A Tighter Analysis of Spectral Clustering, and Beyond

A Tighter Analysis of Spectral Clustering, and Beyond

Peter Macgregor [1] **He Sun** [1]

Abstract

This work studies the classical spectral clustering algorithm which embeds the vertices of some graph $G = (V_G, E_G)$ into $\mathbb{R}^k$ using k eigenvectors of some matrix of G, and applies k-means to partition V_G into k clusters. Our first result is a tighter analysis on the performance of spectral clustering, and explains why it works under some much weaker condition than the ones studied in the literature. For the second result, we show that, by applying fewer than k eigenvectors to construct the embedding, spectral clustering is able to produce better output for many practical instances; this result is the first of its kind in spectral clustering. Besides its conceptual and theoretical significance, the practical impact of our work is demonstrated by the empirical analysis on both synthetic and real-world datasets, in which spectral clustering produces comparable or better results with fewer than k eigenvectors.

1. Introduction

Graph clustering is a fundamental problem in unsupervised learning, and has comprehensive applications in computer science and related scientific fields. Among various techniques to solve graph clustering problems, spectral clustering is probably the easiest one to implement, and has been widely applied in practice. Spectral clustering can be easily described as follows: for any graph $G = (V_G, E_G)$ and some $k \in \mathbb{Z}^+$ as input, spectral clustering embeds the vertices of V_G into $\mathbb{R}^k$ based on the bottom k eigenvectors of the Laplacian matrix of G, and employs k-means on the embedded points to partition V_G into k clusters. Thanks to its simplicity and excellent performance in practice, spectral clustering has been widely applied over the past three decades (Spielman & Teng, 1996).

[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom. Correspondence to: Peter Macgregor <peter.macgregor@ed.ac.uk>, He Sun <h.sun@ed.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

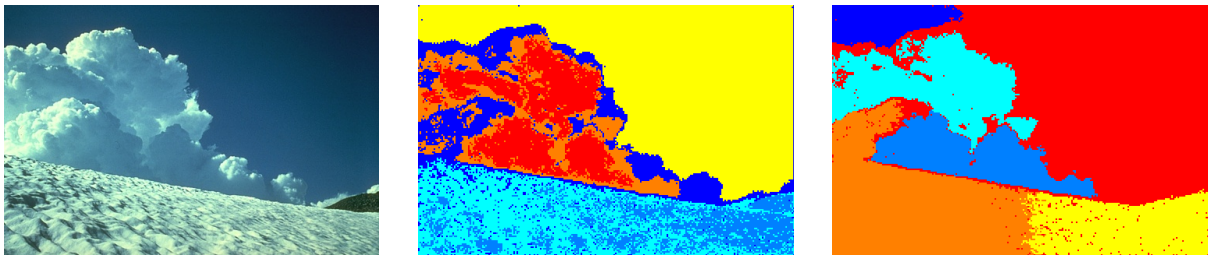
In this work we study spectral clustering, and present two results. Our first result is a tighter analysis of spectral clustering for well-clustered graphs. Informally, we analyse the performance guarantee of spectral clustering under a simple assumption[1] on the input graph. While all the previous work (e.g., (Lee et al., 2014; Kolev & Mehlhorn, 2016; Mizutani, 2021; Peng et al., 2017)) on the same problem suggests that the assumption on the input graph must depend on k, our result demonstrates that the performance of spectral clustering can be rigorously analysed under a general condition independent of k. To the best of our knowledge, our work presents the first result of its kind, and hence we believe that this result and the novel analysis used in its proof are important, and might have further applications in graph clustering.

Secondly, we study the clustering problem in which the crossing edges between the optimal clusters $\{S_i\}_{i=1}^k$ present some noticeable pattern, which we call the *meta-graph* in this work. Notice that, when viewing every cluster S_i as a "giant vertex", our meta-graph captures the intrinsic connection between the optimal clusters, and could be significantly different from a clique graph. We prove that, when this is the case, one can simply apply classical spectral clustering while employing fewer than k eigenvectors to construct the embedding and, surprisingly, this will produce a better clustering result. The significance of this result is further demonstrated by our extensive experimental analysis on the well-known BSDS, MNIST, and USPS datasets (Arbelaez et al., 2011; Hull, 1994; LeCun et al., 1998). While we discuss the experimental details in Section 6, the performance of our algorithm is showcased in Figure 1: in order to find 6 and 45 clusters, spectral clustering with 3 and 7 eigenvectors produce better results than the ones with 6 and 45 eigenvectors according to the default metric of the BSDS dataset.

1.1. Related work

Our first result on the analysis of spectral clustering is tightly related to a number of research that analyses spectral clustering algorithms under various conditions (e.g., (Gharan & Trevisan, 2014; Lee et al., 2014; Kolev & Mehlhorn, 2016; Mizutani, 2021; Ng et al., 2001; Peng et al., 2017)). While

[1]This assumption will be formally defined in Section 3.



(a) Original Image (b) 6 clusters found with 3 vectors (c) 6 clusters found with 6 vectors

(d) Original Image (e) 45 clusters found with 7 vectors (f) 45 clusters found with 45 vectors

Figure 1. Examples of image segmentation using spectral clustering; the original images are from the BSDS. The Rand Index of segmentation (b) is 0.83, while (c) has Rand Index 0.78. Segmentation (e) has Rand Index 0.92, and (f) has Rand Index 0.80. Hence, it's clear that spectral clustering with fewer than k eigenvectors suffices to produce comparable or better output.

we compare in detail between these works and ours in later sections, to the best of our knowledge, our work presents the first result proving spectral clustering works under some general condition independent of n and k. Our work is also related to studies on designing local, and distributed clustering algorithms based on different assumptions (e.g., (Czumaj et al., 2015; Orecchia & Zhu, 2014; Zhu et al., 2013)); due to limited computational resources available, these works require stronger assumptions on input graphs than ours.

Our second result on spectral clustering with fewer eigenvalues is linked to efficient spectral algorithms to find cluster-structures. While it's known that flow and path structures of clusters in digraphs can be uncovered with complex-valued Hermitian matrices (Cucuringu et al., 2020; Laenen & Sun, 2020), our work shows that one can apply real-valued Laplacians of undirected graphs, and find more general patterns of clusters characterised by our structure theorem. Finally, we notice that spectral clustering with more than k eigenvectors is studied in Rebagliati and Verri (2011), although their assumptions on an input graph are different from ours and their result is not directly comparable with ours.

2. Preliminaries

Let $G = (V_G, E_G, w)$ be an undirected graph with n vertices, m edges, and weight function $w : V_G \times V_G \to \mathbb{R}_{\geq 0}$. For any edge $e = \{u, v\} \in E_G$, we write the weight of $\{u, v\}$ by w_{uv} or w_e. For a vertex $u \in V_G$, we denote its *degree* by $d_G(u) \triangleq \sum_{v \in V} w_{uv}$. For any two sets $S, T \subset V_G$, we define the *cut value* $w(S, T) \triangleq \sum_{e \in E_G(S,T)} w_e$, where $E_G(S, T)$ is the set of edges between S and T. For any set $S \subseteq V_G$, the *volume* of S is $\mathrm{vol}_G(S) \triangleq \sum_{u \in S} d_G(u)$, and we write $\mathrm{vol}(G)$ when referring to $\mathrm{vol}(V_G)$. For any nonempty subset $S \subseteq V_G$, we define the *conductance* of S by

$$\Phi_G(S) \triangleq \frac{w(S, V \setminus S)}{\mathrm{vol}_G(S)}.$$

Furthermore, we define the conductance of the graph G by

$$\Phi_G \triangleq \min_{\substack{S \subset V \\ \mathrm{vol}(S) \leq \mathrm{vol}(V)/2}} \Phi_G(S).$$

We call subsets of vertices $A_1, \ldots, A_k$ a *k-way partition* of G if $A_i \cap A_j = \emptyset$ for different i and j, and $\bigcup_{i=1}^{k} A_i = V$. Generalising the definition of conductance, we define *k-way expansion constant* by

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

Next we define the matrices of any $G = (V_G, E_G, w)$. Let $\mathbf{D}_G \in \mathbb{R}^{n \times n}$ be the diagonal matrix defined by $(\mathbf{D}_G)_{uu} = d_G(u)$ for all $u \in V_G$, and we denote by $\mathbf{A}_G \in \mathbb{R}^{n \times n}$ the *adjacency matrix* of G, where $(\mathbf{A}_G)_{uv} = w_{uv}$ for all $u, v \in V_G$. The *normalised Laplacian matrix* of G is defined by $\mathcal{L}_G \triangleq \mathbf{I} - \mathbf{D}_G^{-1/2} \mathbf{A}_G \mathbf{D}_G^{-1/2}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. Since $\mathcal{L}_G$ is symmetric and real-valued, it has n real eigenvalues denoted by $\lambda_1 \leq \ldots \leq \lambda_n$; we use $f_i \in \mathbb{R}^n$

to denote the eigenvectors corresponding to λ_i for any $1 \leq i \leq n$. It is known that $\lambda_1 = 0$ and $\lambda_n \leq 2$ (Chung, 1997).

For any sets S and T, the symmetric difference between S and T is defined by $S \triangle T = (S \setminus T) \cup (T \setminus S)$. For any $k \in \mathbb{Z}^+$, we define $[k] \triangleq \{1, \ldots, k\}$. We sometimes drop the subscript G when it is clear from the context. The following higher-order Cheeger inequality will be used in our analysis.

Lemma 2.1 ((Lee et al., 2014)). *It holds for any $k \in [n]$ that $\lambda_k/2 \leq \rho(k) \leq O\left(k^3\right)\sqrt{\lambda_k}$.*

3. Encoding the Cluster-Structure into the Eigenvectors of $\mathcal{L}$

Let $\{S_i\}_{i=1}^k$ be any optimal k-way partition that achieves $\rho(k)$. We define the indicator vector of cluster S_i by

$$\chi_i(u) \triangleq \begin{cases} 1 & \text{if } u \in S_i, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

and the corresponding *normalised indicator vector* by

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

One of the basic results in spectral graph theory states that G consists of at least k connected components if and only if $\lambda_i = 0$ for any $i \in [k]$, and $\operatorname{span}\left(\{f_i\}_{i=1}^k\right) = \operatorname{span}\left(\{\bar{g}_i\}_{i=1}^k\right)$ (Chung, 1997). Hence, one would expect that, when G consists of k densely connected components (clusters) connected by sparse cuts, the bottom eigenvectors $\{f_i\}_{i=1}^k$ of $\mathcal{L}$ are close to $\{\bar{g}_i\}_{i=1}^k$. This intuition explains the practical success of spectral methods for graph clustering, and forms the basis of many theoretical studies on various spectral clustering algorithms (e.g. (Kwok et al., 2013; Lee et al., 2014; Ng et al., 2001; von Luxburg, 2007)).

Turning this intuition into a mathematical statement, Peng et al. (2017) studies the quantitative relationship between $\{f_i\}_{i=1}^k$ and $\{\bar{g}_i\}_{i=1}^k$ through the function $\Upsilon(k)$ defined by

$$\Upsilon(k) \triangleq \frac{\lambda_{k+1}}{\rho(k)}. \tag{2}$$

To explain the meaning of $\Upsilon(k)$, we assume that G has k well-defined clusters $\{S_i\}_{i=1}^k$. By definition, the values of $\Phi(S_i)$ for every S_i, as well as $\rho(k)$, are low; on the other hand, any $(k+1)$-way partition of V_G would separate the vertices of some S_i, and as such $\rho(k+1)$'s value will be high. Combining this with the higher-order Cheeger inequality, some lower bound on $\Upsilon(k)$ would be sufficient to ensure that G has exactly k clusters. In their work, Peng et al. (2017) assumes $\Upsilon(k) = \Omega(k^2)$, and proves that the space spanned by $\{f_i\}_{i=1}^k$ and the one spanned by $\{\bar{g}_i\}_{i=1}^k$ are close to each other. Specifically, they show that

1. every $\bar{g}_i$ is close to some linear combination of $\{f_i\}_{i=1}^k$, denoted by $\widehat{f}_i$, i.e., it holds that $\|\bar{g}_i - \widehat{f}_i\|^2 \leq 1/\Upsilon(k)$;
2. every f_i is close to some linear combination of $\{\bar{g}_i\}_{i=1}^k$, denoted by $\widehat{g}_i$, i.e., it holds that $\|f_i - \widehat{g}_i\|^2 \leq 1.1k/\Upsilon(k)$.

