i_1)} - c_i \right\|^2}{2} - \left\| F(u) - p^{(i_1)} \right\|^2 \right)$$

$$\geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \frac{\left\| p^{(i_1)} - c_i \right\|^2}{2} - \sum_{i=1}^{k} \sum_{u \in B_i} d_u \left\| F(u) - p^{(i_1)} \right\|^2$$

$$(22) \qquad \geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \frac{\left\| p^{(i_1)} - c_i \right\|^2}{2} - \frac{1.1k^2}{\Upsilon}$$

$$\geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \frac{\left\| p^{(i_1)} - p^{(i_2)} \right\|^2}{8} - \frac{1.1k^2}{\Upsilon}$$

$$(23) \qquad \geqslant \sum_{i=1}^{k} \frac{\zeta^2 \operatorname{vol}(B_i)}{80 \min \left\{ \operatorname{vol}(S_{i_1}), \operatorname{vol}(S_{i_2}) \right\}} - \frac{1.1k^2}{\Upsilon}$$

$$\geqslant \sum_{i=1}^{k} \frac{\zeta^2 \varepsilon_i \min \left\{ \operatorname{vol}(S_{i_1}), \operatorname{vol}(S_{i_2}) \right\}}{80 \min \left\{ \operatorname{vol}(S_{i_1}), \operatorname{vol}(S_{i_2}) \right\}} - \frac{1.1k^2}{\Upsilon}$$

$$\geqslant \sum_{i=1}^{k} \frac{\zeta^2 \varepsilon_i}{80} - \frac{1.1k^2}{\Upsilon}$$

$$\geqslant \frac{\zeta^2 \varepsilon}{80} - \frac{1.1k^2}{\Upsilon} \geqslant \frac{\zeta^2 \varepsilon}{100},$$

where (22) follows from Lemma 4.1, (23) follows from Lemma 4.3, and the last inequality follows from the assumption that $\varepsilon \geqslant 10^5 \cdot k^3/\Upsilon$.

Now, suppose that part 2 of Lemma 4.6 holds, i.e., there are indices i, ℓ such that, for any $j \neq \ell$, it holds that

$$\operatorname{vol}(A_i \cap S_\ell) \geqslant \varepsilon_i \operatorname{vol}(S_\ell),$$
$$\operatorname{vol}(A_i \cap S_j) \geqslant \varepsilon_i \operatorname{vol}(S_\ell)$$

for some $\varepsilon \geqslant 0$, and

$$\sum_{i=1}^{k} \varepsilon_i \geqslant \varepsilon.$$

In this case, we only need to repeat the proof by setting, for any $j \neq i$, $B_j = A_i \cap S_j$, $S_{j_1} = S_\ell$, and $S_{j_2} = S_j$. $\qquad \square$

5. Partitioning well-clustered graphs in nearly linear time. In this section we present a nearly linear time algorithm for partitioning well-clustered graphs and prove Theorem 1.3. At a high level, our algorithm follows the general framework of k-means algorithms and consists of two steps: the seeding step and the grouping step. The seeding step chooses k candidate centers such that each one is close to the actual center of a different cluster. The grouping step assigns the remaining vertices to their individual closest candidate centers.

All the proofs for the seeding and grouping steps assume that we have an embedding $\{x(u)\}_{u \in V[G]}$ satisfying the following two conditions:

$$(24) \qquad \left(1 - \frac{1}{10 \log n}\right) \cdot \|F(u)\|^2 \leqslant \|x(u)\|^2 \leqslant \|F(u)\|^2 + \frac{1}{n^5},$$

$$(25) \quad \left(1 - \frac{1}{10 \log n}\right) \cdot \|F(u) - F(v)\|^2 \leqslant \|x(u) - x(v)\|^2 \leqslant \|F(u) - F(v)\|^2 + \frac{1}{n^5}$$

Notice that these two conditions hold trivially if $\{x(u)\}_{u \in V[G]}$ is the spectral embedding $\{F(u)\}_{u \in V[G]}$ or any embedding produced by good approximations of the first k eigenvectors. However, obtaining such embedding becomes nontrivial for a large value of k, as directly computing the first k eigenvectors takes superlinear time. We will present a nearly linear time algorithm that computes an embedding satisfying (24) and (25). By using standard dimensionality reduction techniques that approximately preserve pairwise distances, such as the Johnson–Lindenstrauss transform (see e.g., [11]), we can also always assume that the dimension of the embedding $\{x(u)\}_{u \in V[G]}$ is $d = O(\log^3 n)$. Throughout the whole section, we assume $k = \omega(\text{poly} \log n)$ and $\Upsilon = \widetilde{\Omega}(k^5)$.

This section is organized as follows: sections 5.1 and 5.2 discuss the seeding and grouping steps, assuming that we have an embedding $\{x(u)\}_{u \in V[G]}$ that satisfies (24) and (25), and section 5.3 analyzes the approximation guarantee of the partition returned by the grouping step. In section 5.4, we present an algorithm that computes all required quantities in nearly linear time, assuming that we know the value of λ_k. This assumption on λ_k will be finally removed in section 5.5, and this leads to our final algorithm, which corresponds to Theorem 1.3.

5.1. The seeding step. We proved in section 4.2 that the approximate center $p^{(i)}$ for every $1 \leqslant i \leqslant k$ satisfies

$$\frac{99}{100 \, \text{vol}(S_i)} \leqslant \left\|p^{(i)}\right\|^2 \leqslant \frac{101}{100 \, \text{vol}(S_i)},$$

and most embedded points $F(u)$ are close to their approximate centers. Together with (24) and (25), these two properties imply that, when sampling points $x(u)$ with probability proportional to $d_u \cdot \|x(u)\|^2$, vertices from different clusters will be approximately sampled with the same probability. We will prove that, when sampling $\Theta(k \log k)$ points in this way, with constant probability there is at least one point sampled from each cluster.

In the next step remove the sampled points which are close to each other and call this resulting set $C^\star$. We prove that with constant probability there is exactly one point in $C^\star$ from a cluster. Algorithm 1 below gives a formal description of the seeding step.

Now we analyze Algorithm 1. For any $1 \leqslant i \leqslant k$, we define $\mathcal{E}_i$ to be the sum of the ℓ_2^2-distance between the embedded points from S_i and $p^{(i)}$, i.e.,

$$\mathcal{E}_i \triangleq \sum_{u \in S_i} d_u \left\|F(u) - p^{(i)}\right\|^2.$$

For any parameter $\rho > 0$, we define the radius of S_i with respect to ρ to be

$$R_i^\rho \triangleq \frac{\rho \cdot \mathcal{E}_i}{\text{vol}(S_i)}$$

Algorithm 1 SEEDANDTRIM($k, \{x(u)\}_{u \in V[G]}$).

1: **input**: the number of clusters k, and the embedding $\{x(u)\}_{u \in V[G]}$.
2: Let $K = \Theta(k \log k)$.
3: **for** $i = 1, \dots, K$ **do**
4: 　　Set $c_i = u$ with probability proportional to $d_u \|x(u)\|^2$.
5: **end for**
6: **for** $i = 2, \dots, K$ **do**
7: 　　Delete all c_j with $j < i$ such that $\|x(c_i) - x(c_j)\|^2 < \frac{\|x(c_i)\|^2}{2 \cdot 10^4 k}$.
8: **end for**
9: **return** the remaining sampled vertices.

