JMLR: Workshop and Conference Proceedings vol 40:1–33, 2015

Partitioning Well-Clustered Graphs: Spectral Clustering Works!

Richard Peng RPENG@MIT.EDU
Massachusetts Institute of Technology, Cambridge, USA

He Sun H.SUN@BRISTOL.AC.UK
Luca Zanetti LUCA.ZANETTI@BRISTOL.AC.UK
University of Bristol, Bristol, UK

Abstract

In this work we study the widely used *spectral clustering* algorithms, i.e. partition a graph into k clusters via (1) embedding the vertices of a graph into a low-dimensional space using the bottom eigenvectors of the Laplacian matrix, and (2) partitioning the embedded points via k-means algorithms. We show that, for a wide class of graphs, spectral clustering algorithms give 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, which is based on heat kernel embeddings and approximate nearest neighbor data structures.

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 (Matula and Shahrokhi, 1990) and, assuming the Small Set Expansion Conjecture (Raghavendra et al., 2012), does not admit a polynomial-time algorithm which achieves a constant factor approximation in the worst case.

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

© 2015 R. Peng, H. Sun & L. Zanetti.

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

$$\rho(k) \triangleq \min_{\text{partition } A_1,\ldots,A_k} \max_{1\leqslant i\leqslant k} \phi_G(A_i). \tag{1.1}$$

Clusters of low conductance in a real network usually capture the 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 (Coleman and Andrews, 1979), which in turn rely on partitioning graphs into multiple subsets (Shi and Malik, 2000). On a theoretical side, decomposing vertex/edge sets into multiple disjoint subsets is used in designing approximation algorithms for Unique Games (Trevisan, 2008), and efficient algorithms for graph problems (Kelner et al., 2014; Leighton and Rao, 1999; Spielman and Teng, 2011).

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 (Lee et al., 2012; Louis et al., 2012). For instance, Lee et al. (2012) proved the following higher-order Cheeger inequality:

$$\frac{\lambda_k}{2} \leqslant \rho(k) \leqslant O(k^2)\sqrt{\lambda_k}, \tag{1.2}$$

where $0 = \lambda_1 \leqslant \ldots \leqslant \lambda_n \leqslant 2$ 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, (1.2) 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* Υ for some k, defined by

$$\Upsilon \triangleq \frac{\lambda_{k+1}}{\rho(k)}, \tag{1.3}$$

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

In this paper we study the well-clustered graphs which satisfy a gap assumption on Υ, and similar assumptions were addressed in previous reference, e.g. (Zhu et al., 2013; Oveis Gharan and Trevisan, 2014). To further reason the rationality of the assumption on Υ, first notice that if the gap assumption does not hold, then G can be partitioned into at least $k+1$ subsets of low conductance; secondly, the assumption on Υ is closely related to the gap between λ_{k+1} and λ_k for some k. The value of k for which there is a gap between λ_{k+1} and λ_k has been observed empirically as an indication of the correct number of clusters (Fortunato, 2010; von Luxburg, 2007).

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 k-means algorithms, 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 locality-sensitive hashing, we give a nearly-linear time algorithm for the k-way partitioning problem.

Our structural results can be summarized as follows. Let $\{S_i\}_{i=1}^k$ be a k-way partition of G achieving $\rho(k)$ defined in (1.1). We define $\{\bar{g}_i\}_{i=1}^k$ to be the normalized indicator vectors of the clusters $\{S_i\}_{i=1}^k$, and $\{f_i\}_{i=1}^k$ to be the eigenvectors corresponding to the k smallest eigenvalues of $\mathcal{L}$. We show that, under the condition of $\Upsilon = \Omega(k^2)$, the span of $\{\bar{g}_i\}_{i=1}^k$ and the span of $\{f_i\}_{i=1}^k$ are close to *each other*, which can be 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 $\Upsilon = \lambda_{k+1}/\rho(k) = \Omega(k^2)$. Assume that $\{f_i\}_{i=1}^k$ are the first k eigenvectors of matrix $\mathcal{L}$, and $\bar{g}_1, \ldots, \bar{g}_k \in \mathbb{R}^n$ are the indicator vectors of $\{S_i\}_{i=1}^k$ with proper normalization[1]. 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 $\|\bar{g}_i - \hat{f}_i\|^2 \leqslant 1/\Upsilon$.*
2. *For every f_i, there is a linear combination of $\{\bar{g}_i\}_{i=1}^k$, called $\hat{g}_i$, such that $\|f_i - \hat{g}_i\|^2 \leqslant 1.1k/\Upsilon$.*

This theorem generalizes the result shown by Arora et al. (Arora et al. (2010), 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 (Davis and Kahan, 1970). We remark that, despite that we use the higher-order Cheeger inequality (1.2) to motivate the definition of Υ, 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 (Kwok et al., 2013; Lee et al., 2012).

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

$$F(u) \triangleq \frac{1}{\mathsf{NormalizationFactor}(u)} \cdot (f_1(u), \ldots, f_k(u))^{\mathsf{T}}, \tag{1.4}$$

where $\mathsf{NormalizationFactor}(u) \in \mathbb{R}$ is the normalization factor for $u \in V$. 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 Υ, the higher concentration the embedded points within the same cluster.

Based on these facts, we analyze the performance of spectral k-means algorithms[2], aiming at answering the following longstanding open question: *Why do spectral k-means algorithms perform well in practice?* We show that the partition $\{A_i\}_{i=1}^k$ produced by the spectral k-means algorithm 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 correspondence S_i. This allows us to apply various k-means algorithms (Kumar et al., 2004; Ostrovsky et al., 2012) to give spectral clustering algorithms, with different time versus approximation tradeoffs. These algorithms have comprehensive applications, and have been the subject of extensive experimental studies for more than 20 years, e.g. (Ng et al.,

1. See the formal definition in Section 3.
2. For simplicity, we use the word "spectral k-means algorithms" to refer to the algorithms which combine a spectral embedding with a k-means algorithm in Euclidean space.

2002; von Luxburg, 2007). While similar results on spectral clustering mainly focus on graphs in the stochastic block model, to the best of our knowledge it is the first rigorous analysis of spectral clustering for general graphs exhibiting a multi-cut structure. Our result is as follows:

Theorem 1.2 (Approximation Guarantee of Spectral k-Means Algorithms) *Let G be a graph satisfying the condition $\Upsilon = \lambda_{k+1}/\rho(k) = \Omega(k^3)$, and $k \in \mathbb{N}$. Let $F : V \to \mathbb{R}^k$ be the embedding defined above. 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) $\mathrm{vol}(A_i \triangle S_i) = O\left(\mathsf{APT} \cdot k^3 \cdot \Upsilon^{-1} \cdot \mathrm{vol}(S_i)\right)$, and (ii) $\phi_G(A_i) = O\left(\phi_G(S_i) + \mathsf{APT} \cdot k^3 \cdot \Upsilon^{-1}\right)$.*

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) obtaining the spectral embedding (1.4) requires $\Omega(mk)$ time for computing k eigenvectors, and (ii) most k-means algorithms run in time $\Omega(nk)$.

To overcome the first obstacle, we study the so-called *heat kernel embedding*, an embedding from V to $\mathbb{R}^n$, and approximates the squared-distance $\|F(u) - F(v)\|^2$ of the embedded points $F(u)$ and $F(v)$ via their *heat-kernel distance*. Since a desired approximation of the heat kernel distances of vertices is computable in nearly-linear time, this approach avoids the computation of eigenvectors. For the second obstacle, instead of applying k-means algorithms as a black-box we apply approximate nearest-neighbor algorithms. Our approach presents an ad-hoc version of k-means algorithms, and indicates that in certain scenarios the standard Lloyd-type heuristic widely used in k-means algorithms can eventually be avoided. Our third 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 a parameter $k \in \mathbb{N}$. Assume that $\Upsilon = \lambda_{k+1}/\rho(k) = \Omega(k^5)$, 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[3] and outputs a k-way partition $\{A_i\}_{i=1}^k$ such that (i) $\mathrm{vol}(A_i \triangle S_i) = O\left(k^3 \log^2 k \cdot \Upsilon^{-1} \cdot \mathrm{vol}(S_i)\right)$, and (ii) $\phi_G(A_i) = O(\phi_G(S_i) + k^3 \log^2 k \cdot \Upsilon^{-1})$.*

We remark that bounds of other expansion parameters of k-way partitioning can be derived from our analysis as well. For instance, one quantity studied extensively in machine learning is the *normalized cut* (Shi and Malik, 2000). While the k-way expansion of a graph is the maximum conductance of the sets in the partition, the corresponding normalized cut is the sum of the conductance of all the sets in this partition. It is easy to see that these two measures differ by at most a factor of k, and the normallized cut for a k-way partition from spectral clustering can be bounded as well.