In essence, their so-called structure theorem gives a quantitative explanation on why spectral methods work for graph clustering when there is a clear cluster-structure in G characterised by $\Upsilon(k)$. As it holds for graphs with clusters of different sizes and edge densities, this structure theorem has been shown to be a powerful tool in analysing clustering algorithms, and inspired many subsequent works (e.g., (Chen et al., 2016; Czumaj et al., 2015; Kloumann et al., 2017; Kolev & Mehlhorn, 2016; Louis & Venkat, 2019; Mizutani, 2021; Peng, 2020; Peng & Yoshida, 2020; Sun & Zanetti, 2019)).

In this section we show that a stronger statement of the original structure theorem holds under a much weaker assumption. Our result is summarised as follows:

Theorem 1 (The Stronger Structure Theorem). *The following statements hold:*

1. *For any $i \in [k]$, there is $\widehat{f}_i \in \mathbb{R}^n$, which is a linear combination of $f_1, \ldots, f_k$, such that $\|\bar{g}_i - \widehat{f}_i\|^2 \leq 1/\Upsilon(k)$.*
2. *There are vectors $\widehat{g}_1, \ldots, \widehat{g}_k$, each of which is a linear combination of $\bar{g}_1, \ldots, \bar{g}_k$, such that $\sum_{i=1}^k \|f_i - \widehat{g}_i\|^2 \leq k/\Upsilon(k)$.*

To examine the significance of Theorem 1, we first highlight that these two statements hold for any $\Upsilon(k)$, while the original structure theorem relies on the assumption that $\Upsilon(k) = \Omega(k^2)$. Since $\Upsilon(k) = \Omega(k^2)$ is a strong and even questionable assumption when k is large, e.g., $k = \Omega(\operatorname{poly}\log(n))$, obtaining these statements for general $\Upsilon(k)$ is important. Secondly, our second statement of Theorem 1 significantly improves the original theorem. Specifically, instead of stating $\|f_i - \widehat{g}_i\|^2 \leq 1.1k/\Upsilon(k)$ for any $i \in [k]$, our second statement shows that $\sum_{i=1}^k \|f_i - \widehat{g}_i\|^2 \leq k/\Upsilon(k)$; hence, it holds in expectation that $\|f_i - \widehat{g}_i\|^2 \leq 1/\Upsilon(k)$, the upper bound of which matches the first statement. This implies that the vectors $f_1, \ldots, f_k$ and $\bar{g}_1, \ldots, \bar{g}_k$ can be linearly approximated by each other with *roughly the same* approximation guarantee. Thirdly, rather than employing the machinery from matrix analysis used by Peng et al. (2017), to prove the original theorem, our proof is simple and purely linear-algebraic. Therefore, we believe that both of our stronger statements and much simplified proof are significant, and could have further applications in graph clustering and related problems.

4. Tighter Analysis of Spectral Clustering

In this section, we analyse the spectral clustering algorithm. For any input graph $G = (V_G, E_G)$ and $k \in [n]$, spectral clustering consists of the three steps below:

1. compute the eigenvectors $f_1, \dots f_k$ of $\mathcal{L}_G$, and embed each $u \in V_G$ to the point $F(u) \in \mathbb{R}^k$ according to
$$F(u) \triangleq \frac{1}{\sqrt{d(u)}} \left(f_1(u), \dots, f_k(u)\right)^\intercal; \qquad (3)$$
2. apply k-means on the embedded points $\{F(u)\}_{u \in V_G}$;
3. partition V_G into k clusters based on the output of k-means.

We will consider spectral clustering for graphs with clusters of *almost-balanced* size defined as follows.

Definition 1. *Let G be a graph with k clusters $\{S_i\}_{i=1}^k$. We say that the clusters are almost-balanced if*
$$(1/2) \cdot \mathrm{vol}(V_G)/k \leq \mathrm{vol}(S_i) \leq 2 \cdot \mathrm{vol}(V_G)/k$$
for all $i \in \{1, \dots, k\}$.

Our main result is summarised in Theorem 2, where we take APT to be the approximation ratio of the k-means algorithm used in spectral clustering. Recall that one can take APT to be some constant (Kumar et al., 2004).

Theorem 2. *Let G be a graph with k clusters $\{S_i\}_{i=1}^k$ of almost-balanced size, and $\Upsilon(k) \geq 2176 \cdot (1 + \mathsf{APT})$. Let $\{A_i\}_{i=1}^k$ be the output of spectral clustering and, without loss of generality, the optimal correspondent of A_i is S_i. Then, it holds that*
$$\sum_{i=1}^{k} \mathrm{vol}\left(A_i \triangle S_i\right) \leq 2176 \cdot (1 + \mathsf{APT}) \cdot \frac{\mathrm{vol}(G)}{\Upsilon(k)}.$$

Notice that some condition on $\Upsilon(k)$ is needed to ensure that an input graph G has k well-defined clusters, so that misclassified vertices can be formally defined. Taking this into account, the most significant feature of Theorem 2 is its upper bound of misclassified vertices with respect to $\Upsilon(k)$: our result holds, and is non-trivial, as long as $\Upsilon(k)$ is lower bounded by some constant[2]. This significantly improves most of the previous results of graph clustering algorithms, which make stronger assumptions on the input graphs. For example, Peng et al. (2017) assumes that $\Upsilon(k) = \Omega(k^3)$, Mizutani (Mizutani, 2021) assumes that $\Upsilon(k) = \Omega(k)$, the algorithm presented in Gharan and Trevisan (2014) assumes that $\lambda_{k+1} = \Omega\left(\mathrm{poly}(k)\lambda_k^{1/4}\right)$, and the one presented in Dey et al. (2019) further assumes some condition with respect to k, λ_k, and the maximum degree of G. While these assumptions require at least a linear dependency on k, making it difficult for the instances with a large value of k to satisfy, our result suggests that the performance of spectral clustering can be rigorously analysed for these graphs. In particular, compared with previous work, our result better justifies the widely used eigen-gap heuristic for spectral clustering (Ng et al., 2001; von Luxburg, 2007). This heuristic suggests that spectral clustering works when the value of $|\lambda_{k+1} - \lambda_k|$ is much larger than $|\lambda_k - \lambda_{k-1}|$, and in practice, the ratio between the two gaps is usually a constant rather than some function of k.

[2]Note that we can take any constant approximation in Definition 1 with a different corresponding constant in Theorem 2.

4.1. Properties of Spectral Embedding

Now we study the properties of the spectral embedding defined in (3), and show in the next subsection how to use these properties to prove Theorem 2. Due to the page limit we refer the reader to the appendix for all the technical details used in our analysis. For every cluster S_i, we define the vector $p^{(i)} \in \mathbb{R}^k$ by
$$p^{(i)}(j) = \frac{1}{\sqrt{\mathrm{vol}(S_i)}} \langle f_j, \bar{g}_i \rangle,$$
and view these $\{p^{(i)}\}_{i=1}^k$ as the approximate centres of the embedded points from the optimal clusters $\{S_i\}_{i=1}^k$. We prove that the total k-means cost of the embedded points can be upper bounded as follows:

Lemma 4.1. *It holds that*
$$\sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| F(u) - p^{(i)} \right\|^2 \leq \frac{k}{\Upsilon(k)}.$$

The importance of Lemma 4.1 is that, although the optimal centres for k-means are unknown, the existence of $\{p^{(i)}\}_{i=1}^k$ is sufficient to show that the cost of an optimal k-means clustering on $\{F(u)\}_{u \in V_G}$ is at most $k/\Upsilon(k)$. Since one can always use an $O(1)$-approximate k-means algorithm for spectral clustering (e.g., (Kanungo et al., 2004; Kumar et al., 2004)), the output of k-means on $\{F(u)\}_{u \in V_G}$ is $O\left(k/\Upsilon(k)\right)$.

In addition, we prove that any pair of different $p^{(i)}$ and $p^{(j)}$ are far away from each other. Moreover, their distance is essentially independent of k and $\Upsilon(k)$, as long as $\Upsilon(k) \geq 20$. This result is as follows:

Lemma 4.2. *It holds for any $i, j \in [k]$ with $i \neq j$ that*
$$\left\| p^{(i)} - p^{(j)} \right\|^2 \geq \frac{1}{\min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}} \left(\frac{1}{2} - \frac{8}{\Upsilon(k)}\right).$$

We remark that, despite the similarity in their formulation, most technical lemmas presented in this subsection and the

appendix are stronger than the ones in Peng et al. (2017). These results are obtained through our stronger structure theorem (Theorem 1), and are crucial for us to prove Theorem 2.

4.2. Proof Sketch of Theorem 2

Now we sketch our proof of Theorem 2. While the technical details can be found in the appendix, our focus here is to give a high-level overview of our proof technique, and explain why a mild condition like $\Upsilon(k) \geq 2174(1 + \mathsf{APT})$ suffices for spectral clustering to perform well in practice.

Let $\{A_i\}_{i=1}^k$ be the output of spectral clustering, and we denote the centre of the embedded points $\{F(u)\}$ for any A_i by c_i. As the starting point of our analysis, we claim that every c_i will be close to its "optimal" correspondent $p^{(\sigma(i))}$ for some $\sigma(i) \in [k]$. That is, the actual centre of embedded points from every A_i is close to the approximate centre of the embedded points from some optimal S_i. To formalise this, we define the function $\sigma : [k] \to [k]$ by

$$\sigma(i) = \arg\min_{j \in [k]} \left\| p^{(j)} - c_i \right\|; \tag{4}$$

that is, cluster A_i should correspond to $S_{\sigma(i)}$ in which the value of $\|p^{(\sigma(i))} - c_i\|$ is the lowest among all the distances between c_i and all of the $p^{(j)}$ for $j \in [k]$. However, one needs to be cautious as (4) wouldn't necessarily define a permutation, and there might exist different $i, i' \in [k]$ such that both of A_i and $A_{i'}$ map to the same $S_{\sigma(i)}$. Taking this into account, for any fixed $\sigma : [k] \to [k]$ and $i \in [k]$, we further define $M_{\sigma,i}$ by

$$M_{\sigma,i} \triangleq \bigcup_{j:\sigma(j)=i} A_j. \tag{5}$$

The following lemma shows that, when mapping every output A_i to $S_{\sigma(i)}$, the total ratio of misclassified volume with respect to each cluster can be upper bounded:

Lemma 4.3. *Let $\{A_i\}_{i=1}^k$ be the output of spectral clustering, and σ and $M_{\sigma,i}$ be defined as in (4) and (5). Assuming $\Upsilon(k) \geq 32$, it holds that*

$$\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 64 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}.$$

It remains to study the case in which σ isn't a permutation. Notice that, if this occurs, there is some $i \in [k]$ such that $M_{\sigma,i} = \emptyset$, and different values of $x, y \in [k]$ such that $\sigma(x) = \sigma(y) = j$ for some $j \neq i$. Based on this, we construct the function $\sigma' : [k] \to [k]$ from σ based on the following procedure:

- Set $\sigma'(z) = i$ if $z = x$;
- Set $\sigma'(z) = \sigma(z)$ for any other $z \in [k] \setminus \{x\}$.

Notice that one can construct σ' in this way as long as σ isn't a permutation, and this constructed σ' reduces the number of $M_{\sigma,i}$ being $\emptyset$ by one. We show one only needs to construct such σ' at most $O(k/\Upsilon(k))$ times to obtain the final permutation called $\sigma^\star$, and it holds for $\sigma^\star$ that

$$\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma^\star,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 1088 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}.$$

Combining this with the fact that the target clusters are balanced proves Theorem 2.

We remark that this method of upper bounding the ratio of misclassified vertices is very different from the ones used in previous references, e.g., (Dey et al., 2019; Mizutani, 2021; Peng et al., 2017). In particular, instead of examining all the possible mappings between $\{A_i\}_{i=1}^k$ and $\{S_i\}_{i=1}^k$, we directly work with some specifically defined function σ, and construct our desired mapping $\sigma^\star$ from σ. This is another key for us to obtain stronger results than the previous work.

5. Beyond the Classical Spectral Clustering

In this section we propose a variant of spectral clustering which employs fewer than k eigenvectors to find k clusters. We prove that, when the structure among the optimal clusters in an input graph satisfies certain conditions, spectral clustering with fewer eigenvectors is able to produce better results than classical spectral clustering. Our result gives a theoretical justification of the surprising showcase in Section 1, and presents a significant speedup on the runtime of spectral clustering in practice, since fewer eigenvectors are used to construct the embedding.