---

and define $\mathsf{CORE}_i^\rho \subseteq S_i$ to be the set of vertices whose ℓ_2^2-distance to $p^{(i)}$ is at most R_i^ρ, i.e.,

$$(26) \qquad \mathsf{CORE}_i^\rho \triangleq \left\{ u \in S_i : \left\| F(u) - p^{(i)} \right\|^2 \leqslant R_i^\rho \right\}.$$

By the averaging argument it holds that

$$\mathrm{vol}(S_i \setminus \mathsf{CORE}_i^\rho) \leqslant \frac{\sum_{u \in S_i} d_u \left\| F(u) - p^{(i)} \right\|^2}{R_i^\rho} = \frac{\mathrm{vol}(S_i)}{\rho}$$

and

$$(27) \qquad \mathrm{vol}(\mathsf{CORE}_i^\rho) \geqslant \max \left\{ \left(1 - \frac{1}{\rho} \right) \mathrm{vol}(S_i), 0 \right\}.$$

From now on, we set the parameter

$$\alpha \triangleq \Theta(K \log K).$$

We first show that most embedded points of the vertices in S_i are contained in the cores CORE_i^α for $1 \leqslant i \leqslant k$.

LEMMA 5.1. *The following statements hold:*
1. $\sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \|F(u)\|^2 \geqslant 1 - \frac{1}{100K}$.
2. $\sum_{i=1}^k \sum_{u \notin \mathsf{CORE}_i^\alpha} d_u \cdot \|F(u)\|^2 \leqslant \frac{k}{100K}$.

Proof. By the definition of CORE_i^α, it holds that

$$\sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \|F(u)\|^2$$

$$\geqslant \frac{1}{\alpha} \int_0^\alpha \sum_{u \in \mathsf{CORE}_i^\rho} d_u \cdot \|F(u)\|^2 d\rho$$

$$(28) \qquad \geqslant \frac{1}{\alpha} \int_0^\alpha \left(\left\| p^{(i)} \right\| - \sqrt{R_i^\rho} \right)^2 \mathrm{vol}(\mathsf{CORE}_i^\rho) d\rho$$

$$(29) \qquad \geqslant \frac{1}{\alpha} \int_0^\alpha \left(\left\| p^{(i)} \right\|^2 - 2\sqrt{R_i^\rho} \cdot \left\| p^{(i)} \right\| \right) \max \left\{ \left(1 - \frac{1}{\rho} \right) \mathrm{vol}(S_i), 0 \right\} d\rho$$

$$(30) \qquad \geqslant \frac{1}{\alpha} \int_0^\alpha \max \left\{ \left(1 - \left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon \right) - 3\sqrt{\mathcal{E}_i \rho} \right) \left(1 - \frac{1}{\rho} \right), 0 \right\} d\rho,$$

where (28) follows from the fact that for all $u \in \mathsf{CORE}_i^\rho$, $\|F(u)\| \geqslant \|p^{(i)}\| - \sqrt{R_i^\rho}$, (29) from (27), and (30) from the definition of R_i^ρ and the fact that

$$\left\|p^{(i)}\right\|^2 \cdot \mathrm{vol}(S_i) \in \left(1 - \left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right), 1 + 2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right).$$

Since $\mathcal{E}_i \leqslant 1.1k^2/\Upsilon$ by Lemma 4.1, it holds that

$$\sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \|F(u)\|^2$$

$$\geqslant \frac{1}{\alpha} \int_0^\alpha \max\left\{\left(1 - \left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right) - 4\sqrt{k^2\rho/\Upsilon}\right)\left(1 - \frac{1}{\rho}\right), 0\right\} d\rho$$

$$\geqslant \frac{1}{\alpha} \int_0^\alpha \max\left\{1 - \left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right) - 4\sqrt{k^2\rho/\Upsilon} - \frac{1}{\rho}, 0\right\} d\rho$$

$$\geqslant 1 - \left(2.2\sqrt{k/\Upsilon} + 1.1k/\Upsilon\right) - 4k\sqrt{\alpha/\Upsilon} - \frac{\ln\alpha}{\alpha}$$

$$\geqslant 1 - \frac{1}{100K},$$

where the last inequality holds by the assumption on α and Υ.

The second statement follows by the fact that

$$\sum_{i=1}^k \sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \|F(u)\|^2 \geqslant k\left(1 - \frac{1}{100K}\right)$$

and $\sum_{u \in V[G]} d_u \cdot \|F(u)\|^2 = k$. $\square$

The next lemma shows that the embedded points from the same core are close to each other, while the embedded points from different cores are far from each other.

LEMMA 5.2. *The following statements hold:*
1. *For any $1 \leqslant i \leqslant k$ and any two vertices $u, v \in \mathsf{CORE}_i^\alpha$, it holds that*

$$\|x(u) - x(v)\|^2 \leqslant \min\left\{\frac{11\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)}, \frac{\|x(u)\|^2}{2 \cdot 10^4 \cdot k}\right\}.$$

2. *For any $i \neq j$, and $u \in \mathsf{CORE}_i^\alpha, v \in \mathsf{CORE}_j^\alpha$, it holds that*

$$\|x(u) - x(v)\|^2 \geqslant \frac{1}{7000k\,\mathrm{vol}(S_i)} > \frac{\|x(u)\|^2}{10^4 k}.$$

Proof. By the definition of CORE_i^α, it holds for any $u \in \mathsf{CORE}_i^\alpha$ that

$$\left\|F(u) - p^{(i)}\right\| \leqslant \sqrt{R_i^\alpha}.$$

By the triangle inequality, it holds for any $u \in \mathsf{CORE}_i^\alpha$ and $v \in \mathsf{CORE}_i^\alpha$ that $\|F(u) - F(v)\| \leqslant 2\sqrt{R_i^\alpha}$ and

$$\|F(u) - F(v)\|^2 \leqslant 4R_i^\alpha = \frac{4\alpha\mathcal{E}_i}{\mathrm{vol}(S_i)} \leqslant \frac{5\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)},$$

where the last inequality follows from Lemma 4.1. Hence, by (25) it holds that

$$\|x(u) - x(v)\|^2 \leqslant \|F(u) - F(v)\|^2 + \frac{1}{n^5} \leqslant \frac{5\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)} + \frac{1}{n^5} \leqslant \frac{11\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)},$$

where we use the fact that $\frac{1}{n^5} \ll \frac{1}{\mathrm{vol}(S_i)}$. On the other hand, we have that

$$\|F(u)\|^2 \geqslant \left(\left\|p^{(i)}\right\| - \sqrt{R_i^\alpha}\right)^2 \geqslant \frac{9}{10\,\mathrm{vol}(S_i)},$$

where the last inequality follow from Lemma 4.2 and the definition of R_i^α. By (24) and the conditions on α, Υ, it also holds that

$$\|x(u) - x(v)\|^2 \leqslant \frac{5\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)} + \frac{1}{n^5} \leqslant \frac{10\alpha k^2}{\Upsilon}\|F(u)\|^2 \leqslant \frac{\|x(u)\|^2}{2\cdot 10^4\cdot k}.$$

With these we proved the first statement.