1.2. Related Work

There is a large amount of literature on graph partitioning under various settings. Oveis Gharan and Trevisan (2014) 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 satisfy the inner- and outer-conductance condition. In order to assure that every A_i has high inner conductance, they assume that $\lambda_{k+1} \geqslant$

3. The $\widetilde{O}(\cdot)$ term hides a factor of $\mathrm{poly}\log n$.

$\mathrm{poly}(k)\lambda_k^{1/4}$, which is much stronger than ours. Moreover, their algorithm runs in polynomial-time, in contrast to our nearly-linear time algorithm.

Dey et al. (2014) studies the properties of the spectral embedding for graphs having a gap between λ_k and λ_{k+1} and presents a k-way partition algorithm, which is based on the 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 (Chung, 2009), balanced separators (Orecchia et al., 2012). It also plays a key role in current efficient approximation algorithms for finding low conductance cuts (Orecchia et al., 2008; Sherman, 2009). However, most of these theoretical guarantees are through the matrix multiplicative weights update framework (Arora et al., 2012; Arora and Kale, 2007). 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) (McSherry, 2001), 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 proven that spectral algorithms give the correct clustering for certain ranges of p and q, e.g. (McSherry, 2001; Rohe et al., 2011; Vu, 2014). 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 (Davis and Kahan, 1970), 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, aka $E(S, T) \triangleq \{\{u, v\} | u \in S \text{ and } v \in T\}$. 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 will work extensively with algebraic objects related to G. We will also 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 and $\mathbf{D}$ is the $n \times n$ diagonal matrix with $\mathbf{D}_{uu} = d_u$ for $u \in V[G]$. 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 will 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 non-zero constant vector.

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

$$\mathcal{R}(f) \triangleq \frac{f^\intercal \mathcal{L} f}{\|f\|_2^2} = \frac{f^\intercal \mathbf{L} f}{\|f\|_{\mathbf{D}}} = \frac{\sum_{\{u,v\}\in E(G)} \left(f(u) - f(v)\right)^2}{\sum_u d_u f(u)^2},$$

where $\|f\|_{\mathbf{D}} \triangleq f^\intercal \mathbf{D} f$. Based on the Rayleigh quotient, the conductance of a set S_i can be expressed as $\phi_G(S_i) = \mathcal{R}(\bar{g}_i)$, and the gap Υ can be written as

$$\Upsilon = \frac{\lambda_{k+1}}{\rho(k)} = \min_{1\leqslant i\leqslant k} \frac{\lambda_{k+1}}{\phi_G(S_i)} = \min_{1\leqslant i\leqslant k} \frac{\lambda_{k+1}}{\mathcal{R}(\bar{g}_i)}. \tag{2.1}$$

Throughout the rest of the paper, we will use $S_1, \ldots, S_k$ to express a k-way partition of G achieving the minimum conductance $\rho(k)$. 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

$$g_i(u) = \begin{cases} 1 & \text{if } u \in S_i, \\ 0 & \text{if } u \notin S_i, \end{cases} \tag{3.1}$$

and define the corresponding normalized indicator vector by

$$\bar{g}_i = \frac{\mathbf{D}^{1/2} g_i}{\|\mathbf{D}^{1/2} g_i\|}. \tag{3.2}$$

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 $f_1, \cdots, f_k$ and $\bar{g}_1, \cdots, \bar{g}_k$ 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 $\{\bar{g}_i\}_{i=1}^k$, and (ii) every indicator vector $\bar{g}_i$ 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.

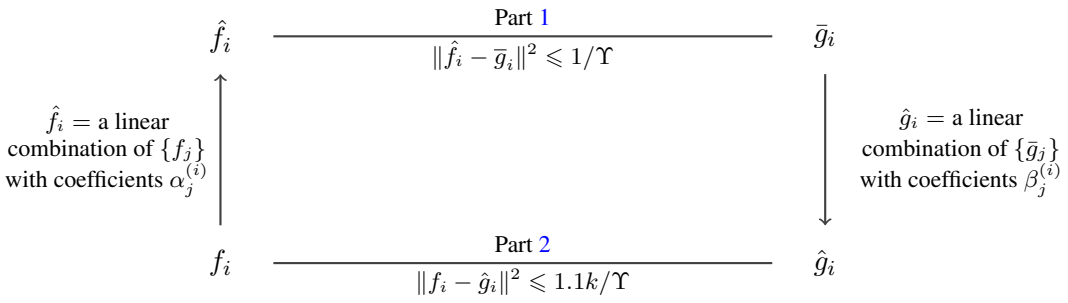

Figure 1: Relations among $\{\hat{f}_i\}$, $\{f_i\}$, $\{\bar{g}_i\}$, and $\{\hat{g}_i\}$ given in Theorem 3.1. Here Υ is the gap defined with respect to λ_{k+1} and $\rho(k)$.

Theorem 3.1 *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 $\left\| \bar{g}_i - \hat{f}_i \right\|^2 \leqslant 1/\Upsilon$.*
2. *There is a linear combination of the vectors $\bar{g}_1, \ldots, \bar{g}_k$ with coefficients $\beta_j^{(i)}$: $\hat{g}_i = \beta_1^{(i)} \bar{g}_1 + \cdots + \beta_k^{(i)} \bar{g}_k$, 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 Theorem 2.2 of Arora et al. (2010).

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 $\{\bar{g}_i\}_{i=1}^k$. To sketch the proof, note that if we could write every $\bar{g}_i$ *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 $\{\bar{g}_i\}_{i=1}^k$. This is because both of $\{f_i\}_{i=1}^k$ and $\{\bar{g}_i\}_{i=1}^k$ are sets of linearly independent vectors of the same dimension and $\operatorname{span}\{\bar{g}_1, \ldots, \bar{g}_k\} \subseteq \operatorname{span}\{f_1, \ldots, f_k\}$. However, the $\bar{g}_i$'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 $\operatorname{span}\{\hat{f}_1, \ldots, \hat{f}_k\} = \operatorname{span}\{f_1, \ldots, f_k\}$, which implies Part 2.

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, S_j there exists an eigenvector having reasonably different values on the coordinates which correspond to S_i and S_j.

Lemma 3.2 *Let $\Upsilon = \Omega(k^3)$. For any $1 \leqslant i \leqslant k$, let $\hat{g}_i = \beta_1^{(i)} \bar{g}_1 + \cdots + \beta_k^{(i)} \bar{g}_k$ 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*

$$\left| \beta_\ell^{(i)} - \beta_j^{(i)} \right| \geqslant \zeta \triangleq \frac{1}{10\sqrt{k}}. \tag{3.3}$$

We point out that it was already shown in Kwok et al. (2013) 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 the entire cluster is concentrated around a certain point.

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

4. Analysis of Spectral k-Means Algorithms

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 $\mathcal{X} \subseteq \mathbb{R}^d$.

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 the k-means algorithm when applied to this embedding, and the proof sketch 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 $c_1, \cdots, c_k$ to minimize the sum of the squared-distance between $x \in \mathcal{X}$ and the center to which it is assigned. Formally, for any partition $\mathcal{X}_1, \cdots, \mathcal{X}_k$ 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 squared-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

$$\Delta_k^2(\mathcal{X}) \triangleq \min_{\text{partition } \mathcal{X}_1,\ldots,\mathcal{X}_k} \mathsf{COST}(\mathcal{X}_1, \ldots, \mathcal{X}_k). \tag{4.1}$$

A typical spectral k-means algorithm on graphs can be described as follows: (i) Compute the bottom k eigenvectors $f_1, \cdots, f_k$ of the normalized Laplacian matrix[4] of graph G. (ii) Map every vertex $u \in V[G]$ to a point $F(u) \in \mathbb{R}^k$ according to

$$F(u) = \frac{1}{\mathsf{NormalizationFactor}(u)} \cdot (f_1(u), \ldots, f_k(u))^{\mathsf{T}}, \tag{4.2}$$

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

4.2. Analysis of the Spectral Embedding

The first step of the k-means clustering described above is to map vertices of a graph into points in Euclidean space, through the spectral embedding (4.2). 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 (4.2) 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

4. 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.

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[5].

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

$$p^{(i)} \triangleq \frac{1}{\sqrt{\mathrm{vol}\,(S_i)}} \left(\beta_i^{(1)}, \ldots, \beta_i^{(k)}\right)^{\mathsf{T}}. \tag{4.3}$$

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 squared-distance between points in the ith cluster $\mathcal{X}_i$ 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* $\sum_{i=1}^{k} \sum_{u \in S_i} d_u \left\|F(u) - p^{(i)}\right\|^2 \leqslant 1.1k^2/\Upsilon$.

We will further show in Lemma 4.2 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 $\mathrm{vol}(S_i)$ are far from points in S_j of bigger $\mathrm{vol}(S_j)$. Notice that, if this was not the case, a small misclassification of points in a bigger cluster S_j could introduce a large error to the cluster of smaller volume.

Lemma 4.2 *For every* $i \neq j$, *it holds that* $\left\|p^{(i)} - p^{(j)}\right\|^2 \geqslant \frac{\zeta^2}{10 \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}$, *where* ζ *is defined in* (3.3).