5.1. Encoding the Cluster-Structure into Meta-Graphs

Suppose that $\{S_i\}_{i=1}^k$ is a k-way partition of V_G for an input graph G that minimises the k-way expansion $\rho(k)$. We define the matrix $\mathbf{A}_M \in \mathbb{R}^{k \times k}$ by

$$\mathbf{A}_M(i,j) = \begin{cases} w(S_i, S_j) & \text{if } i \neq j, \\ 2w(S_i, S_j) & \text{if } i = j \end{cases}$$

and, taking $\mathbf{A}_M$ to be the adjacency matrix, this defines a graph $M = (V_M, E_M, w_M)$ which we refer to as the *meta-graph* of the clusters. We define the normalised adjacency matrix of M by $\mathcal{A}_M \triangleq \mathbf{D}_M^{-1/2} A_M \mathbf{D}_M^{-1/2}$, and the normalised Laplacian matrix of M by $\mathcal{L}_M \triangleq \mathbf{I} - \mathcal{A}_M$. Let the eigenvalues of $\mathcal{L}_M$ be $\gamma_1 \leq \ldots \leq \gamma_k$, and $g_i \in \mathbb{R}^k$ be the eigenvector corresponding to γ_i for any $i \in [k]$.

The starting point of our novel approach is to look at the structure of the meta-graph M defined by $\{S_i\}_{i=1}^k$ of G, and study how the spectral information of $\mathcal{L}_M \in \mathbb{R}^{k \times k}$ is

encoded in the bottom eigenvectors of $\mathcal{L}_G$. To achieve this, for any $\ell \in [k]$ and vertex $i \in V_M$, let

$$\bar{x}^{(i)} \triangleq (g_1(i), \ldots, g_\ell(i))^\intercal ; \tag{6}$$

notice that $\bar{x}^{(i)} \in \mathbb{R}^\ell$ defines the spectral embedding of $i \in V_M$ through the bottom ℓ eigenvectors of $\mathcal{L}_M$.

Definition 2 ((θ, ℓ)-distinguishable graph)**.** *For any $M = (V_M, E_M, w_M)$ with k vertices, $\ell \in [k]$, and $\theta \in \mathbb{R}^+$, we say that M is (θ, ℓ)-distinguishable if*

- *it holds for any $i \in [k]$ that $\left\|\bar{x}^{(i)}\right\|^2 \geq \theta$, and*
- *it holds for any different $i, j \in [k]$ that*

$$\left\| \frac{\bar{x}^{(i)}}{\left\|\bar{x}^{(i)}\right\|} - \frac{\bar{x}^{(j)}}{\left\|\bar{x}^{(j)}\right\|} \right\|^2 \geq \theta.$$

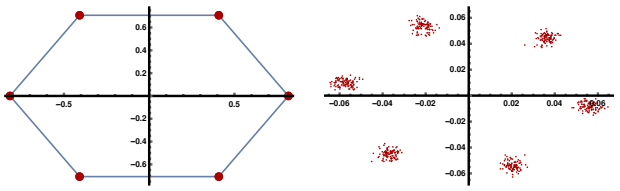
In other words, graph M is (θ, ℓ)-distinguishable if (i) every embedded point $\bar{x}^{(i)}$ has squared length at least θ, and (ii) any pair of embedded points with normalisation are separated by a distance of at least θ. By definition, it is easy to see that, if M is (θ, ℓ)-distinguishable for some large value of θ, then the embedded points $\{\bar{x}^{(i)}\}_{i \in V_M}$ can be easily separated even if $\ell < k$. The two examples below demonstrate that it is indeed the case and, since the meta-graph M is constructed from $\{S_i\}_{i=1}^k$, this well-separation property for $\{\bar{x}^{(i)}\}_{i \in V_M}$ usually implies that the clusters $\{S_i\}_{i=1}^k$ are also well-separated when the vertices are mapped to the points $\{F(u)\}_{u \in V_G}$, in which

$$F(u) \triangleq \frac{1}{\sqrt{d(u)}} \cdot (f_1(u), \ldots f_\ell(u))^\intercal . \tag{7}$$

Example 1. Suppose the meta-graph is C_6, the cycle on 6 vertices. Figure 2(a) shows that the vertices of C_6 are well separated by the second and third eigenvectors of $\mathcal{L}_{C_6}$.[3] Since the minimum distance between any pair of vertices in this embedding is $2/3$, we say that C_6 is $(2/3, 3)$-distinguishable. Figure 2(b) shows that, when using f_2, f_3 of $\mathcal{L}_G$ to embed the vertices of a 600-vertex graph with a cyclical cluster pattern, the embedded points closely match the ones from the meta-graph.

Example 2. Suppose the meta-graph is $P_{4,4}$, which is the 4×4 grid graph. Figure 3(a) shows that the vertices are separated using the second and third eigenvectors of $\mathcal{L}_{P_{4,4}}$. The minimum distance between any pair of vertices in this embedding is roughly 0.1, and so $P_{4,4}$ is $(0.1, 3)$-distinguishable. Figure 3(b) demonstrates that, when using f_2, f_3 of $\mathcal{L}_G$ to construct the embedding, the embedded points closely match the ones from the meta-graph.

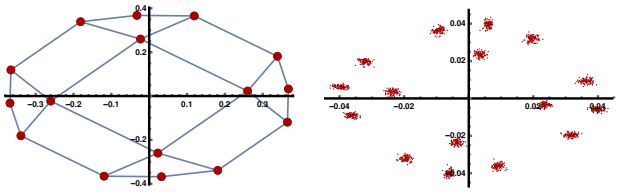
[3]Notice that the first eigenvector is the trivial one and gives no useful information. This is why we visualise the second and third eigenvectors only.



(a) Spectral embedding of C_6 with g_2 and g_3.

(b) Spectral embedding of G with f_2 and f_3.

Figure 2. The spectral embedding of a large graph G whose clusters exhibit a cyclical structure closely matches the embedding of the meta-graph C_6.



(a) Embedding of $P_{4,4}$ with g_2 and g_3.

(b) Embedding of G with f_2 and f_3.

Figure 3. The spectral embedding of a large 1600-vertex graph G whose clusters exhibit a grid structure closely matches the embedding of the meta-graph $P_{4,4}$.

From these examples, it is clear to see that there is a close connection between the embedding $\{\bar{x}^{(i)}\}$ defined in (6) and the embedding $\{F(u)\}$ defined in (7). To formally analyse this connection, we study the structure theorem with meta-graphs.

We define vectors $\bar{g}_1, \ldots, \bar{g}_k \in \mathbb{R}^n$ which represent the eigenvectors of $\mathcal{L}_M$ "blown-up" to n dimensions. Formally, we define $\bar{g}_i$ such that

$$\bar{g}_i = \sum_{j=1}^{k} \frac{D^{1/2}\chi_j}{\left\|D^{1/2}\chi_j\right\|} \cdot g_i(j),$$

where $\chi_j \in \mathbb{R}^n$ is the indicator vector of cluster S_j defined in (1). By definition, it holds for any $u \in S_j$ that

$$\bar{g}_i(u) = \sqrt{\frac{d(u)}{\mathrm{vol}(S_j)}} \cdot g_i(j).$$

The following lemma shows that $\{\bar{g}_i\}_{i=1}^k$ form an orthonormal basis.

Lemma 5.1. *The following statements hold:*

1. *it holds for any $i \in [k]$ that $\|\bar{g}_i\| = 1$;*
2. *it holds for any different $i, j \in [k]$ that $\langle \bar{g}_i, \bar{g}_j \rangle = 0$.*

Next, similar to the function $\Upsilon(k)$ defined in (2), for any input graph $G = (V_G, E_G, w_G)$ and (θ, ℓ)-distinguishable

meta-graph M, we define the function $\Psi(\ell)$ by

$$\Psi(\ell) \triangleq \sum_{i=1}^{\ell} \frac{\gamma_i}{\lambda_{\ell+1}}.$$

Notice that we have by the higher-order Cheeger inequality that $\gamma_i/2 \leq \rho_M(i)$ holds for any $i \in [\ell]$, and $\rho_M(i) \leq \rho_G(k)$ by the construction of matrix $\mathbf{A}_M$. Hence, one can view $\Psi(\ell)$ as a refined definition of $\Upsilon(k)$.

We now show that the vectors $f_1, \ldots, f_\ell$ and $\bar{g}_1, \ldots, \bar{g}_\ell$ are well approximated by each other. In order to show this, we define for any $i \in [\ell]$ the vectors

$$\widehat{f}_i = \sum_{j=1}^{\ell} \langle \bar{g}_i, f_j \rangle f_j \quad \text{and} \quad \widehat{g}_i = \sum_{j=1}^{\ell} \langle f_i, \bar{g}_j \rangle \bar{g}_j,$$

and present the structure theorem with meta-graphs.

Theorem 3 (The Structure Theorem with Meta-Graphs)**.** *The following statements hold:*

1. *it holds for any $i \in [\ell]$ that $\left\| \bar{g}_i - \widehat{f}_i \right\|^2 \leq \gamma_i/\lambda_{\ell+1}$;*
2. *it holds for any $\ell \in [k]$ that*

$$\sum_{i=1}^{\ell} \left\| f_i - \widehat{g}_i \right\|^2 \leq \sum_{i=1}^{\ell} \frac{\gamma_i}{\lambda_{\ell+1}}.$$

5.2. Spectral Clustering with Fewer Eigenvectors

Now we sketch our analysis of spectral clustering with fewer eigenvectors. Our presented algorithm is essentially the same as the standard spectral clustering described in Section 4, with the only difference that every $u \in V_G$ is embedded into a point in $\mathbb{R}^\ell$ by the mapping defined in (7). Our analysis follows from the one from Section 4 at a very high level. However, since we require that $\{F(u)\}_{u \in V_G}$ are well separated in $\mathbb{R}^\ell$ for some $\ell < k$, the proof is more involved.

For any $i \in [k]$, we define the approximate centre $p^{(i)} \in \mathbb{R}^\ell$ of every cluster S_i by

$$p^{(i)}(j) = \frac{1}{\sqrt{\mathrm{vol}(S_i)}} \cdot \sum_{x=1}^{\ell} \langle f_j, \bar{g}_x \rangle \cdot g_x(i),$$

and prove that the total k-means cost for the points $\{F(u)\}_{u \in V_G}$ can be upper bounded.

Lemma 5.2. *It holds that*

$$\sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| F(u) - p^{(i)} \right\|^2 \leq \Psi(\ell).$$

Secondly, we prove that the distance between different $p^{(i)}$ and $p^{(j)}$ can be lower bounded with respect to θ and $\Psi(\ell)$. This result is summarised as follows:

Lemma 5.3. *It holds for different $i, j \in [k]$ that*

$$\left\| p^{(i)} - p^{(j)} \right\|^2 \geq \frac{\theta^2 - 20\sqrt{\theta \cdot \Psi(\ell)}}{16 \min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}.$$

It is important to recognise that the lower bound in Lemma 5.3 implies a condition on θ and $\Psi(\ell)$ under which $p^{(i)}$ and $p^{(j)}$ are well-spread. In other words, spectral clustering with few eigenvectors works when the optimal clusters present a noticeable pattern. Combining these with the other technical ingredients, including our developed technique for constructing the desired mapping σ^* sketched in Section 4.2, we obtain the performance guarantee of our designed algorithm, which is summarised as follows:

Theorem 4. *Let G be a graph with k clusters $\{S_i\}_{i=1}^k$ of almost balanced size, with a (θ, ℓ)-distinguishable meta-graph that satisfies $\Psi(\ell) \leq (2176 \cdot (1 + \mathsf{APT}))^{-1} \cdot \theta^3$. Let $\{A_i\}_{i=1}^k$ be the output of spectral clustering with ℓ eigenvectors, and without loss of generality let the optimal correspondent of A_i be S_i. Then, it holds that*

$$\sum_{i=1}^{k} \mathrm{vol}\left(A_i \triangle S_i\right) \leq 2176 \cdot (1 + \mathsf{APT}) \cdot \frac{\Psi(\ell) \cdot \mathrm{vol}(V_G)}{k \cdot \theta^2}.$$

Notice that if we take $\ell = k$, then we have that $\theta = 1$ and $\Psi(\ell) \leq k/\Upsilon(k)$ which makes the guarantee in Theorem 4 the same as the one in Theorem 2. However, if the meta-graph corresponding to the optimal clusters is (θ, ℓ)-distinguishable for large θ and $\ell \ll k$, then we have that $\Psi(\ell) \ll k/\Upsilon(k)$ and Theorem 4 gives a stronger guarantee than the one from Theorem 2.