Now we turn to the second statement. By the triangle inequality, it holds for any pair of $u \in \mathsf{CORE}_i^\alpha$ and $v \in \mathsf{CORE}_j^\alpha$ that

$$\|F(u) - F(v)\| \geqslant \left\|p^{(i)} - p^{(j)}\right\| - \left\|F(u) - p^{(i)}\right\| - \left\|F(v) - p^{(j)}\right\|.$$

By Lemma 4.3, we have for any $i \neq j$ that

$$\left\|p^{(i)} - p^{(j)}\right\|^2 \geqslant \frac{1}{10^3 k \min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}}.$$

Combining this with the fact that

$$\left\|F(u) - p^{(i)}\right\| \leqslant \sqrt{R_i^\alpha} \leqslant \sqrt{\frac{1.1\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)}},$$

we obtain that

$$\begin{aligned}\|F(u) - F(v)\| &\geqslant \left\|p^{(i)} - p^{(j)}\right\| - \left\|F(u) - p^{(i)}\right\| - \left\|F(v) - p^{(j)}\right\| \\ &\geqslant \sqrt{\frac{1}{10^3 k \min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}}} - \sqrt{\frac{1.1\alpha k^2}{\Upsilon\,\mathrm{vol}(S_i)}} - \sqrt{\frac{1.1\alpha k^2}{\Upsilon\,\mathrm{vol}(S_j)}} \\ &\geqslant \sqrt{\frac{1}{1.1\cdot 10^3 k \min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}}}.\end{aligned}$$

Notice that

$$\|x(u)\|^2 \leqslant \|F(u)\|^2 + \frac{1}{n^5} \leqslant \left(\left\|p^{(i)}\right\| + \sqrt{R_i^\alpha}\right)^2 + \frac{1}{n^5} \leqslant \frac{11}{10\,\mathrm{vol}(S_i)} + \frac{1}{n^5} \leqslant \frac{11}{9\,\mathrm{vol}(S_i)},$$

and therefore we have

$$\|x(u) - x(v)\|^2 \geqslant \left(1 - \frac{1}{10\log n}\right)\|F(u) - F(v)\|^2 \geqslant \frac{1}{7000 k\,\mathrm{vol}(S_i)} > \frac{\|x(u)\|^2}{10^4 k}. \quad \square$$

We next show that, after sampling $\Theta(k\log k)$ vertices, with constant probability the sampled vertices are in the cores $\bigcup_{i=1}^k \mathsf{CORE}_i^\alpha$, and every core contains at least one sampled vertex.

LEMMA 5.3. *Assume that $K = \Omega(k\log k)$ vertices are sampled, in which each vertex is sampled with probability proportional to $d_u \cdot \|x(u)\|^2$. Then, with constant probability the set $C = \{c_1 \ldots c_K\}$ of sampled vertices satisfies the following properties:*

1. *Set C only contains vertices from the cores, i.e., $C \subseteq \bigcup_{i=1}^{k} \mathsf{CORE}_i^{\alpha}$.*
2. *Set C contains at least one vertex from each cluster, i.e., $C \cap S_i \neq \emptyset$ for any $1 \leqslant i \leqslant k$.*

Proof. By (24), it holds for every vertex u that

$$\left(1 - \frac{1}{10 \log n}\right) \cdot \|F(u)\|^2 \leqslant \|x(u)\|^2 \leqslant \|F(u)\|^2 + \frac{1}{n^5}.$$

Since $\sum_{u \in V[G]} d_u \|F(u)\|^2 = k$, it holds that

$$\sum_{u \in V[G]} d_u \|x(u)\|^2 \leqslant \sum_{u \in V[G]} d_u \cdot \left(\|F(u)\|^2 + \frac{1}{n^5}\right) \leqslant k + 1$$

and

$$\sum_{u \in V[G]} d_u \|x(u)\|^2 \geqslant \sum_{u \in V[G]} d_u \cdot \left(1 - \frac{1}{10 \log n}\right) \cdot \|F(u)\|^2 \geqslant \left(1 - \frac{1}{10 \log n}\right) \cdot k,$$

i.e., the total probability mass that we use to sample vertices, i.e., $\sum_{u \in V[G]} d_u \|x(u)\|^2$, is between $(1 - \frac{1}{10 \log n}) \cdot k$ and $k + 1$.

We first bound the probability that we sample at least one vertex from every core. For any fixed $1 \leqslant i \leqslant k$, the probability that a vertex from CORE_i^{α} gets sampled is at least

$$\frac{\sum_{u \in \mathsf{CORE}_i^{\alpha}} d_u \cdot \|x(u)\|^2}{k + 1} \geqslant \frac{\sum_{u \in \mathsf{CORE}_i^{\alpha}} d_u \cdot \|F(u)\|^2}{3(k + 1)} \geqslant \frac{1 - \frac{1}{100K}}{3 \cdot (k + 1)} \geqslant \frac{1}{10k}.$$

Therefore, the probability that we never encounter a vertex from CORE_i^{α} after sampling K vertices is at most $(1 - \frac{1}{10k})^K \leqslant \frac{1}{10k}$. Also, the probability that a sampled vertex is outside the cores of the clusters is at most

$$\frac{\sum_{i=1}^{k} \sum_{u \in S_i \setminus \mathsf{CORE}_i^{\alpha}} d_u \cdot \|x(u)\|^2}{\left(1 - \frac{1}{10 \log n}\right) \cdot k} \leqslant \frac{\sum_{i=1}^{k} \sum_{u \in S_i \setminus \mathsf{CORE}_i^{\alpha}} d_u \cdot \left(\|F(u)\|^2 + n^{-5}\right)}{k/2}$$

$$\leqslant \frac{\frac{k}{100K} + n^{-3}}{k/2} \leqslant \frac{2}{100K} + \frac{1}{n^2}.$$

Taking a union bound over all these events gives that the total probability of undesired events is at most

$$k \cdot \frac{1}{10k} + K \cdot \left(\frac{1}{n^2} + \frac{2}{100K}\right) \leqslant \frac{1}{3}. \qquad \square$$

Based on Lemmas 5.2 and 5.3 we can simply delete one of the two vertices c_i and c_j whose distance is less than $10^{-4} \cdot \|x(c_i)\|^2/(2k)$. The following lemma presents the correctness and runtime of the procedure SEEDANDTRIM, i.e., Algorithm 1.

LEMMA 5.4. *Given the embedding $\{x(u)\}_{u \in V[G]}$ of dimension $d = O(\log^3 n)$ that satisfies (24) and (25), with constant probability the procedure SEEDANDTRIM returns a set $C^{\star}$ of centers $c_1 \ldots c_k$ in $\widetilde{O}(n + k^2)$ time, such that each CORE_i^{α} contains exactly one vertex in $C^{\star}$.*

Proof. Since the sampled set C contains at least one vertex from each core CORE_i^α with constant probability, and only vertices from different cores will remain in $C^\star$ by Lemma 5.2 and the algorithm description, the SEEDANDTRIM procedure returns k centers with constant probability.

Now we analyze the runtime. The procedure takes $\widetilde{O}(n)$ time to compute the norms of $\{x(u)\}_{u \in V[G]}$, since the embedding has dimension $O(\log^3 n)$ by assumption. It takes $\widetilde{O}(k)$ time to sample $\widetilde{O}(k)$ vertices, and trimming the sampling vertices takes $\widetilde{O}(k^2)$ time. Hence, the total runtime is $\widetilde{O}(n + k^2)$. $\qquad\square$

As the end of this subsection, we would like to mention that choosing good candidate centers is crucial for most k-means algorithms and has been studied extensively in the literature (e.g., [6, 31]). Comparing with recent algorithms that obtain good initial centers by iteratively picking points from a *nonuniform* distribution and take $\Omega(nk)$ time, our seeding step (Algorithm 1) runs in $\widetilde{O}(n + k^2)$ time.