4.3. Proof Sketch of Theorem 1.2

Now we analyze why spectral k-means algorithms perform 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

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

Lemma 4.3 *It holds that* $\Delta_k^2 \leqslant 1.1k^2/\Upsilon$.

5. Notice that this embedding is similar with the one used in Lee et al. (2012), 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.

Since $A_1, \cdots, A_k$ is the output of a k-means algorithm with approximation ratio APT, by Lemma 4.3 we have that $\mathsf{COST}(A_1, \ldots, A_k) \leqslant \mathsf{APT} \cdot 1.1k^2/\Upsilon$. We will show that this upper bound of $\mathsf{APT} \cdot 1.1k^2/\Upsilon$ 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 proven 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 to the fact that $\mathsf{COST}(A_1, \ldots, A_k) \leqslant \mathsf{APT} \cdot 1.1k^2/\Upsilon$. The core of of the whole contradiction arguments is the following technical lemma:

Lemma 4.4 *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 $\mathrm{vol}\left(A_i \triangle S_{\sigma(i)}\right) \geqslant 2\varepsilon \,\mathrm{vol}\left(S_{\sigma(i)}\right)$ for $\varepsilon \geqslant 10^5 \cdot k^3/\Upsilon$, then $\mathsf{COST}(A_1, \ldots, A_k) \geqslant 10^{-4} \cdot \varepsilon/k$.*

5. Partitioning Well-Clustered Graphs in Nearly-Linear Time

In this section we present a nearly-linear time algorithm for partitioning well-clustered graphs. 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. We show that, due to the well-separation properties of the embedded points, both steps in a k-means algorithm can be greatly simplified. Together with the heat kernel embedding, we obtain an ad-hoc version of k-means algorithms, which runs in nearly-linear time. Throughout the whole section, we assume that $\Upsilon = \Omega(k^5)$, where $\Omega(\cdot)$ hides a factor of $\log^c k$ for some constant c

5.1. The Seeding Step

The seeding step chooses k points from $\{F(u)\}_{u \in V[G]}$, such that with constant probability every point (i) belongs to a different cluster, and (ii) is close to the center of a cluster. To give an intuition of this seeding step, notice that our approximate center $p^{(i)}$ satisfies $\left\|p^{(i)}\right\|^2 \approx 1/\mathrm{vol}(S_i)$, and most embedded points $F(u)$ are close to their approximate centers, implying that $\|F(u)\|^2$ is approximately equal to $1/\mathrm{vol}(S_i)$ for most vertices $u \in S_i$. Hence, sampling points $F(u)$ with probability proportional to $d_u \cdot \|F(u)\|^2$ ensures that points from different clusters are sampled with approximately the same probability. We prove that, after sampling $\Theta(k \log k)$ points which constitute the sampling set C, with constant probability there is at least one point sampled from each cluster.

Next we remove the sampled points in C which are close to each other, until there are exactly k points left. We call this resulting set $C^\star$, and prove that with constant probability there is exactly one point in $C^\star$ from a cluster. See Algorithm 1 for formal description.

We remark that choosing good candidate centers is crucial for most k-means algorithms, and has been studied extensively in literature (e.g. Arthur and Vassilvitskii (2007); Ostrovsky et al. (2012)). While recent algorithms obtain good initial centers by iteratively picking points from a *non-uniform* distribution and take $\Omega(nk)$ time, Algorithm 1 runs in $\widetilde{O}(k)$ time.

Algorithm 1 SEEDANDTRIM(k, x)

input: the number of clusters k, and the embedding $\{x(u)\}_{u\in V[G]}$
Let $N = \Theta(k \log k)$.
for $i = 1, \ldots, N$ **do**
    Set $c_i = u$ with probability proportional to $d_u \|x(u)\|^2$.
end for
for $i = 2, \ldots, N$ **do**
    Delete all c_j with $j < i$ such that $\|x(c_i) - x(c_j)\|^2 < \frac{\|x(c_i)\|^2}{10^4 k}$.
end for
return $(c_1 \ldots c_k)$

5.2. The Grouping Step

After obtaining $C^\star$, we group the remaining vertices by assigning vertex u to cluster S_i if $F(u)$ is closer to $c_i \in C^\star$ than the other points in $C^\star$. A naive implementation of this step takes $\Omega(nk)$ time. To speed it up, we apply the ε-approximate nearest neighbor data structures (Indyk and Motwani, 1998) which gives us a nearly-linear time algorithm for the grouping step.

5.3. Dealing with Large k

Since computing the spectral embedding $F : V[G] \to \mathbb{R}^k$ takes $O(mk)$ time, SEEDANDTRIM(k, F) and the grouping step together run in nearly-liner time as long as $k = O(\mathrm{poly}\log n)$. However, for larger values of k this approach becomes problematic, as it is not clear how to obtain the embedding F in nearly-linear time. To handle this case, we notice that both steps only use the pairwise distances of the embedded points $F(u)$ and $F(v)$, rather than the spectral embedding itself. We show that these distances can be approximated by the so-called *heat kernel distance*, which can be approximately computed in nearly-linear time.

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

$$\mathbf{H}_t \triangleq \mathrm{e}^{-t\mathcal{L}} = \sum_{i=1}^{n} \mathrm{e}^{-t\lambda_i} f_i f_i^{\mathsf{T}}. \tag{5.1}$$

We view the heat kernel as a geometric embedding from $V[G]$ to $\mathbb{R}^n$, where the jth coordinate of the embedding is given by the jth bottom eigenvector of $\mathbf{H}_t$, i.e.

$$x_t(u) \triangleq \frac{1}{\sqrt{d_u}} \cdot \left(\mathrm{e}^{-t\cdot\lambda_1} f_1(u), \cdots, \mathrm{e}^{-t\cdot\lambda_n} f_n(u)\right). \tag{5.2}$$

We define the squared-distance between the points $x_t(u)$ and $x_t(v)$ by

$$\eta_t(u, v) \triangleq \|x_t(u) - x_t(v)\|^2. \tag{5.3}$$

We prove that, under the condition of $k = \Omega(\log n)$ and the gap assumption of Υ, there is a wide range of t for which $\eta_t(u, v)$ gives a good approximation of $\|F(u) - F(v)\|^2$.

Lemma 5.1 *Let $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$. Then, there is t such that with high probability the embedding $\{x_t(u)\}_{u\in V[G]}$ defined by (5.2) satisfies*

$$\frac{1}{2\mathrm{e}} \cdot \|F(u) - F(v)\|^2 \leqslant \|x_t(u) - x_t(v)\|^2 \leqslant \|F(u) - F(v)\|^2 + \frac{1}{n^c},$$
$$\frac{1}{2\mathrm{e}} \cdot \|F(u)\|^2 \leqslant \|x_t(u)\|^2 \leqslant \|F(u)\|^2 + \frac{1}{n^c}$$

for all u, v, where c is some constant. Moreover, these pairwise distances $\|x_t(u) - x_t(v)\|^2$ for all $\{u, v\} \in E(G)$ can be approximated up to a constant factor in $\widetilde{O}(m)$ time.

Now we use the doubling argument to find a required t: We run the seeding and the grouping steps using the heat kernel embedding for all t of the form $t = 2^i$, where $i = 1, 2, \ldots, O(\log n)$, and keep track of the best partition found so far. It is easy to see that the final partition after enumerating these t's gives a desired approximation. The overall description of our algorithm for large values of k is shown in Algorithm 2.

Algorithm 2 Clustering Algorithm

INPUT: (G, k)
for $i = 1, \ldots, k$ **do**
  $A'_i := \emptyset$
end for
$\mathsf{COST}(A_1, \ldots, A_k) := \infty$;
for $t = 2, 4, 8, \ldots, \mathrm{poly}(n)$ **do**
  $(c_1, \ldots, c_k) \leftarrow$ SEEDANDTRIM(k, x_t)
  Compute a partition $A_1, \ldots, A_k$ of V: for every $v \in V$ assign v to its nearest center c_i using the ε-NNS algorithm with $\varepsilon = \log k$.
  If $\mathsf{COST}(A_1, \ldots, A_k) \leqslant \mathsf{COST}(A'_1, \ldots, A'_k)$ SET$A'_i := A_i$ FOR $1 \leqslant i \leqslant k$
end for
return $(A'_1, \cdots, A'_k)$

Acknowledgments

Part of this work was done while He Sun and Luca Zanetti worked at Max Planck Institute for Informatics, and He Sun was visiting the Simons Institute for the Theory of Computing, UC Berkeley. We are grateful to Luca Trevisan for insightful comments on an early version of our paper, and to Gary Miller for very helpful discussions about heat kernels on graphs.

References

Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In *39th Annual ACM Symposium on Theory of Computing (STOC'07)*, pages 227–236, 2007.