6. Experimental Results

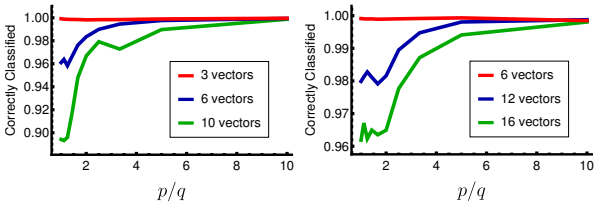
In this section we empirically evaluate the performance of spectral clustering for finding k clusters while using fewer than k eigenvectors. Our results on synthetic data demonstrate that for graphs with a clear pattern of clusters, spectral clustering with fewer than k eigenvectors performs better. This is further confirmed on real-world datasets including BSDS, MNIST, and USPS. We detail the experiment setup in the appendix, and the code to reproduce our results is available at `https://github.com/pmacg/spectral-clustering-meta-graphs`.

6.1. Results on Synthetic Data

We first study the performance of spectral clustering on random graphs whose clusters exhibit a clear pattern. Given the parameters $n \in \mathbb{Z}^+$, $0 \leq q \leq p \leq 1$, and some meta-graph $M = (V_M, E_M)$ with k vertices, we generate a graph with clusters $\{S_i\}_{i=1}^k$, each of size n, as follows. For each pair of vertices $u \in S_i$ and $v \in S_j$, we add the edge (u, v) with probability p if $i = j$ and with probability q if $i \neq j$

and $(i, j) \in E_M$. The metric used for our evaluation is defined by $\frac{1}{nk}\sum_{i=1}^{k}|S_i \cap A_i|$, for the optimal matching between the output $\{A_i\}$ and the ground truth $\{S_i\}$.

In our experiments, we fix $n = 1,000$, $p = 0.01$, and consider the meta-graphs C_{10} and $P_{4,4}$, similar to those illustrated in Examples 1 and 2; this results in graphs with $10,000$ and $16,000$ vertices respectively. We vary the ratio p/q and the number of eigenvectors used to find the clusters. Our experimental result, which is reported as the average score over 10 trials and shown in Figure 4, clearly shows that spectral clustering with fewer than k eigenvectors performs better. This is particularly the case when p and q are close, which corresponds to the more challenging regime in the model.



(a) Meta-Graph C_{10} (b) Meta-Graph $P_{4,4}$

Figure 4. A comparison on the output of spectral clustering with meta-graphs C_{10} and $P_{4,4}$, when different number of eigenvectors are used.

6.2. Results on the BSDS Dataset

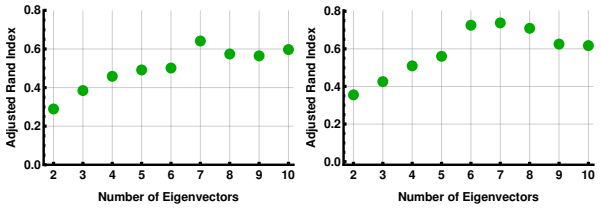
In this experiment, we study the performance of spectral clustering for image segmentation when using different numbers of eigenvectors. We consider the Berkeley Segmentation Data Set (BSDS) (Arbelaez et al., 2011), which consists of 500 images along with their ground-truth segmentations. For each image, we construct a similarity graph on the pixels and take k to be the number of clusters in the ground-truth segmentation. Then we apply spectral clustering, varying the number of eigenvectors used, and evaluate the output using the Rand Index (Rand, 1971). Figure 1 shows two images from the dataset along with the segmentations produced with spectral clustering, and more details and examples are shown in Appendix D. These examples illustrate that spectral clustering with fewer eigenvectors performs better. This is confirmed in our experiments on the entire BSDS dataset. The average Rand Index of the algorithm's output is reported in Table 1, and it is clear to see that using $k/2$ eigenvectors consistently out-performs spectral clustering with k eigenvectors.

6.3. Results on the MNIST and USPS Datasets

We further demonstrate the applicability of our results on the MNIST and USPS datasets (Hull, 1994; LeCun et al., 1998), which consist of images of hand-written digits, and the goal is to cluster the data into 10 clusters corresponding to different digits. For each dataset, we construct the k-nearest neighbour graph for $k = 3$, which results in a graph with $60,000$ vertices for the MNIST dataset and $7,291$ vertices for the USPS dataset. We use spectral clustering to partition the graphs into 10 clusters. We measure the similarity between the found clusters and the ground truth using the Adjusted Rand Index (Gates & Ahn, 2017), and plot the results in Figure 5. Our experiments show that spectral clustering with just 7 eigenvectors gives the best performance on both datasets. Appendix D includes results with additional clustering metrics which also show that using 7 eigenvectors is optimal.

Number of Eigenvectors	Average Rand Index
k	0.71
$k/2$	0.74
OPTIMAL	0.76

Table 1. The average Rand Index across the BSDS dataset for different numbers of eigenvectors. OPTIMAL refers to the algorithm which runs spectral clustering with ℓ eigenvectors for all possible $\ell \in [k]$ and returns the output with the highest Rand Index.



(a) MNIST (b) USPS

Figure 5. Experimental results on the MNIST and USPS datasets. These experiments show that spectral clustering with 7 eigenvectors gives the best partition of the input into 10 clusters.

7. Future Work

Our work leaves a number of interesting questions for future research. For spectral clustering, the only non-trivial assumption remaining in our analysis is that the optimal clusters have almost balanced size. It is unclear whether, under the regime of $\Upsilon(k) = \Omega(1)$, this condition could be eventually removed, or if there's some hard instance showing that our analysis is tight. For spectral clustering with fewer eigenvectors, our presented work is merely the starting point, and leaves many open questions. For example, although one can enumerate the number of used eigenvectors from 1 to k and take the clustering with the minimum k-way expansion, we are interested to know whether the optimal number of eigenvectors can be computed directly,

and rigorously analysed for different graph instances. We believe that the answers to these questions would not only significantly advance our understanding of spectral clustering, but also, as suggested in our experimental studies, have widespread applications in analysing real-world datasets.

Acknowledgements

This work is supported by a Langmuir PhD Scholarship, and an EPSRC Early Career Fellowship (EP/T00729X/1).

References

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

Chen, J., Sun, H., Woodruff, D. P., and Zhang, Q. Communication-optimal distributed clustering. In *30th Advances in Neural Information Processing Systems (NeurIPS'16)*, pp. 3720–3728, 2016.

Chung, F. R. *Spectral graph theory*. American Mathematical Soc., 1997.

Cucuringu, M., Li, H., Sun, H., and Zanetti, L. Hermitian matrices for clustering directed graphs: insights and applications. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20)*, 2020.

Czumaj, A., Peng, P., and Sohler, C. Testing cluster structure of graphs. In *47th Annual ACM Symposium on Theory of Computing (STOC'15)*, pp. 723–732, 2015.

Dey, T. K., Peng, P., Rossi, A., and Sidiropoulos, A. Spectral concentration and greedy k-clustering. *Computational Geometry: Theory and Applications*, 76:19–32, 2019.

Gates, A. J. and Ahn, Y.-Y. The impact of random models on clustering similarity. *The Journal of Machine Learning Research*, 18(1):3049–3076, 2017.

Gharan, S. O. and Trevisan, L. Partitioning into expanders. In *25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'14)*, pp. 1256–1266, 2014.

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

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

Kloumann, I. M., Ugander, J., and Kleinberg, J. Block models and personalized pagerank. *Proceedings of the National Academy of Sciences*, 114(1):33–38, 2017.

Kolev, P. and Mehlhorn, K. A note on spectral clustering. In *24th Annual European Symposium on Algorithms (ESA'16)*, pp. 57:1–57:14, 2016.

Kumar, A., Sabharwal, Y., and Sen, S. A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04)*, pp. 454–462, 2004.

Kwok, T. C., Lau, L. C., Lee, Y. T., Gharan, S. O., and Trevisan, L. 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)*, pp. 11–20, 2013.

Laenen, S. and Sun, H. Higher-order spectral clustering of directed graphs. In *34th Advances in Neural Information Processing Systems (NeurIPS'20)*, 2020.

Lancichinetti, A., Fortunato, S., and Kertész, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3), 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, J. R., Gharan, S. O., and Trevisan, L. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM*, 61(6):37:1–37:30, 2014.

Louis, A. and Venkat, R. Planted models for k-way edge and vertex expansion. In *39th Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS'19)*, pp. 23:1–23:15, 2019.

Mizutani, T. Improved analysis of spectral algorithm for clustering. *Optimization Letters*, 15(4):1303–1325, 2021.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *15th Advances in Neural Information Processing Systems (NeurIPS'01)*, pp. 849–856, 2001.

Orecchia, L. and Zhu, Z. A. Flow-based algorithms for local graph clustering. In *25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'14)*, pp. 1267–1286, 2014.

Peng, P. Robust clustering oracle and local reconstructor of cluster structure of graphs. In *31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'20)*, pp. 2953–2972, 2020.

Peng, P. and Yoshida, Y. Average sensitivity of spectral clustering. In *26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'20)*, pp. 1132–1140, 2020.

Peng, R., Sun, H., and Zanetti, L. Partitioning well-clustered graphs: Spectral clustering works! *SIAM J. Comput.*, 46 (2):710–743, 2017.

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

Rebagliati, N. and Verri, A. Spectral clustering with more than k eigenvectors. *Neurocomputing*, 74(9):1391–1401, 2011.

Spielman, D. A. and Teng, S.-H. Spectral partitioning works: Planar graphs and finite element meshes. In *37th Conference on Foundations of Computer Science (FOCS'96)*, pp. 96–105, 1996.

Sun, H. and Zanetti, L. Distributed graph clustering and sparsification. *ACM Transactions on Parallel Computing*, 6(3):17:1–17:23, 2019.

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

Zhu, Z. A., Lattanzi, S., and Mirrokni, V. S. A local algorithm for finding well-connected clusters. In *30th International Conference on Machine Learning (ICML'13)*, volume 28, pp. 396–404, 2013.

A. Omitted Details from Section 3

In this section, we present the proof of the improved structure theorem.

Proof of Theorem 1. Let $\widehat{f}_i = \sum_{j=1}^{k}\langle \bar{g}_i, f_j\rangle f_j$, and we write $\bar{g}_i$ as a linear combination of the vectors $f_1,\ldots,f_n$ by $\bar{g}_i = \sum_{j=1}^{n}\langle \bar{g}_i, f_j\rangle f_j$. Since $\widehat{f}_i$ is a projection of $\bar{g}_i$, we have that $\bar{g}_i - \widehat{f}_i$ is perpendicular to $\widehat{f}_i$ and

$$\left\|\bar{g}_i - \widehat{f}_i\right\|^2 = \|\bar{g}_i\|^2 - \left\|\widehat{f}_i\right\|^2 = \left(\sum_{j=1}^{n}\langle \bar{g}_i, f_j\rangle^2\right) - \left(\sum_{j=1}^{k}\langle \bar{g}_i, f_j\rangle^2\right) = \sum_{j=k+1}^{n}\langle \bar{g}_i, f_j\rangle^2.$$

Now, let us consider the quadratic form

$$\bar{g}_i^{\mathsf{T}}\mathcal{L}_G\bar{g}_i = \left(\sum_{j=1}^{n}\langle \bar{g}_i, f_j\rangle f_j^{\mathsf{T}}\right)\mathcal{L}_G\left(\sum_{j=1}^{n}\langle \bar{g}_i, f_j\rangle f_j\right) = \sum_{j=1}^{n}\langle \bar{g}_i, f_j\rangle^2\lambda_j \geq \lambda_{k+1}\left\|\bar{g}_i - \widehat{f}_i\right\|^2, \tag{8}$$

where the last inequality follows by the fact that $\lambda_i \geq 0$ holds for any $1 \leq i \leq n$. This gives us that

$$\begin{aligned}
\bar{g}_i^{\mathsf{T}}\mathcal{L}_G\bar{g}_i &= \sum_{(u,v)\in E_G} w(u,v)\left(\frac{\bar{g}_i(u)}{\sqrt{d(u)}} - \frac{\bar{g}_i(v)}{\sqrt{d(v)}}\right)^2 \\
&= \sum_{(u,v)\in E_G} w(u,v)\left(\frac{\chi_i(u)}{\sqrt{\mathrm{vol}(S_i)}} - \frac{\chi_i(v)}{\sqrt{\mathrm{vol}(S_i)}}\right)^2 \\
&= \frac{w(S_i, V\setminus S_i)}{\mathrm{vol}(S_j)} \leq \rho(k).
\end{aligned} \tag{9}$$

Combining (8) with (9), we have that

$$\left\|\bar{g}_i - \widehat{f}_i\right\|^2 \leq \frac{\bar{g}_i^{\mathsf{T}}\mathcal{L}_G\bar{g}_i}{\lambda_{k+1}} \leq \frac{\rho(k)}{\lambda_{k+1}} \leq \frac{1}{\Upsilon(k)},$$

which proves the first statement of the theorem.