5.2. The grouping step. After the seeding step, with constant probability we obtain a set of k vertices $C^\star = \{c_1, \ldots, c_k\}$, and these k vertices belong to k different clusters. Now we assign each remaining vertex u to a cluster S_i if, comparing with all other points $x(c_j)$ with $c_j \in C^\star$, $x(u)$ is closer to $x(c_i)$. A naive implementation of this step requires $\widetilde{\Omega}(nk)$ time. To speed it up, we apply ε-approximate nearest neighbor data structures (ε-NNS) [16], whose formal description is as follows.

PROBLEM 1 (ε-approximate nearest neighbor problem). *Given a set of point $P \subset \mathbb{R}^d$ and a point $q \in \mathbb{R}^d$, find a point $p \in P$ such that, for all $p' \in P$, $\|p - q\| \leqslant (1 + \varepsilon)\|p' - q\|$.*

THEOREM 5.5 (see [16]). *Given a set P of points in $\mathbb{R}^d$, there is an algorithm that solves the ε-approximate nearest neighbor problem with $\widetilde{O}(|P|^{1 + \frac{1}{1+\varepsilon}} + d \cdot |P|)$ preprocessing time and $\widetilde{O}(d \cdot |P|^{\frac{1}{1+\varepsilon}})$ query time.*

Now we set $P = \{x(c_1), \ldots, x(c_k)\}$ and apply the above ε-approximate nearest neighbor data structures to assign the remaining vertices to k clusters $A_1, \ldots, A_k$. By Theorem 5.5 and setting $\varepsilon = \log k - 1$, this step can be finished with $\widetilde{O}(k)$ preprocessing time and $\widetilde{O}(1)$ query time for each query. Hence, the runtime of the grouping step is $\widetilde{O}(n)$. Notice that, with our choice of $\varepsilon = \log k - 1$ and application of ε-NNS, all the remaining vertices in $V \setminus C^\star$ *might not* assign to the cluster A_i with the closest center c_i. We will prove in the next subsection that our choice of ε suffices to obtain a good approximation of the optimal partition. The runtime of the grouping step, and the properties of the returned clusters are summarized in the following lemma.

LEMMA 5.6. *Given a set of centers $C^\star = \{c_1, \ldots, c_k\}$, the grouping step runs in $\widetilde{O}(n)$ time and returns a partition $A_1, \ldots, A_k$ of vertices, such that for any $i \in \{1, \ldots, k\}$, and every $u \in A_i$, it holds for any $j \neq i$ that*

$$\|x(u) - x(c_i)\| \leqslant \log k \cdot \|x(u) - x(c_j)\|.$$

Proof. The statement follows from the definition of ε-NNS, with the choice of $\varepsilon = \log k - 1$, and Theorem 5.5. $\qquad\square$

5.3. Approximation analysis of the algorithm. Now we study the approximation ratio of the k-way partition computed by the seeding and grouping steps. The next lemma analyzes the symmetric difference between the optimal partition and the output of the algorithm.

LEMMA 5.7. *Let $A_1, \ldots, A_k$ be the output of the grouping procedure. Then, under a proper permutation of the indices, with constant probability for any $1 \leqslant i \leqslant k$ it holds that* (i) $\mathrm{vol}(A_i \triangle S_i) = \widetilde{O}(k^3/\Upsilon)\,\mathrm{vol}(S_i)$ *and* (ii) $\phi_G(A_i) = 1.1 \cdot \phi_G(S_i) + \widetilde{O}(k^3/\Upsilon)$.

Proof. We assume that $c_1, \ldots, c_k \in V$ are the centers returned by SEEDANDTRIM, and $\{x(u)\}_{u \in V[G]}$ is the embedding we used in the algorithm. Moreover, $\{x(u)\}_{u \in V[G]}$ satisfies (24) and (25). We further assume that these sampled $c_1, \ldots, c_k \subseteq \bigcup_{i=1}^{k} \mathsf{CORE}_i^\alpha$. By Lemma 5.3, this holds with constant probability, and we assume that this event happens in the following analysis. Then, by the second statement of Lemma 5.2 it holds for any $i \neq j$ that

$$(31) \qquad \|x(c_i) - x(c_j)\|^2 = \Omega\left(\frac{1}{k \cdot \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}\right).$$

By Lemma 5.6, it holds for any $1 \leqslant i \leqslant k$ that

$$\mathrm{vol}(S_i \setminus A_i)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \|x(c_i) - x(v)\| > \frac{\|x(c_j) - x(v)\|}{\log k}\right\}\right)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \|x(c_i) - x(v)\| > \frac{\|x(c_i) - x(c_j)\| - \|x(c_i) - x(v)\|}{\log k}\right\}\right)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : 2\|x(c_i) - x(v)\| > \frac{\|x(c_i) - x(c_j)\|}{\log k}\right\}\right)$$
$$= \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \|x(c_i) - x(v)\| > \frac{\|x(c_i) - x(c_j)\|}{2 \log k}\right\}\right).$$

By (25) and the triangle inequality, we have that

$$\|x(c_i) - x(v)\| \leqslant \|F(c_i) - F(v)\| + \frac{1}{n^{2.5}} \leqslant \left\|F(c_i) - p^{(i)}\right\| + \left\|p^{(i)} - F(v)\right\| + \frac{1}{n^{2.5}},$$

and hence

$$\mathrm{vol}(S_i \setminus A_i)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \left\|F(c_i) - p^{(i)}\right\| + \left\|p^{(i)} - F(v)\right\| + \frac{1}{n^{2.5}} > \frac{\|x(c_i) - x(c_j)\|}{2 \log k}\right\}\right)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \left\|p^{(i)} - F(v)\right\| > \frac{\|x(c_i) - x(c_j)\|}{2 \log k} - \left\|F(c_i) - p^{(i)}\right\| - \frac{1}{n^{2.5}}\right\}\right)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \left\|p^{(i)} - F(v)\right\| > \frac{\|x(c_i) - x(c_j)\|}{2 \log k} - \sqrt{R_i^\alpha} - \frac{1}{n^{2.5}}\right\}\right)$$
$$\leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_i : \left\|p^{(i)} - F(v)\right\|^2 = \Omega\left(\frac{1}{k \log^2 k \cdot \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}\right)\right\}\right)$$
$$= \widetilde{O}\left(k^3/\Upsilon\right) \mathrm{vol}(S_i),$$

where the last equality follows from Lemma 4.1. For the same reason, we have

$$\mathrm{vol}(A_i \setminus S_i) \leqslant \sum_{i \neq j} \mathrm{vol}\left(\left\{v \in S_j : \|x(c_j) - x(v)\| \geqslant \frac{\|x(c_i) - x(v)\|}{\log k}\right\}\right)$$
$$= \widetilde{O}\left(k^3/\Upsilon\right) \mathrm{vol}(S_i),$$

and therefore

$$\mathrm{vol}(S_i \triangle A_i) = \mathrm{vol}(S_i \setminus A_i) + \mathrm{vol}(A_i \setminus S_i) = \widetilde{O}\left(k^3/\Upsilon\right) \mathrm{vol}(S_i).$$

This yields the first statement of the lemma.

The second statement follows by the same argument used in proving Theorem 1.2. $\square$

5.4. Fast computation of the required embedding. So far we assumed the existence of the embedding $\{x(u)\}_{u \in V[G]}$ satisfying (24) and (25) and analyzed the performance of the seeding and grouping steps. In this subsection, we will present a nearly linear time algorithm to compute all the required distances used in the seeding and grouping steps. Our algorithm is based on the so-called heat kernal of a graph.