Sanjeev Arora, Boaz Barak, and David Steurer. Subexponential algorithms for unique games and related problems. In *51st Annual IEEE Symposium on Foundations of Computer Science (FOCS'10)*, pages 563–572, 2010.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07)*, pages 1027–1035, 2007.

Fan R. K. Chung. A local graph partitioning algorithm using heat kernel pagerank. *Internet Mathematics*, 6(3):315–330, 2009.

Guy B. Coleman and Harry C. Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Tamal K. Dey, Alfred Rossi, and Anastasios Sidiropoulos. Spectral concentration, robust k-center, and simple clustering. *arXiv:1404.3918*, 2014.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *30th Annual ACM Symposium on Theory of Computing (STOC'98)*, pages 604–613, 1998.

Jonathan A. Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'14)*, pages 217–226, 2014.

Ioannis Koutis, Alex Levin, and Richard Peng. Improved Spectral Sparsification and Numerical Algorithms for SDD Matrices. In *29th International Symposium on Theoretical Aspects of Computer Science (STACS'12)*, pages 266–277, 2012.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+ ε)-approximation algorithm for geometric k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04)*, pages 454–462, 2004.

Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. Improved Cheeger's inequality: analysis of spectral partitioning algorithms through higher order spectral gap. In *45th Annual ACM Symposium on Theory of Computing (STOC'13)*, pages 11–20, 2013.

James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multi-way spectral partitioning and higher-order Cheeger inequalities. In *44th Annual ACM Symposium on Theory of Computing (STOC'12)*, pages 1117–1130, 2012.

Frank T. Leighton and Satish Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787–832, 1999.

Anand Louis, Prasad Raghavendra, Prasad Tetali, and Santosh Vempala. Many sparse cuts via higher eigenvalues. In *44th Annual ACM Symposium on Theory of Computing (STOC'12)*, pages 1131–1140, 2012.

David W. Matula and Farhad Shahrokhi. Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, 27(1-2):113–123, 1990.

Frank McSherry. Spectral partitioning of random graphs. In *42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS'01)*, pages 529–537, 2001.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.

Lorenzo Orecchia, Leonard J. Schulman, Umesh V. Vazirani, and Nisheeth K. Vishnoi. On partitioning graphs via single commodity flows. In *40th Annual ACM Symposium on Theory of Computing (STOC'08)*, pages 461–470, 2008.

Lorenzo Orecchia, Sushant Sachdeva, and Nisheeth K Vishnoi. Approximating the exponential, the Lanczos method and an $\widetilde{O}(m)$-time spectral algorithm for balanced separator. In *44th Annual ACM Symposium on Theory of Computing (STOC'12)*, pages 1141–1160, 2012.

Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the k-means problem. *Journal of the ACM*, 59(6):28, 2012.

Shayan Oveis Gharan and Luca Trevisan. Partitioning into expanders. In *25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'14)*, pages 1256–1266, 2014.

Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *27th Conference on Computational Complexity (CCC'12)*, pages 64–73, 2012.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Jonah Sherman. Breaking the multicommodity flow barrier for $O(\sqrt{\log n})$-approximations to sparsest cut. In *50th Annual IEEE Symposium on Foundations of Computer Science (FOCS'09)*, pages 363–372, 2009.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Luca Trevisan. Approximation algorithms for unique games. *Theory of Computing*, 4(1):111–128, 2008.

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

Van Vu. A simple SVD algorithm for finding hidden partitions. *arXiv:1404.3918*, 2014.

Zeyuan Allen Zhu, Silvio Lattanzi, and Vahab S. Mirrokni. A local algorithm for finding well-connected clusters. In *30th International Conference on Machine Learning (ICML'13)*, pages 396–404, 2013.

Appendix A. Omitted Proofs of Section 3

In this section we prove Theorem 3.1 and Lemma 3.2. We begin proving the first direction of Theorem 3.1.

Proof of Theorem 3.1 Part 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

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

where $\alpha' \triangleq \left(\alpha_2^{(i)}\right)^2 + \cdots + \left(\alpha_k^{(i)}\right)^2$. 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. ■

Next we prove Part 2 of Theorem 3.1. We first state two classical results that will be used in our proof.

Theorem A.1 (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*

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

Theorem A.2 (Corollary 6.3.4, Horn and Johnson (2012)) *Let* $\mathbf{A}$ *be an* $n \times n$ *normal 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 Theorem 3.1 Part 2 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 $\left\{\alpha_i^{(j)}\right\}_{i=1}^{k}$ representing $\hat{f}_j$. We define a vector $\alpha^{(j)}$ by

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

Notice that (i) the norm of every column of $\mathbf{A}$ is close to 1, and (ii) different columns are *almost* orthogonal to each other, since by Cauchy-Schwarz inequality it holds that

$$\left|\left\langle \alpha^{(i)}, \alpha^{(j)} \right\rangle\right| \leqslant \left\|\alpha^{(i)}\right\| \cdot \left\|\alpha^{(j)}\right\| \leqslant \max\left\{\mathcal{R}(\bar{g}_i)/\lambda_{k+1}, \mathcal{R}(\bar{g}_j)/\lambda_{k+1}\right\} \leqslant 1/\Upsilon, \qquad \text{for } i \neq j.$$

This implies that $\mathbf{A}$ is almost an orthogonal matrix. Moreover, it holds for any $i \neq j$ that

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

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 A.1), it holds that all the eigenvalues of $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ are at least

$$1 - 1/\Upsilon - (k-1)\cdot 1/\Upsilon = 1 - k/\Upsilon.$$

Therefore, $\mathbf{A}$ has no eigenvalue with value 0 as long as $\Upsilon > k$, i.e., the vectors $\left\{\alpha^{(j)}\right\}_{j=1}^{k}$ are linearly independent. Combining this with the fact that $\operatorname{span}\{\hat{f}_1, \ldots, \hat{f}_k\} \subseteq \operatorname{span}\{f_1, \ldots, f_k\}$ and $\dim(\operatorname{span}(\{f_1, \ldots, f_k\})) = k$, it holds that $\operatorname{span}\{\hat{f}_1, \ldots, \hat{f}_k\} = \operatorname{span}\{f_1, \ldots, f_k\}$. 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.,

$$f_i = \beta_1^{(i)}\hat{f}_1 + \beta_2^{(i)}\hat{f}_2 + \cdots + \beta_k^{(i)}\hat{f}_k. \tag{A.1}$$

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

$$\hat{g}_i = \beta_1^{(i)}\overline{g}_1 + \beta_2^{(i)}\overline{g}_2 + \cdots + \beta_k^{(i)}\overline{g}_k. \tag{A.2}$$

Without loss of generality we assume that $\left|\beta_j^{(i)}\right| = \max\left\{\left|\beta_1^{(i)}\right|, \ldots, \left|\beta_k^{(i)}\right|\right\}$. Then, it holds

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

which implies (for a large enough value of Υ) that

$$\left(\beta_j^{(i)}\right)^2 \leqslant \left(1-\frac{1}{\Upsilon}-\frac{k^2}{\Upsilon}\right)^{-1} \leqslant 1.1.$$

Combining this with Part 1 of Theorem 3.1, it is easy to see that

$$\|f_i-\hat{g}_i\|^2 \leqslant k \max_{1\leqslant j\leqslant k}\left(\beta_j^{(i)}\right)^2 \|\hat{f}_j-\overline{g}_j\|^2 \leqslant 1.1k/\Upsilon.$$

■

Now we prove Lemma 3.2, which shows that for any two different clusters S_i, S_j there exists an eigenvector having reasonably different values on the coordinates which correspond to S_i and S_j.

Proof of Lemma 3.2 Let $\beta^{(i)}=\left(\beta_1^{(i)},\ldots,\beta_k^{(i)}\right)^{\mathsf{T}}$, for $1\leqslant i\leqslant k$. Since $\bar{g}_i\perp\bar{g}_j$ for any $i\neq j$, we have by the orthonormality of $\overline{g}_1,\cdots,\overline{g}_k$ 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

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

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

$$\left\|\beta^{(i)}\right\|^2 \in \left(1-(2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon),1+2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon\right). \tag{A.3}$$

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 A.1), all eigenvalues λ of $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ satisfies

$$|\lambda-(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{i,i}| \leqslant (k-1)\cdot(2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon) \tag{A.4}$$

for any i. Combing this with (A.3), we have that $\mathbf{B}$' eigenvalues have modulus 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

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

Since $\mathbf{B}^\intercal\mathbf{B}$ is symmetric and 0 is an eigenvalue of $\mathbf{Q}^\intercal\mathbf{Q}$, by Theorem A.2 we know that, if $\hat{\lambda}$ is an eigenvalue of $\mathbf{Q}^\intercal\mathbf{Q}$, then there is an eigenvalue λ of $\mathbf{B}^\intercal\mathbf{B}$ such that

$$\begin{aligned}
|\hat{\lambda} - \lambda| &\leqslant \|\mathbf{B}^\intercal\mathbf{E} + \mathbf{E}^\intercal\mathbf{B} + \mathbf{E}^\intercal\mathbf{E}\| \\
&\leqslant \|\mathbf{B}^\intercal\mathbf{E}\| + \|\mathbf{E}^\intercal\mathbf{B}\| + \|\mathbf{E}^\intercal\mathbf{E}\| \\
&\leqslant 4\sqrt{k}\zeta' + k\zeta'^2,
\end{aligned}$$

which implies that

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

due to (A.3) and (A.4). By setting $\hat{\lambda} = 0$, we have that

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

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. ■

Appendix B. Omitted Proofs of Section 4

In this section we list all the proofs omitted from Section 4. We start proving Lemma 4.1, which shows that the embedded points of a cluster S_i are concentrated around the corresponding point $p^{(i)}$.