Now we prove the second statement. We define for any $1 \leq i \leq k$ that $\widehat{g}_i = \sum_{j=1}^{k}\langle f_i, \bar{g}_j\rangle\bar{g}_j$, and have that

$$\begin{aligned}
\sum_{i=1}^{k}\|f_i - \widehat{g}_i\|^2 &= \sum_{i=1}^{k}\left(\|f_i\|^2 - \|\widehat{g}_i\|^2\right) = k - \sum_{i=1}^{k}\sum_{j=1}^{k}\langle \bar{g}_j, f_i\rangle^2 \\
&= \sum_{j=1}^{k}\left(1 - \sum_{i=1}^{k}\langle \bar{g}_j, f_i\rangle^2\right) = \sum_{j=1}^{k}\left(\|\bar{g}_j\|^2 - \left\|\widehat{f}_j\right\|^2\right) \\
&= \sum_{j=1}^{k}\left\|\bar{g}_j - \widehat{f}_j\right\|^2 \leq \sum_{j=1}^{k}\frac{1}{\Upsilon(k)} = \frac{k}{\Upsilon(k)},
\end{aligned}$$

where the last inequality follows by the first statement of Theorem 1. □

B. Omitted Details from Section 4

This section discusses the details omitted from Section 4, and the section is organised as follows: we first analyse the properties of spectral embedding in Section B.1; in Section B.2, we analyse the approximation guarantee of spectral clustering, and prove Theorem 2. Throughout the paper, the COST function for any partition $A_1,\ldots,A_k$ of the vertices of G is defined by

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

where d is the dimension used for the embedding; i.e., the COST function minimises the total ℓ_2-distance between the points $F(u)$'s and their individually closest centre c_i, where $c_1,\ldots,c_k$ are chosen arbitrarily in $\mathbb{R}^d$.

B.1. Omitted Analysis on the Spectral Embedding

We first prove Lemma 4.1, which states that the total ℓ_2^2-distance between all the embedded points $\{F(u)\}_u \in V_G$ and their corresponding approximate centres $\{p^{(i)}\}_{i=1}^k$ is upper bounded with respect to $\Upsilon(k)$.

Proof of Lemma 4.1. We have

$$\begin{aligned}\sum_{i=1}^{k}\sum_{u\in S_i} d(u)\left\|F(u)-p^{(i)}\right\|^2 &= \sum_{i=1}^{k}\sum_{u\in S_i} d(u)\left[\sum_{j=1}^{k}\left(\frac{f_j(u)}{\sqrt{d(u)}}-\frac{\langle \bar{g}_i, f_j\rangle}{\sqrt{\mathrm{vol}(S_i)}}\right)^2\right]\\ &= \sum_{i=1}^{k}\sum_{u\in S_i}\sum_{j=1}^{k}\left(f_j(u)-\langle \bar{g}_i, f_j\rangle \bar{g}_i(u)\right)^2\\ &= \sum_{i=1}^{k}\sum_{u\in S_i}\sum_{j=1}^{k}\left(f_j(u)-\widehat{g}_j(u)\right)^2\\ &= \sum_{j=1}^{k}\left\|f_j-\widehat{g}_j\right\|^2 \leq \frac{k}{\Upsilon(k)},\end{aligned}$$

where the final inequality follows by the second statement of Theorem 1 and it holds for $u \in S_x$ that $\widehat{g}_i(u) = \sum_{j=1}^{k}\langle f_i, \bar{g}_j\rangle \bar{g}_j(u) = \langle f_i, \bar{g}_x\rangle \bar{g}_x(u)$. □

Next, we show that the $\|p^{(i)}\|^2$ approximately equals to $1/\mathrm{vol}(S_i)$, and this fact will be used in our subsequent analysis.

Lemma B.1. *It holds for any $i \in [k]$ that*

$$\frac{1}{\mathrm{vol}(S_i)}\left(1-\frac{1}{\Upsilon(k)}\right) \leq \left\|p^{(i)}\right\|^2 \leq \frac{1}{\mathrm{vol}(S_i)}.$$

Proof. By definition, we have

$$\mathrm{vol}(S_i)\left\|p^{(i)}\right\|^2 = \sum_{j=1}^{k}\langle f_j, \bar{g}_i\rangle^2 = \left\|\widehat{f}_i\right\|^2 = 1-\left\|\widehat{f}_i-\bar{g}_i\right\|^2 \geq 1-\frac{1}{\Upsilon(k)},$$

where the inequality follows by Theorem 1. The other direction of the inequality follows similarly. □

Lemma B.2. *It holds for any different $i, j \in [k]$ that*

$$\left\|\sqrt{\mathrm{vol}(S_i)}\cdot p^{(i)}-\sqrt{\mathrm{vol}(S_j)}\cdot p^{(j)}\right\|^2 \geq 2-\frac{8}{\Upsilon(k)}.$$

Proof. We have

$$\begin{aligned}\left\|\sqrt{\mathrm{vol}(S_i)}\cdot p^{(i)}-\sqrt{\mathrm{vol}(S_j)}\cdot p^{(j)}\right\|^2 &= \sum_{x=1}^{k}\left(\langle f_x, \bar{g}_i\rangle - \langle f_x, \bar{g}_j\rangle\right)^2\\ &= \left(\sum_{x=1}^{k}\langle f_x, \bar{g}_i\rangle^2\right)+\left(\sum_{x=1}^{k}\langle f_x, \bar{g}_j\rangle^2\right)-2\sum_{x=1}^{k}\langle f_x, \bar{g}_i\rangle\langle f_x, \bar{g}_j\rangle\\ &\geq \left\|\widehat{f}_i\right\|^2+\left\|\widehat{f}_j\right\|^2-2\left|\widehat{f}_i^{\intercal}\widehat{f}_j\right|\\ &\geq 2\left(1-\frac{1}{\Upsilon(k)}\right)-2\left|(\bar{g}_i+\widehat{f}_i-\bar{g}_i)^{\intercal}(\bar{g}_j+\widehat{f}_j-\bar{g}_j)\right|\\ &= 2\left(1-\frac{1}{\Upsilon(k)}\right)-2\left(\left|\langle \bar{g}_i, \widehat{f}_j-\bar{g}_j\rangle+\langle \bar{g}_j, \widehat{f}_i-\bar{g}_i\rangle+\langle \widehat{f}_i-\bar{g}_i, \widehat{f}_j-\bar{g}_j\rangle\right|\right)\\ &\geq 2\left(1-\frac{1}{\Upsilon(k)}\right)-6\cdot\frac{1}{\Upsilon(k)} \geq 2-\frac{8}{\Upsilon(k)}.\end{aligned}$$

□

Lemma B.3. *It holds for any different $i.j \in [k]$ that*

$$\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\|^2 \geq 2 - \frac{20}{\Upsilon(k)}.$$

Proof. Assume without loss of generality that $\sqrt{\mathrm{vol}(S_i)}\left\|p^{(i)}\right\| \leq \sqrt{\mathrm{vol}(S_j)}\left\|p^{(j)}\right\|$. Let $\alpha_i = \sqrt{\mathrm{vol}(S_i)}p^{(i)}$ and $\alpha_j = \sqrt{\mathrm{vol}(S_j)}p^{(j)}$ and notice that $\|\alpha_i\| \leq \|\alpha_j\| \leq 1$. Then, using Lemma B.1 we have

$$\begin{aligned}
\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\| &\geq \left\| \alpha_i - \frac{\|\alpha_i\|}{\|\alpha_j\|}\alpha_j \right\| \\
&\geq \|\alpha_i - \alpha_j\| - (\|\alpha_j\| - \|\alpha_i\|) \\
&\geq \sqrt{2 - \frac{8}{\Upsilon(k)}} - \left(\sqrt{\mathrm{vol}(S_j)}\left\|p^{(j)}\right\| - \sqrt{\mathrm{vol}(S_i)}\left\|p^{(i)}\right\| \right) \\
&\geq \sqrt{2}\left(1 - \frac{4}{\Upsilon(k)}\right) + \sqrt{1 - \frac{1}{\Upsilon(k)}} - 1 \\
&\geq \sqrt{2} - \frac{4\sqrt{2}}{\Upsilon(k)} - \frac{1}{\Upsilon(k)} \geq \sqrt{2} - \frac{7}{\Upsilon(k)},
\end{aligned}$$

where the second inequality follows by the triangle inequality, and the third and fourth use Lemma B.2. We also use the fact that for $x \leq 1$, it is the case that $\sqrt{1-x} \geq 1 - x$. This gives

$$\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\|^2 \geq 2 - \frac{14\sqrt{2}}{\Upsilon(k)} \geq 2 - \frac{20}{\Upsilon(k)},$$

which proves the lemma. □

With these lemmas above, we're ready to prove Lemma 4.2, which lower bounds the distance between different $p^{(i)}$ and $p^{(j)}$.

Proof of Lemma 4.2. Assume without loss of generality that $\left\|p^{(i)}\right\| \geq \left\|p^{(j)}\right\|$. Then, let $\left\|p^{(j)}\right\| = \alpha\left\|p^{(i)}\right\|$ for some $\alpha \in [0, 1]$. By Lemma B.1, it holds that

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

Additionally, notice that we have by the proof of Lemma B.3 that

$$\left\langle \frac{p^{(i)}}{\left\|p^{(i)}\right\|}, \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\rangle \leq \sqrt{2} - \frac{1}{2}\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\| \leq \frac{\sqrt{2}}{2} + \frac{7}{2\Upsilon(k)},$$

where we use the fact that if $x^2 + y^2 = 1$, then $x + y \leq \sqrt{2}$. One can understand the equation above by considering the right-angled triangle with one edge given by $p^{(i)}/\left\|p^{(i)}\right\|$ and another edge given by $(p^{(i)}/\left\|p^{(i)}\right\|).(p^{(j)}/\left\|p^{(j)}\right\|)$. Then,

$$\begin{aligned}
\left\|p^{(i)} - p^{(j)}\right\|^2 &= \left\|p^{(i)}\right\|^2 + \left\|p^{(j)}\right\|^2 - 2\left\langle \frac{p^{(i)}}{\left\|p^{(i)}\right\|}, \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\rangle \left\|p^{(i)}\right\|\left\|p^{(j)}\right\| \\
&\geq (1+\alpha)\left\|p^{(i)}\right\|^2 - \left(\sqrt{2} + \frac{7}{\Upsilon(k)}\right)\alpha\left\|p^{(i)}\right\|^2 \\
&\geq \left(1 - (\sqrt{2}-1)\alpha - \frac{7}{\Upsilon(k)}\right)\left\|p^{(i)}\right\|^2 \\
&\geq \frac{1}{\min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}\left(\frac{1}{2} - \frac{7}{\Upsilon(k)}\right)\left(1 - \frac{1}{\Upsilon(k)}\right) \\
&\geq \frac{1}{\min\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\}}\left(\frac{1}{2} - \frac{8}{\Upsilon(k)}\right),
\end{aligned}$$

which completes the proof. □

B.2. Omitted Analysis on Approximation Guarantee of Spectral Clustering

In this subsection, we analyse the total volume of misclassified vertices by spectral clustering, and prove Theorem 2. We will first prove that, under the function $\sigma : [k] \to [k]$ defined in (4), the value of $\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)}$ will be upper bounded.