Formally, the heat kernel of G with parameter $t \geqslant 0$ is defined by

$$(32) \qquad\qquad \mathbf{H}_t \triangleq \mathrm{e}^{-t\mathcal{L}} = \sum_{i=1}^n \mathrm{e}^{-t\lambda_i} f_i f_i^{\mathsf{T}}.$$

We view the heat kernel as a geometric embedding from $V[G]$ to $\mathbb{R}^n$ defined by

$$(33) \qquad\qquad x_t(u) \triangleq \frac{1}{\sqrt{d_u}} \cdot \left(\mathrm{e}^{-t \cdot \lambda_1} f_1(u), \dots, \mathrm{e}^{-t \cdot \lambda_n} f_n(u)\right),$$

and define the ℓ_2^2-distance between the points $x_t(u)$ and $x_t(v)$ by

$$(34) \qquad\qquad \eta_t(u, v) \triangleq \|x_t(u) - x_t(v)\|^2.$$

The following lemma shows that, when $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$, the values of $\eta_t(u, v)$ for all edges $\{u, v\} \in E[G]$ can be approximately computed in $\widetilde{O}(m)$ time.

LEMMA 5.8. *Let $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$. Then, there is $t = O(\mathrm{poly}(n))$ such that the embedding $\{x_t(u)\}_{u \in V[G]}$ defined in (33) satisfies (24) and (25). Moreover, the values of $\eta_t(u, v)$ for all $\{u, v\} \in E[G]$ can be approximately computed in $\widetilde{O}(m)$ time, such that with high probability the conditions (24) and (25) hold for all edges $\{u, v\} \in E[G]$.*

Our proof of Lemma 5.8 uses the algorithm for approximating the matrix exponential in [29] as a subroutine, whose performance is summarised in Theorem 5.9. Recall that any $n \times n$ real and symmetric matrix $\mathbf{A}$ is diagonally dominant (SDD) if $\mathbf{A}_{ii} \geqslant \sum_{j \neq i} |\mathbf{A}_{ij}|$ for each $i = 1, \dots, n$. It is easy to see that the Laplacian matrix of any undirected graph is diagonally dominant.

THEOREM 5.9 (see [29]).   *Given an $n \times n$ SDD matrix $\mathbf{A}$ with $m_{\mathbf{A}}$ nonzero entries, a vector v, and a parameter $\delta > 0$, there is an algorithm that can compute a vector x such that $\|\mathrm{e}^{-\mathbf{A}} v - x\| \leqslant \delta \|v\|$ in time $\widetilde{O}((m_{\mathbf{A}} + n) \log(2 + \|\mathbf{A}\|))$, where the $\widetilde{O}(\cdot)$ notation hides $\mathrm{poly} \log n$ and $\mathrm{poly} \log(1/\delta)$ factors.*

Proof of Lemma 5.8. By the higher-order Cheeger inequality (2), we have that

$$\Upsilon = \frac{\lambda_{k+1}}{\rho(k)} \leqslant \frac{2\lambda_{k+1}}{\lambda_k}.$$

Since $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$, it holds that $400 \cdot \log^2 n \leqslant \lambda_{k+1}/\lambda_k$, and there is t such that

$$t \in \left(\frac{10 \cdot \log n}{\lambda_{k+1}}, \quad \frac{1}{20 \cdot \lambda_k \cdot \log n} \right).$$

We first show that the embedding $\{x_t(u)\}_{u \in V[G]}$ with this t satisfies (24) and (25).

By the definition of $\eta_t(u, v)$, we have that

$$\eta_t(u, v) = \sum_{i=1}^{n} e^{-2t\lambda_i} \left(\frac{f_i(u)}{\sqrt{d_u}} - \frac{f_i(v)}{\sqrt{d_v}} \right)^2$$

$$(35) \qquad = \sum_{i=1}^{k} e^{-2t\lambda_i} \left(\frac{f_i(u)}{\sqrt{d_u}} - \frac{f_i(v)}{\sqrt{d_v}} \right)^2 + \sum_{i=k+1}^{n} e^{-2t\lambda_i} \left(\frac{f_i(u)}{\sqrt{d_u}} - \frac{f_i(v)}{\sqrt{d_v}} \right)^2.$$

Notice that it holds for $1 \leqslant i \leqslant k$ that

$$(36) \qquad 1 - \frac{1}{10 \log n} \leqslant e^{-1/(10 \log n)} \leqslant e^{-\lambda_i/(10\lambda_k/\log n)} \leqslant e^{-2t\lambda_i} \leqslant 1,$$

and it holds for $k + 1 \leqslant i \leqslant n$ that

$$(37) \qquad e^{-2t \cdot \lambda_i} \leqslant e^{-2\lambda_i \cdot 10 \log n/\lambda_{k+1}} \leqslant e^{-10 \log n \lambda_{k+1}/\lambda_{k+1}} = \frac{1}{n^{20}}.$$

Combining (35), (36), and (37), it holds for any $\{u, v\} \in E[G]$ that

$$\left(1 - \frac{1}{10 \cdot \log n} \right) \cdot \|F(u) - F(v)\|^2 \leqslant \eta_t(u, v) \leqslant \|F(u) - F(v)\|^2 + \frac{1}{n^5},$$

which proves the first statement.

Now we show that the distances of $\|x_t(u) - x_t(v)\|$ for all edges $\{u, v\} \in E[G]$ can be approximately computed in nearly linear time. For any vertex $u \in V[G]$, we define $\xi_u \in \mathbb{R}^n$, where $(\xi_u)_v = 1/\sqrt{d_u}$ if $v = u$, and $(\xi_u)_v = 0$ otherwise. Combining (32) with (33) and (34), we have that $\eta_t(u, v) = \|\mathbf{H}_t(\xi_u - \xi_v)\|^2$. We define $\mathbf{Z}$ to be the operator of error δ which corresponds to the algorithm described in Theorem 5.9, and replacing $\mathbf{H}_t$ with $\mathbf{Z}$ we get

$$\left| \|\mathbf{Z}(\xi_u - \xi_v)\| - \eta_t^{1/2}(u, v) \right| \leqslant \delta \|\xi_u - \xi_v\| \leqslant \delta,$$

where the last inequality follows from $d_u, d_v \geqslant 1$. Hence, it holds that

$$(38) \qquad \eta_t^{1/2}(u, v) - \delta \leqslant \|\mathbf{Z}(\xi_u - \xi_v)\| \leqslant \eta_t^{1/2}(u, v) + \delta.$$

By applying the Johnson–Lindenstrauss transform in a way analogous to the computation of effective resistances (e.g., [20] and [37]), we obtain an $O(\varepsilon^{-2} \cdot \log n) \times n$ Gaussian matrix $\mathbf{Q}$, such that with high probability it holds for all u, v that

$$(39) \qquad (1 - \varepsilon) \|\mathbf{Z}(\xi_u - \xi_v)\| \leqslant \|\mathbf{QZ}(\xi_u - \xi_v)\| \leqslant (1 + \varepsilon) \|\mathbf{Z}(\xi_u - \xi_v)\|.$$

Combining (38) and (39) gives us that

$$(1 - \varepsilon)\left(\eta_t^{1/2}(u, v) - \delta\right) \leqslant \|\mathbf{QZ}\left(\xi_u - \xi_v\right)\| \leqslant (1 + \varepsilon)\left(\eta_t^{1/2}(u, v) + \delta\right).$$