Proof of Lemma 4.1 Since $\|x\|^2 = \|\mathbf{D}^{-1/2}x\|_{\mathbf{D}}$ holds for any $x \in \mathbb{R}^n$, by Theorem 3.1 we have for any $1 \leqslant j \leqslant k$ that

$$\sum_{i=1}^{k}\sum_{u \in S_i} d_u \left(F(u)_j - p_j^{(i)}\right)^2 = \left\|\mathbf{D}^{-1/2}f_j - \mathbf{D}^{-1/2}\hat{g}_j\right\|_{\mathbf{D}} = \|f_j - \hat{g}_j\|^2 \leqslant 1.1k/\Upsilon.$$

Summing over all j for $1 \leqslant j \leqslant k$ implies that

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

■

The next lemma shows that the ℓ_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 B.1 *It holds for every* $1 \leqslant i \leqslant k$ *that*

$$\frac{9}{10\,\mathrm{vol}(S_i)} \leqslant \left\|p^{(i)}\right\|^2 \leqslant \frac{11}{10\,\mathrm{vol}(S_i)}.$$

Proof By (4.3), we have that

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

Notice that $p^{(i)}$ is just the ith row of the matrix $\mathbf{B}$ defined in the proof of Lemma 3.2, normalized by $\sqrt{\mathrm{vol}(S_i)}$. Since $\mathbf{B}$ and $\mathbf{B}^{\mathsf{T}}$ share the same singular values (this follows from the SVD decomposition), by (A.4) 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 squared norm of the ith row of $\mathbf{B}$, we have that

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

which implies the statement. ■

We can now prove Lemma 4.2 which shows that, for every $i \neq j$, $p^{(i)}$ is far from $p^{(j)}$. Moreover, their distance is inversely proportional to the volume of the smaller cluster between S_i and S_j.

Proof of Lemma 4.2 Let S_i and S_j be two arbitrary clusters. By Lemma 3.2, 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 (B.1), 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 $\left\|p^{(i)}\right\|^2 \geqslant \left\|p^{(j)}\right\|^2$. By Lemma B.1, 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

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

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

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

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

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

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

and the lemma follows. ■

The next lemma gives an upper bound to the cost of the optimal k-means clustering which depends on the gap Υ.

Proof of Lemma 4.3 Since Δ_k^2 is obtained by minimizing over all partitions $A_1, \ldots, A_k$ and $c_1, \ldots, c_k$, we have that

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

Hence the statement follows by applying Lemma 4.1. ■

By Lemma 4.3 and the assumption that $A_1, \cdots, A_k$ is an APT-factor approximation of an optimal clustering, we have that $\mathsf{COST}(A_1, \ldots, A_k) \leqslant 1.1 \cdot \mathsf{APT} \cdot k^2/\Upsilon$. Now we prove Theorem 1.2, which is based on the upper bound of COST and Lemma 4.4.

Proof of Theorem 1.2 Let $A_1, \ldots, A_k$ be a k-way partition that achieves an approximation ratio of APT. We first show that there exists a permutation σ of the indices such that

$$\mathrm{vol}\left(A_i \triangle S_{\sigma(i)}\right) \leqslant \frac{2 \cdot 10^5 \cdot k^3 \cdot \mathsf{APT}}{\Upsilon} \mathrm{vol}(S_{\sigma(i)})$$

for any $1 \leqslant i \leqslant k$. Assume for contradiction that there is $1 \leqslant i \leqslant k$ such that

$$\mathrm{vol}(A_i \triangle S_{\sigma(i)}) > \frac{2 \cdot 10^5 \cdot k^3 \cdot \mathsf{APT}}{\Upsilon} \mathrm{vol}\left(S_{\sigma(i)}\right).$$

This implies by Lemma 4.4 that

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

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

Next we bound the conductance of every cluster A_i. Let

$$\varepsilon = \frac{2 \cdot 10^5 \cdot k^3 \cdot \mathsf{APT}}{\Upsilon} = O\left(\frac{k^3 \cdot \mathsf{APT}}{\Upsilon}\right).$$

For any $1 \leqslant i \leqslant k$, the number of leaving edges of A_i is upper bounded by

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

On the other hand, we have that

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

Hence,

$$\phi_G(A_i) \leqslant \frac{(2\varepsilon + \phi_G(S_{\sigma(i)}))\,\mathrm{vol}(S_{\sigma(i)})}{(1 - 2\varepsilon)\,\mathrm{vol}(S_{\sigma(i)})} = \frac{2\varepsilon + \phi_G(S_{\sigma(i)})}{1 - 2\varepsilon} = O(\phi_G(S_{\sigma(i)}) + \mathsf{APT} \cdot k^3/\Upsilon).$$

■

The rest of this section is devoted to prove Lemma 4.4, which 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 B.2). However, we already proved in Lemma 4.2 that any two clusters are far from each other. This implies that the COST value of $A_1, \ldots, A_k$ must be high, which leads to a contradiction.

Lemma B.2 *Suppose for every permutation $\pi : \{1, \dots, k\} \to \{1, \dots, k\}$ there exists an index i such that*

$$\operatorname{vol}\left(A_i \triangle S_{\pi(i)}\right) \geqslant 2\varepsilon \operatorname{vol}\left(S_{\pi(i)}\right).$$

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

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

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

Proof Let $\sigma : \{1, \dots, k\} \to \{1, \dots, k\}$ be the function defined by

$$\sigma(i) = \operatorname*{argmax}_{1 \leqslant j \leqslant k} \frac{\operatorname{vol}(A_i \cap S_j)}{\operatorname{vol}(S_j)}.$$

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 $\operatorname{vol}(A_i \triangle S_{\sigma(i)}) \geqslant 2\varepsilon \operatorname{vol}(S_{\sigma(i)})$. Without loss of generality, we assume that $i = 1$. Notice that

$$\operatorname{vol}\left(A_1 \triangle S_{\sigma(1)}\right) = \sum_{j \neq 1} \operatorname{vol}\left(A_j \cap S_{\sigma(1)}\right) + \sum_{j \neq \sigma(1)} \operatorname{vol}\left(A_1 \cap S_j\right). \tag{B.3}$$

Hence, one of the summations on the right hand side of (B.3) is at least $\varepsilon \operatorname{vol}\left(S_{\sigma(1)}\right)$. Now the proof is based on the case distinction.

Case 1: Assume that $\sum_{j \neq 1} \operatorname{vol}\left(A_j \cap S_{\sigma(1)}\right) \geqslant \varepsilon \operatorname{vol}(S_{\sigma(1)})$. We define τ_j for $2 \leqslant j \leqslant k$ to be

$$\tau_j = \frac{\operatorname{vol}\left(A_j \cap S_{\sigma(1)}\right)}{\operatorname{vol}\left(S_{\sigma(1)}\right)}.$$

We have that

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

and by the definition of σ it holds that

$$\frac{\operatorname{vol}\left(A_j \cap S_{\sigma(1)}\right)}{\operatorname{vol}\left(S_{\sigma(1)}\right)} \geqslant \frac{\operatorname{vol}\left(A_j \cap S_{\sigma(j)}\right)}{\operatorname{vol}\left(S_{\sigma(j)}\right)} = \tau_j$$

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

$$\sum_{j \neq \sigma(1)} \operatorname{vol}\left(A_1 \cap S_j\right) \geqslant \varepsilon \operatorname{vol}(S_{\sigma(1)}). \tag{B.4}$$

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

$$\tau'_j = \frac{\operatorname{vol}(A_1 \cap S_j)}{\operatorname{vol}\left(S_{\sigma(1)}\right)}.$$

Then, (B.4) implies that

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

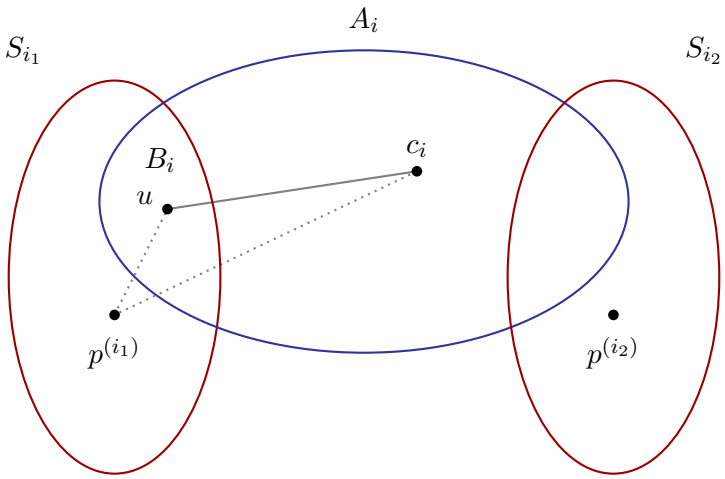

Figure 2: We use the fact that $\|p^{(i_1)} - c_i\| \geqslant \|p^{(i_2)} - c_i\|$, and lower bound the value of COST function by only looking at the contribution of points $u \in B_i$ for all $1 \leqslant i \leqslant k$.