Proof of Lemma 4.3. Let us define $B_{ij} = A_i \cap S_j$ to be the vertices in A_i which belong to the true cluster S_j. Then, we have that

$$\begin{aligned}\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} &= \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq \sigma(i)}}^{k} \mathrm{vol}(B_{ij}) \left(\frac{1}{\mathrm{vol}(S_{\sigma(i)})} + \frac{1}{\mathrm{vol}(S_j)} \right) \\ &\leq 2 \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq \sigma(i)}}^{k} \frac{\mathrm{vol}(B_{ij})}{\min\{\mathrm{vol}(S_{\sigma(i)}), \mathrm{vol}(S_j)\}}, \end{aligned} \tag{10}$$

and that

$$\begin{aligned}\mathrm{COST}(A_1, \dots A_k) &= \sum_{i=1}^{k} \sum_{u \in A_i} d(u) \left\| F(u) - c_i \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} d(u) \left\| F(u) - c_i \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} d(u) \left(\frac{\left\| p^{(j)} - c_i \right\|^2}{2} - \left\| p^{(j)} - F(u) \right\|^2 \right) \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} \frac{d(u) \left\| p^{(j)} - p^{(\sigma(i))} \right\|^2}{8} - \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq i}} \sum_{u \in B_{ij}} d(u) \left\| p^{(j)} - F(u) \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \mathrm{vol}(B_{ij}) \frac{\left\| p^{(j)} - p^{(\sigma(i))} \right\|^2}{8} - \sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| p^{(i)} - F(u) \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \frac{\mathrm{vol}(B_{ij})}{8 \min\{\mathrm{vol}(S_{\sigma(i)}), \mathrm{vol}(S_j)\}} \left(\frac{1}{2} - \frac{8}{\Upsilon(k)} \right) - \sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| p^{(i)} - F(u) \right\|^2 \\ &\geq \frac{1}{16} \cdot \left(\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \right) \left(\frac{1}{2} - \frac{8}{\Upsilon(k)} \right) - \frac{k}{\Upsilon(k)}, \end{aligned}$$

where the second inequality follows by the inequality of $\|a-b\|^2 \geq \frac{\|b-c\|^2}{2} - \|a-c\|^2$, the third inequality follows since c_i is closer to $p^{(\sigma(i))}$ than $p^{(j)}$, the fifth one follows from Lemma 4.2, and the last one follows by (10).

On the other side, since $\mathsf{COST}(A_1, \dots, A_k) \leq \mathsf{APT} \cdot k/\Upsilon(k)$, we have that

$$\frac{1}{16} \cdot \left(\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \right) \left(\frac{1}{2} - \frac{8}{\Upsilon(k)} \right) - \frac{k}{\Upsilon(k)} \leq \mathsf{APT} \cdot \frac{k}{\Upsilon(k)}.$$

This implies that

$$\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 16 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)} \cdot \left(\frac{1}{2} - \frac{8}{\Upsilon(k)} \right)^{-1} \leq 64 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)},$$

where the last inequality holds by the assumption that $\Upsilon(k) \geq 32$. □

Next, we show how to apply Lemma 4.3 to prove Theorem 2. In particular, we will present a method to construct our desired permutation from the function σ defined in (4) without significantly increasing the overall cost, and show why the final constructed permutation suffices for our purpose.

Proof of Theorem 2. By Lemma 4.1, we have

$$\mathsf{COST}(S_1, \ldots, S_k) \leq \frac{k}{\Upsilon(k)}.$$

Combining this with the fact that one can apply an approximate k-means clustering algorithm with approximation ratio APT for spectral clustering, we have that

$$\mathsf{COST}(A_1, \ldots, A_k) \leq \mathsf{APT} \cdot \frac{k}{\Upsilon(k)}.$$

Then, let $\sigma : [1, k] \to [1, k]$ be the function which assigns the clusters $A_1, \ldots, A_k$ to the ground truth clusters such that $\sigma(i) = \arg\min_{j \in [k]} \left\| p^{(j)} - c_i \right\|$. Then, it holds by Lemma 4.3 that

$$\epsilon(\sigma) \triangleq \sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 64 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}. \tag{11}$$

Now, assume that σ is not a permutation, and we'll apply the following procedure inductively to construct a permutation from σ. Since σ isn't a permutation, there is $i \in [k]$ such that $M_{\sigma,i} = \emptyset$. Therefore, there are different values of $x, y \in [k]$ such that $\sigma(x) = \sigma(y) = j$ for some $j \neq i$. Based on this, we construct the function $\sigma' : [k] \to [k]$ such that $\sigma'(z) = i$ if $z = x$, and $\sigma'(z) = \sigma(z)$ for any other $z \in [k] \setminus \{x\}$. Notice that we can construct σ' in this way as long as σ isn't a permutation. By the definition of $\epsilon(\cdot)$ and function σ', the difference between $\epsilon(\sigma')$ and $\epsilon(\sigma)$ can be written as

$$\epsilon(\sigma') - \epsilon(\sigma) = \underbrace{\left(\frac{\mathrm{vol}(M_{\sigma',i} \triangle S_i)}{\mathrm{vol}(S_i)} - \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \right)}_{=:\alpha} + \underbrace{\left(\frac{\mathrm{vol}(M_{\sigma',j} \triangle S_j)}{\mathrm{vol}(S_j)} - \frac{\mathrm{vol}(M_{\sigma,j} \triangle S_j)}{\mathrm{vol}(S_j)} \right)}_{=:\beta}. \tag{12}$$

Let us consider 4 cases based on the sign of α, β defined above. In each case, we bound the cost introduced by the change from σ to σ', and then consider the total cost introduced throughout the entire procedure of constructing a permutation.

Case 1: $\alpha < 0, \beta < 0$**.** In this case, it is clear that $\epsilon(\sigma') - \epsilon(\sigma) \leq 0$, and hence the total introduced cost is at most 0.

Case 2: $\alpha > 0, \beta < 0$**.** In this case, we have

$$\begin{aligned}
\epsilon(\sigma') - \epsilon(\sigma) &\leq \frac{1}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \left(\mathrm{vol}(M_{\sigma',i} \triangle S_i) - \mathrm{vol}(M_{\sigma,i} \triangle S_i) + |\mathrm{vol}(M_{\sigma',j} \triangle S_j) - \mathrm{vol}(M_{\sigma,j} \triangle S_j)| \right) \\
&= \frac{1}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \left(\mathrm{vol}(M_{\sigma',i} \triangle S_i) - \mathrm{vol}(M_{\sigma,i} \triangle S_i) + \mathrm{vol}(M_{\sigma,j} \triangle S_j) - \mathrm{vol}(M_{\sigma',j} \triangle S_j) \right) \\
&= \frac{1}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \left(\mathrm{vol}(A_x \setminus S_i) - \mathrm{vol}(A_x \cap S_i) + \mathrm{vol}(A_x \setminus S_j) - \mathrm{vol}(A_x \cap S_j) \right) \\
&\leq \frac{2 \cdot \mathrm{vol}(A_x \setminus S_j)}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \\
&\leq \frac{8 \cdot \mathrm{vol}(A_x \setminus S_j)}{\mathrm{vol}(S_j)},
\end{aligned}$$

where the last inequality follows by the fact that the clusters are almost balanced. Since each set A_x is moved at most once in order to construct a permutation, the total introduced cost due to this case is at most

$$\sum_{j=1}^{k} \sum_{A_x \in M_{\sigma,j}} \frac{8 \cdot \mathrm{vol}(A_x \setminus S_j)}{\mathrm{vol}(S_j)} \leq 8 \cdot \sum_{j=1}^{k} \frac{\mathrm{vol}(M_{\sigma,j} \triangle S_j)}{\mathrm{vol}(S_j)} \leq 512 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}.$$

Case 3: $\alpha > 0, \beta > 0$**.** In this case, we have

$$\begin{aligned}
\epsilon(\sigma') - \epsilon(\sigma) &\leq \frac{1}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \left(\mathrm{vol}(M_{\sigma',i} \triangle S_i) - \mathrm{vol}(M_{\sigma,i} \triangle S_i) + \mathrm{vol}(M_{\sigma',j} \triangle S_j) - \mathrm{vol}(M_{\sigma,j} \triangle S_j)\right) \\
&= \frac{1}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \left(\mathrm{vol}(A_x \setminus S_i) - \mathrm{vol}(A_x \cap S_i) + \mathrm{vol}(A_x \cap S_j) - \mathrm{vol}(A_x \setminus S_j)\right) \\
&\leq \frac{1}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \left(2 \cdot \mathrm{vol}(A_x \cap S_j)\right) \\
&\leq \frac{2 \cdot \mathrm{vol}(S_j)}{\min(\mathrm{vol}(S_i), \mathrm{vol}(S_j))} \\
&\leq 8,
\end{aligned}$$

where the last inequality follows by the fact that the clusters are almost balanced. We will consider the total introduced cost due to this case and Case 4 together, and so let's first examine Case 4.

Case 4: $\alpha < 0, \beta > 0$**.** In this case, we have

$$\begin{aligned}
\epsilon(\sigma') - \epsilon(\sigma) &\leq \frac{1}{\mathrm{vol}(S_j)} \left(\mathrm{vol}(M_{\sigma',j} \triangle S_j) - \mathrm{vol}(M_{\sigma,j} \triangle S_j)\right) \\
&\leq \frac{1}{\mathrm{vol}(S_j)} \left(\mathrm{vol}(A_x \cap S_j) - \mathrm{vol}(A_x \setminus S_j)\right) \\
&\leq \frac{\mathrm{vol}(S_j)}{\mathrm{vol}(S_j)} \\
&= 1.
\end{aligned}$$

Now, let us bound the total number of times we need to construct σ' in order to obtain a permutation. For any i with $M_{\sigma,i} = \emptyset$, we have

$$\frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} = \frac{\mathrm{vol}(S_i)}{\mathrm{vol}(S_i)} = 1,$$

so the total number of required iterations is upper bounded by

$$|\{i : M_{\sigma,i} = \emptyset\}| \leq \sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 64 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}.$$

As such, the total introduced cost due to Cases 3 and 4 is at most

$$8 \cdot 64 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)} = 512 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}.$$

Putting everything together, we have that

$$\epsilon(\sigma^\star) \leq \epsilon(\sigma) + 1024 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)} \leq 1088 \cdot (1 + \mathsf{APT}) \cdot \frac{k}{\Upsilon(k)}.$$

This implies that

$$\sum_{i=1}^{k} \mathrm{vol}(M_{\sigma^\star,i} \triangle S_i) \leq 2176 \cdot (1 + \mathsf{APT}) \cdot \frac{\mathrm{vol}(V_G)}{\Upsilon(k)}$$

and completes the proof. □

C. Omitted Details from Section 5

This section presents the omitted details from Section 5.

C.1. Omitted Details for Section 5.1

We first analyse the properties of $\{\bar{g}_i\}_{i=1}^k$, and prove Lemma 5.1.

Proof of Lemma 5.1. By definition, we have that

$$\|\bar{g}_i\|^2 = \bar{g}_i^\intercal \bar{g}_i = \sum_{j=1}^k \sum_{u\in S_j} \bar{g}_i(u)^2 = \sum_{j=1}^k \sum_{u\in S_j} \left(\frac{\sqrt{d(u)}}{\sqrt{\mathrm{vol}(S_j)}}\cdot g_i(j)\right)^2 = \sum_{j=1}^k g_i(j)^2 = \|g_i\|^2 = 1,$$

which proves the first statement.

To prove the second statement, we have for any $i \neq j$ that

$$\langle \bar{g}_i, \bar{g}_j\rangle = \sum_{x=1}^k \sum_{u\in S_x} \bar{g}_i(u)\bar{g}_j(u) = \sum_{x=1}^k \sum_{u\in S_x} \frac{d(u)}{\mathrm{vol}(S_x)}\cdot g_i(x)g_j(x) = \sum_{x=1}^k g_i(x)g_j(x) = g_i^\intercal g_j = 0,$$

which proves the second statement. □

We will later use the following important relationship between the eigenvalues of $\mathcal{L}_M$ and $\mathcal{L}_G$.