Squaring both sides and invoking the inequality $(1 - \varepsilon)\alpha^2 - (1 + \varepsilon^{-1})b^2 \leqslant (a + b)^2 \leqslant (1 + \varepsilon)\alpha^2 + (1 + \varepsilon^{-1})b^2$ gives

$$(1 - 5\varepsilon)\,\eta_t(u, v) - 2\delta^2\varepsilon^{-1} \leqslant \|\mathbf{QZ}\left(\xi_u - \xi_v\right)\|^2 \leqslant (1 + 5\varepsilon)\,\eta_t(u, v) + 2\delta^2\varepsilon^{-1}.$$

Scaling $\mathbf{QZ}$ by a factor of $(1 + 5\varepsilon)^{-1}$ and appending an extra entry in each vector to create an additive distortion of $2\delta\varepsilon^{-1}$ then gives the desired bounds when δ is set to εn^{-6}. To satisfy the conditions (24) and (25) we just need to set $\varepsilon = O(1/\log n)$.

To analyze the runtime of computing $\|\mathbf{QZ}(\xi_u - \xi_v)\|^2$ for all edges $\{u, v\} \in E[G]$, notice that $\mathbf{Q}$ has only $O(\log^3 n)$ rows. We can then run the approximate exponential algorithm from [29] $O(\log^3 n)$ times, where each time we use a different row of $\mathbf{Q}$ as input. Since $\|\mathcal{L}\| \leqslant 2$, by Theorem 5.9 we can compute $\mathbf{QZ}$ in $\widetilde{O}(m)$ time. Notice that $\mathbf{QZ}\xi_u$ is some column of $\mathbf{QZ}$ after rescaling, and therefore we can compute all the required distances in time $\widetilde{O}(m)$.   □

We remark that the proof above shows an interesting property about the embedding (33), i.e., for a large value of k and a certain condition on Υ, there is always a t such that the values of $\eta_t(u, v)$ gives a good approximation of $\|F(u) - F(v)\|^2$ for all edges $\{u, v\} \in E[G]$. A similar intuition which views the heat kernel embedding as a weighted combination of multiple eigenvectors was discussed in [29].

5.5. Proof of Theorem 1.3. We proved in section 5.4 that if $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$, there is a

$$(40) \qquad\qquad t \in \left(\frac{10 \log n}{\lambda_{k+1}}, \quad \frac{1}{20 \cdot \lambda_k \cdot \log n}\right)$$

such that $\{x_t(u)\}_{u \in V[G]}$ satisfies the conditions (24) and (25). Moreover, the values of $\|x_t(u) - x_t(v)\|$ for $\{u, v\} \in E[G]$ can be approximately computed in nearly linear time.[3] However, it is unclear how to approximate λ_k. Furthermore, without this approximation of λ_k, obtaining the desired embedding $\{x(u)\}_{u \in V[G]}$ becomes highly nontrivial.

To overcome this obstacle, we run the seeding and grouping steps for all possible t of the form 2^i, where $t \in \mathbb{N}_{\geqslant 0}$, as it allows us to run the seeding and grouping steps with the right values of t at some point. However, by (34) the distance between any pair of embedded vertices decreases when we increase the value of t. Moreover, all these embedded points $\{x_t(u)\}_{u \in V[G]}$ tend to "concentrate" around a single point for an arbitrary large value of t. To avoid this situation, for every possible t we compute the value of $\sum_{v \in V[G]} d_v \|x_t(v)\|^2$, and the algorithm only moves to the next iteration if

$$(41) \qquad\qquad \sum_{v \in V[G]} d_v \|x_t(v)\|^2 \geqslant k\left(1 - \frac{2}{\log n}\right).$$

---

[3]Lemma 5.8 shows that both of the embedding $\{x_t(u)\}_{u \in V[G]}$ and the embedding that the algorithm computes in nearly linear time satisfy the conditions (24) and (25) with high probability. For ease of discussion, we use $\{x_t(u)\}_{u \in V[G]}$ to express the embedding that the algorithm actually uses.

By Lemma 5.1, (41) is satisfied for all values of t in the right range (40), and the algorithm will not terminate before $t = \lfloor \log n / \lambda_{k+1} \rfloor$. See Algorithm 2 for the formal description of our final algorithm.

Algorithm 2 A nearly linear time graph clustering algorithm, $k = \Omega(\log n)$.

1: **input**: the input graph G, and the number of clusters k
2: Let $t = 2$.
3: **repeat**
4: Let $(c_1, \ldots, c_k) = \text{SEEDANDTRIM}\big(k, \{x_t(u)\}_{u \in V[G]}\big)$.
5: **if** SEEDANDTRIM returns exactly k points **then**
6:    Compute a partition $A_1, \ldots, A_k$ of $V[G]$: for every $v \in V[G]$ assign v to its nearest center c_i using the ε-NNS algorithm with $\varepsilon = \log k - 1$.
7: **end if**
8: Let $t = 2t$
9: **until** $t > n^{10}$ or $\sum_{v \in V[G]} d_v \|x_t\|^2 < k \left(1 - \frac{2}{\log n}\right)$.
10: **return** $(A_1, \ldots, A_k)$.

LEMMA 5.10. *Let $t = \Omega(1/(\lambda_k \cdot \log n))$, and t satisfies (41). Suppose that* SEEDANDTRIM *uses the embedding $\{x_t(u)\}_{u \in V[G]}$ and returns k centers $c_1, \ldots, c_k$. Then, with constant probability, the following statements hold:*

1. *It holds that*

$$\{c_1, \ldots, c_k\} \subseteq \bigcup_{i=1}^{k} \mathsf{CORE}_i^\alpha.$$

2. *These k centers belong to different cores, and it holds for any different i, j that*

$$\|x_t(c_i) - x_t(c_j)\|^2 = \widetilde{\Omega}\left(\frac{1}{k \cdot \mathrm{vol}(S_i)}\right).$$

3. *For any $i = 1, \ldots, k$, it holds that*

$$\sum_{i=1}^{k} \sum_{u \in S_i} d_u \cdot \|x(u) - x(c_i)\|^2 = \widetilde{O}\left(\frac{k^3}{\Upsilon}\right).$$

Proof. Since $\|x_t(u)\|$ is decreasing with respect to the value of t for any vertex u, by Lemma 5.1 for any $t = \Omega(1/(\lambda_k \cdot \log n))$ we have

$$\sum_{i=1}^{k} \sum_{u \notin \mathsf{CORE}_i^\alpha} d_u \cdot \|x_t(u)\|^2 \leqslant \sum_{i=1}^{k} \sum_{u \notin \mathsf{CORE}_i^\alpha} d_u \cdot \left(\|F(u)\|^2 + \frac{1}{n^5}\right) \leqslant \frac{k}{100K} + \frac{kn^2}{n^5} \leqslant \frac{1}{\log k}.$$

On the other hand, we only consider values of t satisfying (41). Since every vertex u is sampled with probability proportional to $d_u \cdot \|x_t(u)\|^2$, with constant probability it holds that

$$\{c_1, \ldots, c_k\} \subseteq \bigcup_{i=1}^{k} \mathsf{CORE}_i^\alpha,$$

which proves the first statement.