The statement in this case holds by assuming $\mathrm{vol}\left(A_1 \cap S_{\sigma(1)}\right) \geqslant \varepsilon \,\mathrm{vol}\left(S_{\sigma(1)}\right)$, since otherwise we have

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

and this case was proven in Case 1.

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 $j \notin \{\sigma(1), \cdots, \sigma(k)\}$. 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

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

■

Proof of Lemma 4.4 By the assumption of Lemma 4.4 and Lemma B.2, for every i there exist $i_1 \neq i_2$ such that

$$\begin{aligned} \mathrm{vol}(A_i \cap S_{i_1}) &\geqslant \varepsilon_i \min\{\mathrm{vol}(S_{i_1}), \mathrm{vol}(S_{i_2})\}, \\ \mathrm{vol}(A_i \cap S_{i_2}) &\geqslant \varepsilon_i \min\{\mathrm{vol}(S_{i_1}), \mathrm{vol}(S_{i_2})\}, \end{aligned} \tag{B.5}$$

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 4.2). 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

$$\begin{aligned}
\mathsf{COST}(A_1, \ldots, A_k) &\geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \left(\frac{\left\|p^{(i_1)} - c_i\right\|^2}{2} - \left\|F(u) - p^{(i_1)}\right\|^2 \right) \\
&\geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \frac{\left\|p^{(i_1)} - c_i\right\|^2}{2} - \sum_{i=1}^{k} \sum_{u \in B_i} d_u \left\|F(u) - p^{(i_1)}\right\|^2 \\
&\geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \frac{\left\|p^{(i_1)} - c_i\right\|^2}{2} - \frac{1.1k^2}{\Upsilon} \qquad \text{(B.6)} \\
&\geqslant \sum_{i=1}^{k} \sum_{u \in B_i} d_u \frac{\left\|p^{(i_1)} - p^{(i_2)}\right\|^2}{8} - \frac{1.1k^2}{\Upsilon} \\
&\geqslant \sum_{i=1}^{k} \frac{\zeta^2 \operatorname{vol}(B_i)}{80 \min\left\{\operatorname{vol}(S_{i_1}), \operatorname{vol}(S_{i_2})\right\}} - \frac{1.1k^2}{\Upsilon} \qquad \text{(B.7)} \\
&\geqslant \sum_{i=1}^{k} \frac{\zeta^2 \varepsilon_i \min\left\{\operatorname{vol}(S_{i_1}), \operatorname{vol}(S_{i_2})\right\}}{80 \min\left\{\operatorname{vol}(S_{i_1}), \operatorname{vol}(S_{i_2})\right\}} - \frac{1.1k^2}{\Upsilon} \\
&\geqslant \sum_{i=1}^{k} \frac{\zeta^2 \varepsilon_i}{80} - \frac{1.1k^2}{\Upsilon} \\
&\geqslant \frac{\zeta^2 \varepsilon}{80} - \frac{1.1k^2}{\Upsilon} \geqslant \frac{\zeta^2 \varepsilon}{100}
\end{aligned}$$

where (B.6) follows from Lemma 4.1, (B.7) follows from Lemma 4.2 and the last inequality follows from the assumption that $\varepsilon \geqslant 10^5 \cdot k^3/\Upsilon$.

■

Appendix C. Omitted Proofs of Section 5

In this section we give a detailed discussion about our nearly-linear time algorithm for partitioning well-clustered graphs. We will start by analyzing the seeding step as described in Algorithm 1, and

give the correctness proof of the grouping step. 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:

$$\frac{1}{2\mathrm{e}}\cdot\|F(u)\|^2 \leqslant \|x(u)\|^2 \leqslant \|F(u)\|^2 + \frac{1}{n^5}, \tag{C.1}$$

$$\frac{1}{2\mathrm{e}}\cdot\|F(u)-F(v)\|^2 \leqslant \|x(u)-x(v)\|^2 \leqslant \|F(u)-F(v)\|^2 + \frac{1}{n^5} \tag{C.2}$$

Notice that these two conditions hold trivially if $\{x(u)\}_{u\in V[G]}$ is the spectral embedding, or any embedding produced by good approximations of the first k eigenvectors. However, obtaining such embedding becomes non-trivial when k becomes large, as directly computing the first k eigenvectors takes super-linear time. We will show that with proper choice of t the heat kernel embedding defined in (5.2) satisfies (C.1) and (C.2). This leads to a nearly-linear time algorithm for partitioning well-clustered graphs, even when the number of clusters k is super logarithmic.

Throughout the whole section, we assume that $\Omega(k^5)$, where the $\Omega(\cdot)$ notation here hides a factor of $\log^c k$ for some constant c.

C.1. Analysis of SEEDANDTRIM

In this subsection we analyse the correctness of Algorithm 1, which works as follows: We first sample $N \triangleq \Theta(k\log k)$ vertices, each with probability proportional to $d_u\|x(u)\|^2$. Next, we delete the sampled vertices that are close to each other until there are exactly k vertices left. We will show these k points are close to the actual centers of k clusters.

For any $1 \leqslant i \leqslant k$, we define $\mathcal{E}_i$ to be

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

i.e., $\mathcal{E}_i$ is the approximate contribution of vertices in S_i to the COST function. We define the radius of S_i by

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

for some parameter α, i.e., R_i^α is the approximate mean square error in cluster S_i. We define $\mathsf{CORE}_i^\alpha \subseteq S_i$ to be the set of vertices whose ℓ_2^2-distance to $p^{(i)}$ is at most R_i^α, i.e.,

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

By the averaging argument it holds that

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

and therefore $\mathrm{vol}(\mathsf{CORE}_i^\alpha) \geqslant \left(1-\frac{1}{\alpha}\right)\mathrm{vol}(S_i)$. From now on, we assume that $\alpha = \Theta(N\log N)$.

Lemma C.1 *The following statements hold:*

- $\sum_{u\in\mathsf{CORE}_i^\alpha} d_u\cdot\|F(u)\|^2 \geqslant 1-\frac{1}{100N}$.

- $\sum_{i=1}^{k}\sum_{u\notin \mathsf{CORE}_i^{\alpha}} d_u\cdot\|F(u)\|^2 \leqslant \frac{k}{100N}$.

Proof By the definition of CORE_i^{α}, we have that

$$\begin{aligned}
&\sum_{u\in\mathsf{CORE}_i^{\alpha}} d_u\cdot\|F(u)\|^2\\
&\geqslant \frac{1}{\alpha}\int_0^{\alpha}\sum_{u\in\mathsf{CORE}_i^{\rho}} d_u\cdot\|F(u)\|^2 d\rho\\
&\geqslant \frac{1}{\alpha}\int_0^{\alpha}\left(\left\|p^{(i)}\right\|-\sqrt{R_i^{\rho}}\right)^2\mathrm{vol}(\mathsf{CORE}_i^{\rho})d\rho &\text{(C.3)}\\
&\geqslant \frac{1}{\alpha}\int_0^{\alpha}\left(\left\|p^{(i)}\right\|^2-2\sqrt{R_i^{\rho}}\cdot\left\|p^{(i)}\right\|\right)\max\left\{\left(1-\frac{1}{\rho}\right)\mathrm{vol}(S_i),0\right\}d\rho &\text{(C.4)}\\
&\geqslant \frac{1}{\alpha}\int_0^{\alpha}\max\left\{\left(1-(2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon)-3\sqrt{\mathcal{E}_i\rho}\right)\left(1-\frac{1}{\rho}\right),0\right\}d\rho &\text{(C.5)}
\end{aligned}$$

where (C.3) follows from the fact that for all $u\in\mathsf{CORE}_i^{\rho}$, $\|F(u)\|\geqslant\|p^{(i)}\|-\sqrt{R_i^{\rho}}$, (C.4) from $\mathrm{vol}(\mathsf{CORE}_i^{\rho})\geqslant\max\left\{\left(1-\frac{1}{\rho}\right)\mathrm{vol}(S_i),0\right\}$, and (C.5) from the definition of R_i^{ρ} and the fact that

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

Therefore, we have that

$$\begin{aligned}
&\sum_{u\in\mathsf{CORE}_i^{\alpha}} d_u\cdot\|F(u)\|^2\\
&\geqslant \frac{1}{\alpha}\int_0^{\alpha}\max\left\{\left(1-(2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon)-4\sqrt{k^2\rho/\Upsilon}\right)\left(1-\frac{1}{\rho}\right),0\right\}d\rho\\
&\geqslant \frac{1}{\alpha}\int_0^{\alpha}\max\left\{1-(2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon)-4\sqrt{k^2\rho/\Upsilon}-\frac{1}{\rho},0\right\}d\rho\\
&\geqslant 1-(2.2\sqrt{k/\Upsilon}+1.1k/\Upsilon)-4k\sqrt{\alpha/\Upsilon}-\frac{\ln\alpha}{\alpha}\\
&\geqslant 1-\frac{1}{100N}, &\text{(C.6)}
\end{aligned}$$

where the last inequality holds by our assumptions on α and Υ.