Lemma C.1. *It holds for any $1 \leq i \leq k$ that $\lambda_i \leq \gamma_i$.*

Proof. Notice that we have for any $j \leq k$ that

$$\begin{aligned}
\bar{g}_j^\intercal \mathcal{L}_G \bar{g}_j &= \sum_{(u,v)\in E_G} w_G(u,v)\left(\frac{\bar{g}_j(u)}{\sqrt{d(u)}} - \frac{\bar{g}_j(v)}{\sqrt{d(v)}}\right)^2 \\
&= \sum_{x=1}^{k-1}\sum_{y=x+1}^{k}\sum_{a\in S_x}\sum_{b\in S_y} w(a,b)\left(\frac{\bar{g}_j(a)}{\sqrt{d(a)}} - \frac{\bar{g}_j(b)}{\sqrt{d(b)}}\right)^2 \\
&= \sum_{x=1}^{k-1}\sum_{y=x+1}^{k} w(S_x,S_y)\left(\frac{g_j(x)}{\sqrt{\mathrm{vol}(S_x)}} - \frac{g_j(y)}{\sqrt{\mathrm{vol}(S_y)}}\right)^2 \\
&= g_j^\intercal \mathcal{L}_M g_j \\
&= \gamma_j.
\end{aligned}$$

By Lemma 5.1, we have an i-dimensional subspace $S_i \subset \mathbb{R}^n$ such that

$$\max_{x\in S_i} \frac{x^\intercal \mathcal{L}_G x}{x^\intercal x} = \gamma_i,$$

from which the statement of the lemma follows by the Courant-Fischer theorem. □

Now we prove the structure theorem with meta-graphs.

Proof of Theorem 3. For the first statement, we write $\bar{g}_i$ as a linear combination of the vectors $f_1,\ldots,f_n$, i.e., $\bar{g}_i = \sum_{j=1}^n \langle \bar{g}_i, f_j\rangle f_j$. Since $\widehat{f}_i$ is a projection of $\bar{g}_i$, we have that $\bar{g}_i - \widehat{f}_i$ is perpendicular to $\widehat{f}_i$, and that

$$\left\|\bar{g}_i - \widehat{f}_i\right\|^2 = \|\bar{g}_i\|^2 - \left\|\widehat{f}_i\right\|^2 = \left(\sum_{j=1}^n \langle \bar{g}_i, f_j\rangle^2\right) - \left(\sum_{j=1}^{\ell} \langle \bar{g}_i, f_j\rangle^2\right) = \sum_{j=\ell+1}^{n} \langle \bar{g}_i, f_j\rangle^2.$$

Now, we study the quadratic form $\bar{g}_i^\intercal \mathcal{L}_G \bar{g}_i$ and have that

$$\bar{g}_i^\intercal \mathcal{L}_G \bar{g}_i = \left(\sum_{j=1}^n \langle \bar{g}_i, f_j\rangle f_j^\intercal\right)\mathcal{L}_G\left(\sum_{j=1}^n \langle \bar{g}_i, f_j\rangle f_j\right) = \sum_{j=1}^n \langle \bar{g}_i, f_j\rangle^2 \lambda_j \geq \lambda_{\ell+1}\left\|\bar{g}_i - \widehat{f}_i\right\|^2.$$

By the proof of Lemma C.1, we have that $\bar{g}_i^\intercal \mathcal{L}_G \bar{g}_i \leq \gamma_i$, from which the first statement follows.

Now we prove the second statement. We define the vectors $\bar{g}_{k+1}, \ldots, \bar{g}_n$ to be an arbitrary orthonormal basis of the space orthogonal to the space spanned by $\bar{g}_1, \ldots, \bar{g}_k$. Then, we can write any f_i as $f_i = \sum_{j=1}^{n} \langle f_i, \bar{g}_j \rangle \bar{g}_j$, and have that

$$\begin{aligned}\sum_{i=1}^{\ell} \|f_i - \widehat{g}_i\|^2 &= \sum_{i=1}^{\ell} \left(\|f_i\|^2 - \|\widehat{g}_i\|^2 \right) = \ell - \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \langle f_i, \bar{g}_j \rangle^2 = \sum_{j=1}^{\ell} \left(1 - \sum_{i=1}^{\ell} \langle \bar{g}_j, f_i \rangle^2 \right) \\ &= \sum_{j=1}^{\ell} \left(\|\bar{g}_j\|^2 - \left\|\widehat{f}_j\right\|^2 \right) = \sum_{j=1}^{\ell} \left\|\bar{g}_j - \widehat{f}_j\right\|^2 \leq \sum_{j=1}^{\ell} \frac{\gamma_j}{\lambda_{\ell+1}},\end{aligned}$$

where the last inequality follows by the first statement of the theorem. □

C.2. Omitted Details from Section 5.2

In this section we present all the technical details used for analysing the spectral embedding (7). We first prove Lemma 5.2, which upper bounds the total distance between all the embedded points and their corresponding approximate centres.

Proof of Lemma 5.2. By definition, it holds that

$$\begin{aligned}\sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| F(u) - p^{(i)} \right\|^2 &= \sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left[\sum_{j=1}^{\ell} \left(\frac{f_j(u)}{\sqrt{d(u)}} - \left(\sum_{x=1}^{\ell} \langle f_j, \bar{g}_x \rangle \frac{g_x(i)}{\sqrt{\mathrm{vol}(S_i)}} \right) \right)^2 \right] \\ &= \sum_{i=1}^{k} \sum_{u \in S_i} \sum_{j=1}^{\ell} \left(f_j(u) - \left(\sum_{x=1}^{\ell} \langle f_j, \bar{g}_x \rangle \bar{g}_x(u) \right) \right)^2 \\ &= \sum_{i=1}^{k} \sum_{u \in S_i} \sum_{j=1}^{\ell} \left(f_j(u) - \widehat{g}_j(u) \right)^2 \\ &= \sum_{j=1}^{\ell} \|f_j - \widehat{g}_j\|^2 \\ &\leq \Psi(\ell),\end{aligned}$$

where the final inequality follows from the second statement of Theorem 3. □

Lemma C.2. *It holds for $i \in [k]$ that*

$$\left(1 - \frac{4\sqrt{\Psi(\ell)}}{\theta} \right) \frac{\left\|\bar{x}^{(i)}\right\|^2}{\mathrm{vol}(S_i)} \leq \left\|p^{(i)}\right\|^2 \leq \left(1 + \frac{2\sqrt{\Psi(\ell)}}{\theta} \right) \frac{\left\|\bar{x}^{(i)}\right\|^2}{\mathrm{vol}(S_i)}.$$

Proof. It holds by definition that

$$\begin{aligned}\mathrm{vol}(S_i) \cdot \left\|p^{(i)}\right\|^2 &= \sum_{x=1}^{\ell} \left(\sum_{y=1}^{\ell} \langle f_x, \bar{g}_y \rangle g_y(i) \right)^2 \\ &= \sum_{x=1}^{\ell}\sum_{y=1}^{\ell}\sum_{z=1}^{\ell} \langle f_x, \bar{g}_y \rangle \langle f_x, \bar{g}_z \rangle g_y(i) g_z(i) \\ &= \sum_{x=1}^{\ell}\sum_{y=1}^{\ell} \langle f_x, \bar{g}_y \rangle^2 g_y(i)^2 + \sum_{x=1}^{\ell}\sum_{y=1}^{\ell}\sum_{\substack{z=1 \\ z \neq y}}^{\ell} \langle f_x, \bar{g}_y \rangle \langle f_x, \bar{g}_z \rangle g_y(i) g_z(i) \\ &= \sum_{x=1}^{\ell}\sum_{y=1}^{\ell} \langle f_x, \bar{g}_y \rangle^2 g_y(i)^2 + \sum_{x=1}^{\ell}\sum_{y=1}^{\ell}\sum_{\substack{z=1 \\ z \neq y}}^{\ell} g_y(i) g_z(i) \cdot \left(\widehat{f}_y^\intercal \widehat{f}_z \right). \qquad (13)\end{aligned}$$

We study the two terms of (13) separately. For the second term, we have that

$$\begin{aligned}
\sum_{x=1}^{\ell}\sum_{y=1}^{\ell}\sum_{\substack{z=1\\z\neq y}}^{\ell} g_y(i)g_z(i)\cdot\left(\widehat{f}_y^{\intercal}\widehat{f}_z\right) &\leq \sum_{y=1}^{\ell}|g_y(i)|\sum_{\substack{1\leq z\leq \ell\\z\neq y}}|g_z(i)|\left|\widehat{f}_y^{\intercal}\widehat{f}_z\right| \\
&= \sum_{y=1}^{\ell}|g_y(i)|\sum_{\substack{1\leq z\leq \ell\\z\neq y}}|g_z(i)|\left|\left(\bar{g}_y+\widehat{f}_y-\bar{g}_y\right)^{\intercal}\left(\bar{g}_z+\widehat{f}_z-\bar{g}_z\right)\right| \\
&= \sum_{y=1}^{\ell}|g_y(i)|\sum_{\substack{1\leq z\leq \ell\\z\neq y}}|g_z(i)|\left|\langle\widehat{f}_y-\bar{g}_y,\bar{g}_z\rangle+\langle\widehat{f}_z-\bar{g}_z,\bar{g}_y\rangle+\langle\widehat{f}_y-\bar{g}_y,\widehat{f}_z-\bar{g}_z\rangle\right| \\
&= \sum_{y=1}^{\ell}|g_y(i)|\sum_{\substack{z=1\\z\neq y}}^{\ell}|g_z(i)|\left|\langle\widehat{f}_y-\bar{g}_y,\bar{g}_z\rangle\right| \\
&\leq \sqrt{\left(\sum_{y=1}^{\ell}g_y(i)^2\right)\sum_{y=1}^{\ell}\left(\sum_{\substack{1\leq z\leq \ell\\z\neq y}}|g_z(i)|\left|\langle\widehat{f}_y-\bar{g}_y,\bar{g}_z\rangle\right|\right)^2} \\
&\leq \sqrt{\sum_{y=1}^{\ell}\left(\sum_{\substack{1\leq z\leq \ell\\z\neq y}}g_z(i)^2\right)\left(\sum_{\substack{1\leq z\leq \ell\\z\neq y}}\langle\widehat{f}_y-\bar{g}_y,\bar{g}_z\rangle^2\right)} \\
&\leq \sqrt{\sum_{y=1}^{\ell}\sum_{\substack{1\leq z\leq \ell\\z\neq y}}\langle\widehat{f}_y-\bar{g}_y,\bar{g}_z\rangle^2} \\
&\leq \sqrt{\sum_{y=1}^{\ell}\left\|\widehat{f}_y-\bar{g}_y\right\|^2} \\
&\leq \sqrt{\Psi(\ell)},
\end{aligned}$$

where we used the fact that $\sum_{y=1}^{k} g_y(i)^2 = 1$ for all $i \in [k]$. Therefore, we have that

$$\begin{aligned}
\mathrm{vol}(S_i)\left\|p^{(i)}\right\|^2 &\leq \sum_{y=1}^{\ell}\left(\sum_{x=1}^{\ell}\langle f_x,\bar{g}_y\rangle^2\right)g_y(i)^2+\sqrt{\Psi(\ell)} \leq \sum_{y=1}^{\ell}g_y(i)^2+\sqrt{\Psi(\ell)} \leq \left\|\bar{x}^{(i)}\right\|^2+\sqrt{\Psi(\ell)} \\
&\leq \left\|\bar{x}^{(i)}\right\|^2\left(1+\frac{2\sqrt{\Psi(\ell)}}{\theta}\right).
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
\mathrm{vol}(S_i)\left\|p^{(i)}\right\|^2 &\geq \sum_{y=1}^{\ell}\left(\sum_{x=1}^{\ell}\langle f_x,\bar{g}_y\rangle^2\right)g_y(i)^2-2\sqrt{\Psi(\ell)} \geq \sum_{y=1}^{\ell}\left\|\widehat{f}_y\right\|^2 g_y(i)^2-2\sqrt{\Psi(\ell)} \\
&\geq (1-\Psi(\ell))\left\|\bar{x}^{(i)}\right\|^2-2\sqrt{\Psi(\ell)} \\
&\geq \left\|\bar{x}^{(i)}\right\|^2\left(1-\frac{4\sqrt{\Psi(\ell)}}{\theta}\right),
\end{aligned}$$

where the last inequality holds by the fact that $\left\|\bar{x}^{(i)}\right\| \leq 1$ and $\Psi(\ell) < 1$. Hence, the statement holds. □