Now we prove that these k centers belong to different cores. We fix an index i and assume that $c_i \in S_i$. We will prove that

$$(42) \qquad \|x_t(c_i)\|^2 = \widetilde{\Omega}\left(\frac{1}{\mathrm{vol}(S_i)}\right).$$

Assume by contradiction that (42) does not hold, i.e.,

$$\|x_t(c_i)\|^2 = o\left(\frac{1}{\log^c k \, \mathrm{vol}(S_i)}\right)$$

for any constant c. Then, we have that

$$\sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \|x_t(u)\|^2 \leqslant \sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \left(\|x_t(c_i)\| + \sqrt{R_i^\alpha}\right)^2$$

$$\leqslant 2 \cdot \sum_{u \in \mathsf{CORE}_i^\alpha} \left(d_u \cdot \|x_t(c_i)\|^2 + d_u \cdot R_i^\alpha\right)$$

$$= o\left(\frac{1}{\log^c k}\right) + o\left(\frac{1}{k^2}\right)$$

$$= o\left(\frac{1}{\log^c k}\right).$$

Combining this with (41), the probability that vertices get sampled from CORE_i^α is

$$\frac{\sum_{u \in \mathsf{CORE}_i^\alpha} d_u \cdot \|x_t(u)\|^2}{\sum_{v \in V[G]} d_v \|x_t(v)\|^2} = o\left(\frac{1}{k \cdot \log^c k}\right).$$

This means if we sample $K = \Theta(k \log k)$ vertices, vertices in CORE_i^α will not get sampled with probability at least $1 - 1/\log^5 k$. This contradicts the fact that $c_i \in \mathsf{CORE}_i^\alpha$. Therefore (42) holds.

Now, by description of Algorithm 1, we have for any $j \neq i$

$$\|x_t(c_i) - x_t(c_j)\|^2 \geqslant \frac{\|x(c_i)\|^2}{2 \cdot 10^4 \cdot k} = \widetilde{\Omega}\left(\frac{1}{k \cdot \mathrm{vol}(S_i)}\right),$$

where the last equality follows from (42). Since any vertex in CORE_i^α has distance at most R_i^α from c_i, c_j and c_i belong to different cores. Therefore, the second statement holds.

Finally we turn our attention to the third statement. We showed in Lemma 5.8 that, when $t = \Theta(1/(\lambda_k \cdot \log n))$, the embedding $\{x_t(u)\}_{u \in V[G]}$ satisfies the conditions (24) and (25). Hence, it holds that

$$\sum_{i=1}^k \sum_{u \in S_i} d_u \cdot \|x(u) - x(c_i)\|^2$$

$$\leqslant \sum_{i=1}^k \sum_{u \in S_i} \left(d_u \cdot \|F(u) - F(c_i)\|^2 + \frac{1}{n^5}\right)$$

$$\leqslant \sum_{i=1}^k \sum_{u \in S_i} \left(d_u \cdot (\|F(u) - p_i\| + \|F(c_i) - p_i\|)^2 + \frac{1}{n^5}\right)$$

$$(43) \qquad \leqslant \sum_{i=1}^k \sum_{u \in S_i} \left(2 \cdot d_u \cdot \left(\|F(u) - p_i\|^2 + \|F(c_i) - p_i\|^2\right) + \frac{1}{n^5}\right).$$

Notice that by Lemma 4.1 we have

$$(44) \qquad \sum_{i=1}^{k} \sum_{u \in S_i} d_u \cdot \|F(u) - p_i\|^2 \leqslant 1.1 k^2/\Upsilon.$$

On the other hand, we have $\|F(c_i) - p_i\|^2 \leqslant R_i^\alpha$ as $c_i \in \mathsf{CORE}_i^\alpha$, and

$$(45) \quad \sum_{i=1}^{k} \sum_{u \in S_i} 2 \cdot d_u \cdot \|F(c_i) - p_i\|^2 \leqslant \sum_{i=1}^{k} 2 \operatorname{vol}(S_i) \cdot \frac{\alpha \cdot \mathcal{E}_i}{\operatorname{vol}(S_i)} = \sum_{i=1}^{k} 2\alpha \cdot \mathcal{E}_i = \widetilde{O}\left(\frac{k^3}{\Upsilon}\right).$$

Combining (43) with (44) and (45), we have that

$$\sum_{i=1}^{k} \sum_{u \in S_i} d_u \cdot \|x(u) - x(c_i)\|^2 \leqslant \widetilde{O}\left(\frac{k^3}{\Upsilon}\right) + \sum_{u \in V[G]} \frac{d_u}{n^5} = \widetilde{O}\left(\frac{k^3}{\Upsilon}\right).$$

Moreover, by (33) and (34) it is straightforward to see that the distance between any embedded vertices decreases as we increase the value of t. Hence, the statement holds for any $t = \Omega(1/(\lambda_k \cdot \log n))$. $\qquad\blacksquare$

LEMMA 5.11. *Let $A_1, \ldots, A_k$ be a k-way partition returned by Algorithm 2. Then, under a proper permutation of the indices, with constant probability for any $1 \leqslant i \leqslant k$ it holds that* (i) $\operatorname{vol}(A_i \triangle S_i) = \widetilde{O}(k^4/\Upsilon) \operatorname{vol}(S_i)$, *and* (ii) $\phi_G(A_i) = 1.1 \cdot \phi_G(S_i) + \widetilde{O}(k^4/\Upsilon)$.

Proof. We assume that $c_1, \ldots, c_k$ are the centers returned by SEEDANDTRIM when obtaining $A_1, \ldots, A_k$. By Lemma 5.10, with constant probability it holds that $\{c_1, \ldots, c_k\} \subseteq \bigcup_{i=1}^{k} \mathsf{CORE}_i^\alpha$, and c_i and c_j belong to different cores for $i \neq j$. Without loss of generality, we assume that $c_i \in \mathsf{CORE}_i^\alpha$. Then, it holds that

$$\operatorname{vol}(S_i \setminus A_i)$$

$$\leqslant \sum_{i \neq j} \operatorname{vol}\left(\left\{v \in S_i : \|x(c_i) - x(v)\| \geqslant \frac{\|x(c_j) - x(v)\|}{\log k}\right\}\right)$$

$$\leqslant \sum_{i \neq j} \operatorname{vol}\left(\left\{v \in S_i : \|x(c_i) - x(v)\| \geqslant \frac{\|x(c_i) - x(c_j)\| - \|x(c_i) - x(v)\|}{\log k}\right\}\right)$$

$$\leqslant \sum_{i \neq j} \operatorname{vol}\left(\left\{v \in S_i : 2\|x(c_i) - x(v)\| \geqslant \frac{\|x(c_i) - x(c_j)\|}{\log k}\right\}\right)$$

$$(46) \qquad \leqslant \sum_{i \neq j} \operatorname{vol}\left(\left\{v \in S_i : \|x(c_i) - x(v)\|^2 = \widetilde{\Omega}\left(\frac{1}{k \min\{\operatorname{vol}(S_j), \operatorname{vol}(S_i)\}}\right)\right\}\right)$$

$$(47) \qquad = \widetilde{O}\left(k^4/\Upsilon\right) \operatorname{vol}(S_i),$$

where (46) follows from the second statement of Lemma 5.10.