The second statement follows by the fact that

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

and

$$\sum_{u\in V[G]} d_u\|F(u)\|^2=\sum_{u\in V[G]}\sum_{i=1}^{k} f_i^2(u)=k.$$

■

We next show that, after sampling $\Theta(k \log k)$ vertices, with constant probability the sampled vertices are from the cores of k clusters, and every core contains at least one sampled vertex.

Lemma C.2 *Assume that $N = \Omega(k \log k)$ vertices are sampled, in which every vertex is sampled with probability proportional to $d_u \cdot \|x(u)\|^2$. Then, with constant probability the set $C = \{c_1 \dots c_N\}$ of sampled vertices has the following properties:*

1. *Set C only contains vertices from the cores, i.e. $C \subseteq \bigcup_{i=1}^{k} \mathsf{CORE}_i^{\alpha}$, and*
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 (C.2), it holds for every vertex u that

$$\frac{1}{2\mathrm{e}} \cdot \|F(u)\|^2 \leqslant \|x(u)\|^2 \leqslant \|F(u)\|^2 + \frac{1}{n^5}.$$

Since

$$\sum_{u \in V[G]} d_u \|F(u)\|^2 = \sum_{u \in V[G]} \sum_{i=1}^{k} f_i^2(u) = k,$$

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

We first bound the probability that we sample at least one vertex from every core. For every $1 \leqslant i \leqslant k$, we have that the probability of each sample coming from CORE_i^{α} is at least

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

Therefore, the probability that we never encounter a vertex from CORE_i^{α} sampling N vertices is at most

$$\left(1 - \frac{1}{10k}\right)^N \leqslant \frac{1}{10k}.$$

Also, the probability that a sampled vertex is outside the cores of the clusters is at most

$$\begin{aligned}\frac{\sum_{u \notin \mathsf{CORE}_i^{\alpha}, \forall i} d_u \cdot \|x(u)\|^2}{k/6} &\leqslant \frac{\sum_{u \notin \mathsf{CORE}_i^{\alpha}, \forall i} d_u \cdot \left(\|F(u)\|^2 + n^{-5}\right)}{k/6} \\ &\leqslant \frac{\frac{k}{100N} + n^{-3}}{k/6} \leqslant \frac{1}{n^2} + \frac{6}{100N}.\end{aligned}$$

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

$$k \cdot \frac{1}{10k} + N \cdot \left(\frac{1}{n^2} + \frac{6}{100N}\right) \leqslant \frac{1}{2}.$$

■

We now show that points from the same core are much closer between each other than points from different cores. This allows us to show that the procedure SEEDANDTRIM succeeds with constant probability.

Lemma C.3 *For any two vertices $u, v \in \mathsf{CORE}_i^\alpha$, it holds that*

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

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

$$\|F(u) - F(v)\| \leqslant 2\sqrt{R_i^\alpha},$$

or

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

where the last inequality follows from the fact that $\sum_{i=1}^k \mathcal{E}_i \leqslant 1.1k^2/\Upsilon$. On the other hand, we also have

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

and

$$\|F(u)\|^2 \leqslant \left(\left\|p^{(i)}\right\| + \sqrt{R_i^\alpha}\right)^2 \leqslant \left(\left\|p^{(i)}\right\| + \sqrt{\frac{10\alpha k^2 \left\|p^{(i)}\right\|^2}{\Upsilon}}\right)^2 \leqslant \frac{11}{10 \operatorname{vol}(S_i)}.$$

Therefore by (C.2) it follows

$$\|x(u) - x(v)\|^2 \leqslant \|F(u) - F(v)\|^2 + \frac{1}{n^5} \leqslant \frac{5\alpha k^2}{\Upsilon \operatorname{vol}(S_i)} + \frac{1}{n^5} \leqslant \frac{10\alpha k^2}{\Upsilon}\|F(u)\|^2 \leqslant \frac{11\alpha k^2}{\Upsilon \operatorname{vol}(S_i)}.$$

By the conditions on α, Υ, and the fact that $\frac{1}{2\mathrm{e}} \cdot \|F(u)\|^2 \leqslant 2\|x(u)\|^2$, it also holds

$$\|x(u) - x(v)\|^2 \leqslant \frac{10\alpha k^2}{\Upsilon}\|F(u)\|^2 \leqslant \frac{\|x(u)\|^2}{2 \cdot 10^4 \cdot k}.$$

■

The next lemma shows the opposite: If two vertices belong to different cores, then they are far away from each other.

Lemma C.4 *For any $i \neq j$, and $u \in \mathsf{CORE}_i^\alpha, v \in \mathsf{CORE}_j^\alpha$, it holds that*

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

Proof 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 4.2, we have for any $i \neq j$,

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

Combing 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

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

Hence, we have that

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

■

Based on Lemma C.3 and Lemma C.4, we can simply delete one of the two vertices c_i and c_j whose distance is less than $10^{-4} \cdot \|x(c_i)\|^2/k$. The correctness and runtime of the procedure SEEDANDTRIM is summarized as follows:

Lemma C.5 *Given the embedding $\{x(u)\}$ satisfying (C.2) and (C.1), the procedure* SEEDANDTRIM *returns in $\widetilde{O}(k^2)$ time a set $C^\star$ of centers $c_1 \ldots c_k$ such that each CORE_i^α contains exactly one vertex in $C^\star$.*

C.2. Analysis of the Grouping Step

After the seeding step, we obtain k vertices $c_1, \cdots, c_k$, and with constant probability these k vertices belong to k different clusters. Now we assign the remaining $n - k$ vertices to different clusters. Based on the well-separation property, this step can be done by using the algorithm for solving the ε-approximate nearest neighbor problem (ε-NNS), which is formally described as follows:

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

Indyk and Motwani (1998) presents an algorithm for solving the ε-NNS problem, which uses $\widetilde{O}\left(|P|^{1+\frac{1}{1+\varepsilon}}+d|P|\right)$ preprocessing and requires $\widetilde{O}\left(d|P|^{\frac{1}{1+\varepsilon}}\right)$ query time, for a set of points $P\subset\mathbb{R}^d$, and $\varepsilon>0$. Our grouping step uses Indyk and Motwani's algorithm as a black box, where $P=\{x(c_1),\ldots,x(c_k)\}$, $\varepsilon=\Theta(\log k)$ and d is the dimension of the embedding $\{x(u)\}_{u\in V[G]}$. By using the standard dimensionality reduction techniques that approximately preserve pairwise distances, such as the Johnson-Lindenstrauss transform (see e.g. Dasgupta and Gupta (2003)), we can always assume $d=O(\log n)$. For this reason we can implement the grouping step in $\widetilde{O}(n)$ time.

C.3. Approximation Analysis of the Algorithm

Now we analyze the approximation ratio of the k-way partition computed by the seeding and grouping steps. The next lemma shows that the symmetric difference between the optimal partition and the ones returned by our algorithm can be bounded. This result is comparable with Theorem 1.2.

Lemma C.6 *Let $A_1,\ldots,A_k$ be a k-way partition of G returned by the seeding and grouping procedures. Then, under a proper permutation of the indices, for any $1\leqslant i\leqslant k$ it holds that (i) $\operatorname{vol}(A_i\triangle S_i)=O\left(k^3\log^2 k\cdot\Upsilon^{-1}\operatorname{vol}(S_i)\right)$, and (ii) $\phi_G(A_i)=O\left(\phi_G(S_i)+k^3\log^2 k\cdot\Upsilon^{-1}\right)$.*

Proof Let $c_1,\ldots,c_k$ be the approximate centers produced by the seeding and grouping steps. Then, for any $i\neq j$ it holds that

$$\|c_i-c_j\|^2\geqslant\left(\left\|p^{(i)}-p^{(j)}\right\|-\sqrt{R_i^\alpha}-\sqrt{R_j^\alpha}\right)^2\geqslant\frac{1}{2\cdot 10^4 k\min\{\operatorname{vol}(S_j),\operatorname{vol}(S_i)\}}\tag{C.7}$$

where the first inequality follows from the fact that c_i and $p^{(i)}$, as well as c_j and $p^{(j)}$ belong to the to the same cores respectively, while the second by Lemma 4.2 and the definition of R_i^α. Hence,

$$\begin{aligned}
\operatorname{vol}(S_i\setminus A_i)&\leqslant\sum_{i\neq j}\operatorname{vol}\left(\left\{v\in S_i:\|c_i-x(v)\|\geqslant\frac{\|c_j-x(v)\|}{\log k}\right\}\right)&\text{(C.8)}\\
&\leqslant\sum_{i\neq j}\operatorname{vol}\left(\left\{v\in S_i:\|c_i-x(v)\|\geqslant\frac{\|c_i-c_j\|-\|c_i-x(v)\|}{\log k}\right\}\right)\\
&\leqslant\sum_{i\neq j}\operatorname{vol}\left(\left\{v\in S_i:2\|c_i-x(v)\|\geqslant\frac{\|c_i-c_j\|}{\log k}\right\}\right)\\
&\leqslant\sum_{i\neq j}\operatorname{vol}\left(\left\{v\in S_i:2\left\|p^{(i)}-x(v)\right\|\geqslant\frac{\|c_i-c_j\|}{\log k}-2\left\|c_i-p^{(i)}\right\|\right\}\right)\\
&\leqslant\sum_{i\neq j}\operatorname{vol}\left(\left\{v\in S_i:\left\|p^{(i)}-x(v)\right\|^2\geqslant\frac{1}{10^5 k\log^2 k\operatorname{vol}(S_i)}\right\}\right)&\text{(C.9)}\\
&\leqslant\frac{2\cdot 10^5\cdot k^3\log^2 k}{\Upsilon}\operatorname{vol}(S_i),&\text{(C.10)}
\end{aligned}$$

where (C.8) follows by solving the ε-NNS problem with $\varepsilon=\log k-1$, (C.9) follows from (C.7) and Lemma C.4, and (C.10) from Lemma 4.1. Similarly, we also have that

$$\operatorname{vol}(A_i\setminus S_i)\leqslant\sum_{i\neq j}\operatorname{vol}\left(\left\{v\in S_j:\|c_j-x(v)\|\geqslant\frac{\|c_i-x(v)\|}{\log k}\right\}\right)\leqslant\frac{2\cdot 10^5\cdot k^3\log^2 k}{\Upsilon}\operatorname{vol}(S_i),$$

yielding the first statement of the lemma.

The second statement follows by the same argument in Theorem 1.2. ■

C.4. Dealing with Large k

In this subsection we will look at the case for which $k = \omega(\log n)$, and prove Lemma 5.1. As discussed in Section 5.3, when k is in this regime computing the spectral embedding becomes expensive, and we need to use another embedding instead. We will analyze the heat kernel embedding discussed in Section 5.3, and prove that the heat kernel embedding satisfies the assumptions (C.1) and (C.2). We will further show that such embedding can be approximately computed in nearly-nearly time. Our algorithm uses the algorithm for approximating the matrix exponential in Orecchia et al. (2012) as a subroutine, whose performance is listed below for completeness.

Theorem C.7 (Orecchia et al. (2012)) *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 $\|\mathrm{e}^{-\mathbf{A}}y - x\| \leqslant \delta\|y\|$ in time $\widetilde{O}((m_{\mathbf{A}}+n)\log(2+\|\mathbf{A}\|))$, where the $\widetilde{O}(\cdot)$ notation hides $\operatorname{poly}\log n$ and $\operatorname{poly}\log(1/\delta)$ factors.*

Proof of Lemma 5.1 Since

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

by assuming $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$ there is t such that $t \in (c\log n/\lambda_{k+1}, 1/(2\lambda_k))$ for a constant $c > 1$. We first show that the heat kernel embedding with this t satisfies the assumptions (C.1) and (C.2).

By the definition of the heat kernel distance in (5.3), we have that

$$\begin{aligned}\eta_t(u,v) &= \sum_{i=1}^{n} \mathrm{e}^{-2t\lambda_i}\left(\frac{f_i(u)}{\sqrt{d_u}} - \frac{f_i(v)}{\sqrt{d_v}}\right)^2 \\ &= \sum_{i=1}^{k} \mathrm{e}^{-2t\lambda_i}\left(\frac{f_i(u)}{\sqrt{d_u}} - \frac{f_i(v)}{\sqrt{d_v}}\right)^2 + \sum_{i=k+1}^{n} \mathrm{e}^{-2t\lambda_i}\left(\frac{f_i(u)}{\sqrt{d_u}} - \frac{f_i(v)}{\sqrt{d_v}}\right)^2. \end{aligned} \tag{C.11}$$

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

$$1 \geqslant \mathrm{e}^{-2t\lambda_i} \geqslant \mathrm{e}^{-\lambda_i/\lambda_k} \geqslant \frac{1}{\mathrm{e}}, \tag{C.12}$$

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

$$\mathrm{e}^{-2t\lambda_i} \leqslant \mathrm{e}^{-2c\log n\lambda_i/\lambda_{k+1}} \leqslant \mathrm{e}^{-2c\log n\lambda_{k+1}/\lambda_{k+1}} = \frac{1}{n^{2c}}. \tag{C.13}$$

By (C.12), the first summation in (C.11) is $[1/\mathrm{e}, 1]\cdot\|F(u) - F(v)\|^2$, and by (C.13) the second summation in (C.11) is at most n^{-2c+1}. The second assumption holds by noticing that $\|F(u)\|$ and $\|x_t(u)\|$ are the distances between $F(u)$, $x_t(u)$ to the origin respectively. Hence, the first statement holds.

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 (5.1) with (5.2) and (5.3), 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 C.7, 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 $d_u, d_v \geqslant 1$. This is equivalent to

$$\eta_t^{1/2}(u, v) - \delta \leqslant \|\mathbf{Z} (\xi_u - \xi_v)\| \leqslant \eta_t^{1/2}(u, v) + \delta. \tag{C.14}$$

We invoke the Johnson-Lindenstrauss transform in a way analogous to the computation of effective resistances from Spielman and Srivastava (2011) and Koutis et al. (2012). For an $O(\varepsilon^{-2} \cdot \log n) \times n$ Gaussian matrix $\mathbf{Q}$, with high probability it holds for all u, v that

$$(1 - \varepsilon) \|\mathbf{Z} (\xi_u - \xi_v)\| \leqslant \|\mathbf{QZ} (\xi_u - \xi_v)\| \leqslant (1 + \varepsilon) \|\mathbf{Z} (\xi_u - \xi_v)\| . \tag{C.15}$$

Combining (C.14) and (C.15) gives us that

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

Square both sides, and invoking the inequality

$$(1 - \varepsilon)\alpha^2 - (1 + \varepsilon^{-1})b^2 \leqslant (a + b)^2 \leqslant (1 + \varepsilon)\alpha^2 + (1 + \varepsilon^{-1})b^2,$$

then gives

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

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^{-c}. By setting ε to be a arbitrary small constant, the running time then follows from $\|\mathcal{L}\| \leqslant 2$ and the performance of the approximate exponential algorithm from (Orecchia et al., 2012) described in Theorem C.7. ■

The proof above provides an interesting observation about the heat kernel embedding: under the condition of $k = \Omega(\log n)$ and $\Upsilon = \Omega(k^3)$, we can always find a parameter t such that, when viewing $\|x_t(u) - x_t(v)\|^2$, the contribution to $\|x_t(u) - x_t(v)\|^2$ from the first k coordinates of $x_t(u)$ and $x_t(v)$ gives a $(1/\mathrm{e})$-approximation of $\|F(u) - F(v)\|^2$, while the contribution to $\|x_t(u) - x_t(v)\|^2$ from the remaining $n - k$ coordinates of $x_t(u)$ and $x_t(v)$ is $O(n^{-c})$, for a constant c. A similar intuition which views the heat kernel embedding as a weighted combination of multiple eigenvectors was discussed in Orecchia et al. (2012).

Appendix D. Generalization For Weighted Graphs

Our result can be easily generalized to undirected and weighted graphs for which the edge weights are polynomially bounded. Formally, for any weighted graph $G = (V, E, w)$ with the weight function $w : E \to \mathbb{R}$, we define the adjacency matrix $\mathbf{A}$ of G by

$$\mathbf{A}_{u,v} = \begin{cases} w(u,v) & \text{if } \{u,v\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

where $w(u,v) = w(v,u)$ is the weight on the edge $\{u,v\}$. For every vertex $u \in V$ we define the *weighted degree* of u as $d_u = \sum_{\{u,v\} \in E} w(u,v)$, and the degree matrix $\mathbf{D}$ is defined by $\mathbf{D}_{u,u} = d_u$. We can define the Laplacian matrix $\mathcal{L}$ and the heat kernel $\mathbf{H}_t$ in the same way as in the case of unweighted graphs. Then, it is easy to verify that all the results in Section 5 hold.