Similarly, we also have that

$$\operatorname{vol}(A_i \setminus S_i) \leqslant \sum_{i \neq j} \operatorname{vol}\left(\left\{v \in S_j : \|x(c_j) - x(v)\| \geqslant \frac{\|x(c_i) - x(v)\|}{\log k}\right\}\right)$$

$$= \widetilde{O}\left(k^4/\Upsilon\right) \operatorname{vol}(S_i).$$

This yields the first statement of the lemma. The second statement follows by the same argument used in proving Theorem 1.2. $\qquad\blacksquare$

Proof of Theorem 1.3. The approximation guarantee of the returned partition is shown in Lemma 5.11. For the runtime, notice that we enumerate at most $O(\text{poly} \log n)$ possible values of t. Furthermore, and for every such possible value of t, the algorithm runs in $\widetilde{O}(m)$ time. This includes computing the distances of embedded points and the seeding/grouping steps. Hence, the total runtime is $\widetilde{O}(m)$.   ☐

REFERENCES

[1] Z. Allen-Zhu, S. Lattanzi, and V. S. Mirrokni, *A local algorithm for finding well-connected clusters*, in Proceedings of the 30th International Conference on Machine Learning (ICML'13), 2013, pp. 396–404.
[2] S. Arora, B. Barak, and D. Steurer, *Subexponential algorithms for unique games and related problems*, in Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS'10), 2010, pp. 563–572.
[3] S. Arora, E. Hazan, and S. Kale, *The multiplicative weights update method: A meta-algorithm and applications*, Theory Comput., 8 (2012), pp. 121–164.
[4] S. Arora and S. Kale, *A combinatorial, primal-dual approach to semidefinite programs*, in Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC'07), 2007, pp. 227–236.
[5] S. Arora, S. Rao, and U. V. Vazirani, *Expander flows, geometric embeddings and graph partitioning*, J. ACM, 56 (2009).
[6] D. Arthur and S. Vassilvitskii, *k-means++: The advantages of careful seeding*, in Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07), 2007, pp. 1027–1035.
[7] P. Awasthi, A. Blum, and O. Sheffet, *Center-based clustering under perturbation stability*, Inform. Process. Lett., 112 (2012), pp. 49–54.
[8] S. Ben-David, *A theoretical approach to the clustering selection problem*, in Proceedings of the 4th MultiClust Workshop on Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering, 2013.
[9] F. R. K. Chung, *A local graph partitioning algorithm using heat kernel pagerank*, Internet Math., 6 (2009), pp. 315–330.
[10] G. B. Coleman and H. C. Andrews, *Image segmentation by clustering*, Proc. IEEE, 67 (1979), pp. 773–785.
[11] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures Algorithms, 22 (2003), pp. 60–65.
[12] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. III*, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
[13] T. K. Dey, A. Rossi, and A. Sidiropoulos, *Spectral Concentration, Robust k-Center, and Simple Clustering*, arXiv:1404.3918, 2014.
[14] S. Fortunato, *Community detection in graphs*, Phys. Rep., 486 (2010), pp. 75–174.
[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 2012.
[16] P. Indyk and R. Motwani, *Approximate nearest neighbors: Towards removing the curse of dimensionality*, in Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC'98), 1998, pp. 604–613.
[17] R. Kannan, S. Vempala, and A. Vetta, *On clusterings: Good, bad and spectral*, J. ACM, 51 (2004), pp. 497–515.
[18] J. A. Kelner, Y. T. Lee, L. Orecchia, and A. Sidford, *An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generaliza-*

tions, in Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'14), 2014, pp. 217–226.

[19] P. KOLEV AND K. MEHLHORN, *A Note on Spectral Clustering*, CoRR, abs/1509.09188, 2015.

[20] I. KOUTIS, A. LEVIN, AND R. PENG, *Improved spectral sparsification and numerical algorithms for SDD matrices*, in Proceedings of the 29th International Symposium on Theoretical Aspects of Computer Science (STACS'12), 2012, pp. 266–277.

[21] T. C. KWOK, L. C. LAU, Y. T. LEE, S. OVEIS GHARAN, AND L. TREVISAN, *Improved Cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap*, in Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC'13), 2013, pp. 11–20.

[22] J. R. LEE, S. OVEIS GHARAN, AND L. TREVISAN, *Multi-way spectral partitioning and higher-order Cheeger inequalities*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC'12), 2012, pp. 1117–1130.

[23] F. T. LEIGHTON AND S. RAO, *Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms*, J. ACM, 46 (1999), pp. 787–832.

[24] A. LOUIS, P. RAGHAVENDRA, P. TETALI, AND S. VEMPALA, *Many sparse cuts via higher eigenvalues*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC'12), 2012, pp. 1131–1140.

[25] K. MAKARYCHEV, Y. MAKARYCHEV, AND A. VIJAYARAGHAVAN, *Correlation clustering with noisy partial information*, in Proceedings of the 28th Conference on Learning Theory (COLT'15), 2015, pp. 1321–1342.

[26] D. W. MATULA AND F. SHAHROKHI, *Sparsest cuts and bottlenecks in graphs*, Discrete Appl. Math., 27 (1990), pp. 113–123.

[27] F. MCSHERRY, *Spectral partitioning of random graphs*, in Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS'01), 2001, pp. 529–537.

[28] A. Y. NG, M. I. JORDAN, AND Y. WEISS, *On spectral clustering: Analysis and an algorithm*, Adv. Neural Inform. Process. Syst., 2 (2002), pp. 849–856.

[29] L. ORECCHIA, S. SACHDEVA, AND N. K. VISHNOI, *Approximating the exponential, the Lanczos method and an $\widetilde{O}(m)$-time spectral algorithm for balanced separator*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC'12), 2012, pp. 1141–1160.

[30] L. ORECCHIA, L. J. SCHULMAN, U. V. VAZIRANI, AND N. K. VISHNOI, *On partitioning graphs via single commodity flows*, in Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC'08), 2008, pp. 461–470.

[31] R. OSTROVSKY, Y. RABANI, L. J. SCHULMAN, AND C. SWAMY, *The effectiveness of Lloyd-type methods for the k-means problem*, J. ACM, 59 (2012).

[32] S. OVEIS GHARAN AND L. TREVISAN, *Partitioning into expanders*, in Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'14), 2014, pp. 1256–1266.

[33] K. ROHE, S. CHATTERJEE, AND B. YU, *Spectral clustering and the high-dimensional stochastic blockmodel*, Ann. Statist., 39 (2011), pp. 1878–1915.

[34] L. SALOFF-COSTE, *Lectures on finite Markov chains*, in Lectures on Probability Theory and Statistics, P. Bernard, ed., Lecture Notes in Math. 1665, Springer-Verlag, Berlin, 1997, pp. 301–413.

[35] J. SHERMAN, *Breaking the multicommodity flow barrier for $O(\sqrt{\log n})$-approximations to sparsest cut*, in Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS'09), 2009, pp. 363–372.

[36] J. SHI AND J. MALIK, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intelligence, 22 (2000), pp. 888–905.

[37] D. A. SPIELMAN AND N. SRIVASTAVA, *Graph sparsification by effective resistances*, SIAM J. Comput., 40 (2011), pp. 1913–1926.

[38] D. A. SPIELMAN AND S.-H. TENG, *Spectral sparsification of graphs*, SIAM J. Comput., 40 (2011), pp. 981–1025.

[39] L. TREVISAN, *Approximation algorithms for unique games*, Theory Comput., 4 (2008), pp. 111–128.

[40] U. VON LUXBURG, *A tutorial on spectral clustering*, Statist. Comput., 17 (2007), pp. 395–416.

[41] V. VU, *A Simple SVD Algorithm for Finding Hidden Partitions*, arXiv:1404.3918, 2014.