Lemma C.3. *It holds for $i \neq j$ that*

$$\left\| \frac{\sqrt{\mathrm{vol}(S_i)}}{\left\|\bar{x}^{(i)}\right\|} \cdot p^{(i)} - \frac{\sqrt{\mathrm{vol}(S_j)}}{\left\|\bar{x}^{(j)}\right\|} \cdot p^{(j)} \right\|^2 \geq \theta - 3\sqrt{\Psi(\ell)}.$$

Proof. By definition, it holds that

$$\begin{aligned}
&\left\| \frac{\sqrt{\mathrm{vol}(S_i)}}{\left\|\bar{x}^{(i)}\right\|} \cdot p^{(i)} - \frac{\sqrt{\mathrm{vol}(S_j)}}{\left\|\bar{x}^{(j)}\right\|} \cdot p^{(j)} \right\|^2 \\
=&\sum_{t=1}^{\ell} \left(\sum_{y=1}^{\ell} \langle f_t, \bar{g}_y \rangle \left(\frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right) \right)^2 \\
=&\sum_{t=1}^{\ell} \sum_{y=1}^{\ell} \langle f_t, \bar{g}_y \rangle^2 \left(\frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right)^2 \\
&\qquad + \sum_{t=1}^{\ell} \sum_{y=1}^{\ell} \sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \langle f_t, \bar{g}_y \rangle \langle f_t, \bar{g}_z \rangle \left(\frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right) \left(\frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right).
\end{aligned}$$

We upper bound the second term by

$$\begin{aligned}
&\sum_{y=1}^{\ell} \left| \frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \left| \frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \sum_{t=1}^{\ell} \left| \langle f_t, \bar{g}_y \rangle \langle f_t, \bar{g}_x \rangle \right| \\
=&\sum_{y=1}^{\ell} \left| \frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \left| \frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \left| \widehat{f}_y^{\mathsf{T}} \widehat{f}_z \right| \\
=&\sum_{y=1}^{\ell} \left| \frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \left| \frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \left| \left(\bar{g}_y + \widehat{f}_y - \bar{g}_y \right)^{\mathsf{T}} \left(\bar{g}_z + \widehat{f}_z - \bar{g}_z \right) \right| \\
=&\sum_{y=1}^{\ell} \left| \frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \left| \frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \left| \langle \widehat{f}_y - \bar{g}_y, \bar{g}_z \rangle \right| \\
\leq& \sqrt{\left(\sum_{y=1}^{\ell} \left(\frac{g_y(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_y(j)}{\left\|\bar{x}^{(j)}\right\|} \right)^2 \right) \sum_{y=1}^{\ell} \left(\sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \left| \frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right| \left| \langle \widehat{f}_y - \bar{g}_y, \bar{g}_z \rangle \right| \right)^2} \\
\leq& \sqrt{2 \sum_{y=1}^{\ell} \left(\sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \left(\frac{g_z(i)}{\left\|\bar{x}^{(i)}\right\|} - \frac{g_z(j)}{\left\|\bar{x}^{(j)}\right\|} \right)^2 \right) \left(\sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \langle \widehat{f}_y - \bar{g}_y, \bar{g}_z \rangle^2 \right)} \\
\leq& 2 \sqrt{\sum_{y=1}^{\ell} \sum_{\substack{1 \leq z \leq \ell \\ z \neq y}} \langle \widehat{f}_y - \bar{g}_y, \bar{g}_z \rangle^2} \\
\leq& 2 \sqrt{\sum_{y=1}^{\ell} \left\| \widehat{f}_y - \bar{g}_y \right\|^2} \\
\leq& 2\sqrt{\Psi(\ell)},
\end{aligned}$$

from which we can conclude that

$$\left\| \frac{\sqrt{\mathrm{vol}(S_i)}}{\left\|\bar{x}^{(i)}\right\|} \cdot p^{(i)} - \frac{\sqrt{\mathrm{vol}(S_j)}}{\left\|\bar{x}^{(j)}\right\|} \cdot p^{(j)} \right\|^2 \geq (1 - \Psi(\ell))\,\theta - 2\sqrt{\Psi(\ell)} \geq \theta - 3\sqrt{\Psi(\ell)}.$$

With this we proved the statement. □

Lemma C.4. *It holds for any different $i, j \in [k]$ that* $\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\|^2 \geq \frac{\theta}{4} - 8\sqrt{\frac{\Psi}{\theta}}$.

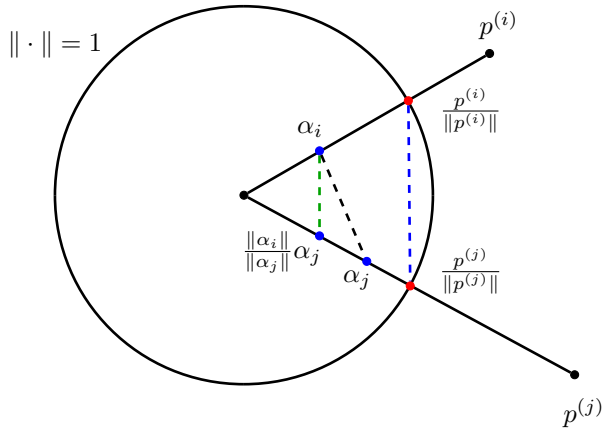


Figure 6. Illustration of the proof of Lemma C.4. Our goal is to give a lower bound on the length of $\left(\frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|}\right)$, which is the blue dotted line in the figure. We instead calculate a lower bound on the length of $\left(\alpha_i - \frac{\|\alpha_i\|}{\|\alpha_j\|}\right)$, which is the green dotted line, and use the fact that $\|\alpha_i\| \leq 1$ and $\|\alpha_j\| \leq 1$.

Proof. We set the parameter $\epsilon = 4\sqrt{\Psi}/\theta$, and define $\alpha_i = \frac{\sqrt{\mathrm{vol}(S_i)}}{(1+\epsilon)\left\|\bar{x}^{(i)}\right\|} \cdot p^{(i)}$, as well as $\alpha_j = \frac{\sqrt{\mathrm{vol}(S_j)}}{(1+\epsilon)\left\|\bar{x}^{(j)}\right\|} \cdot p^{(j)}$. By the definition of ϵ and Lemma C.2, it holds that $\|\alpha_i\| \leq 1$, and $\|\alpha_j\| \leq 1$. We can also assume without loss of generality that $\|\alpha_i\| \leq \|\alpha_j\|$. Then, as illustrated in Figure 6, we have

$$\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\|^2 \geq \left\| \alpha_i - \frac{\|\alpha_i\|}{\|\alpha_j\|}\alpha_j \right\|^2,$$

and so it suffices to lower bound the right-hand side of the inequality above. By the triangle inequality, we have

$$\begin{aligned}
\left\| \alpha_i - \frac{\|\alpha_i\|}{\|\alpha_j\|}\alpha_j \right\| &\geq \|\alpha_i - \alpha_j\| - \left\| \alpha_j - \frac{\|\alpha_i\|}{\|\alpha_j\|}\alpha_j \right\| \\
&= \frac{1}{1+\epsilon} \left\| \frac{\sqrt{\mathrm{vol}(S_i)}}{\left\|\bar{x}^{(i)}\right\|} \cdot p^{(i)} - \frac{\sqrt{\mathrm{vol}(S_j)}}{\left\|\bar{x}^{(j)}\right\|} \cdot p^{(j)} \right\| - \left(\|\alpha_j\| - \|\alpha_i\|\right).
\end{aligned}$$

Now, we have that

$$\|\alpha_j\| - \|\alpha_i\| = \frac{\sqrt{\mathrm{vol}(S_j)}}{(1+\epsilon)\left\|\bar{x}^{(j)}\right\|} \cdot \left\|p^{(j)}\right\| - \frac{\sqrt{\mathrm{vol}(S_i)}}{(1+\epsilon)\left\|\bar{x}^{(i)}\right\|} \cdot \left\|p^{(i)}\right\| \leq 1 - \frac{1-\epsilon}{1+\epsilon} = \frac{2\epsilon}{1+\epsilon},$$

and have by Lemma C.3 that

$$\left\| \frac{\sqrt{\mathrm{vol}(S_i)}}{\left\|\bar{x}^{(i)}\right\|} \cdot p^{(i)} - \frac{\sqrt{\mathrm{vol}(S_j)}}{\left\|\bar{x}^{(j)}\right\|} \cdot p^{(j)} \right\| \geq \sqrt{\theta - 3\sqrt{\Psi(\ell)}} \geq \sqrt{\theta} - \sqrt{2\epsilon} \geq \sqrt{\theta} - 2\epsilon.$$

since $\epsilon = 4\sqrt{\Psi}/\theta < 1$ by the assumption on Ψ. This gives us that

$$\left\| \alpha_i - \frac{\|\alpha_i\|}{\|\alpha_j\|}\alpha_j \right\| \geq \frac{\sqrt{\theta} - 2\epsilon}{1+\epsilon} - \frac{2\epsilon}{1+\epsilon} \geq \frac{1}{2}\left(\sqrt{\theta} - 4\epsilon\right).$$

Finally, we have that

$$\left\| \frac{p^{(i)}}{\left\|p^{(i)}\right\|} - \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\|^2 \geq \frac{1}{4}\left(\sqrt{\theta} - 16\frac{\sqrt{\Psi(\ell)}}{\theta}\right)^2 \geq \frac{\theta}{4} - 8\sqrt{\frac{\Psi(\ell)}{\theta}},$$

which completes the proof. □

Proof of Lemma 5.3. We assume without loss of generality that $\left\|p^{(i)}\right\|^2 \geq \left\|p^{(j)}\right\|^2$. Then, by Lemma C.2 and the fact that $\|\bar{x}^{(i)}\|^2 \geq \theta$ holds for any $i \in [k]$, we have

$$\left\|p^{(i)}\right\|^2 \geq \left(1 - \frac{4\sqrt{\Psi(\ell)}}{\theta}\right) \cdot \frac{\left\|\bar{x}^{(i)}\right\|^2}{\mathrm{vol}(S_i)}, \qquad \left\|p^{(j)}\right\|^2 \geq \left(1 - \frac{4\sqrt{\Psi(\ell)}}{\theta}\right) \cdot \frac{\left\|\bar{x}^{(j)}\right\|^2}{\mathrm{vol}(S_j)},$$

which implies that

$$\left\|p^{(i)}\right\|^2 \geq \frac{\theta - 4\sqrt{\Psi(\ell)}}{\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}}.$$

Now, we will proceed by case distinction.

Case 1: $\left\|p^{(i)}\right\| \geq 4\left\|p^{(j)}\right\|$. In this case, we have $\left\|p^{(i)} - p^{(j)}\right\| \geq \left\|p^{(i)}\right\| - \left\|p^{(j)}\right\| \geq \frac{3}{4}\left\|p^{(i)}\right\|$, and

$$\left\|p^{(i)} - p^{(j)}\right\|^2 \geq \frac{9}{16} \cdot \frac{\theta - 4\sqrt{\Psi(\ell)}}{\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}} \geq \frac{\theta\left(\theta - 20\sqrt{\Psi(\ell)/\theta}\right)}{16\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}} = \frac{\theta^2 - 20\sqrt{\theta \cdot \Psi(\ell)}}{16\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}},$$

since $\theta < 1$.

Case 2: $\left\|p^{(j)}\right\| = \alpha\left\|p^{(i)}\right\|$ for some $\alpha \in \left(\frac{1}{4}, 1\right]$. By Lemma C.4, we have

$$\left\langle \frac{p^{(i)}}{\left\|p^{(i)}\right\|}, \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\rangle \leq 1 - \frac{1}{2}\left(\frac{\theta}{4} - 8\sqrt{\frac{\Psi}{\theta}}\right) \leq 1 - \frac{\theta}{8} + 2\sqrt{\frac{\Psi}{\theta}}.$$

Then, it holds that

$$\begin{aligned}
\left\|p^{(i)} - p^{(j)}\right\|^2 &= \left\|p^{(i)}\right\|^2 + \left\|p^{(j)}\right\|^2 - 2\left\langle \frac{p^{(i)}}{\left\|p^{(i)}\right\|}, \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\rangle \left\|p^{(i)}\right\| \left\|p^{(j)}\right\| \\
&\geq \left(1 + \alpha^2\right)\left\|p^{(i)}\right\|^2 - 2\left(1 - \frac{\theta}{8} + 2\sqrt{\frac{\Psi(\ell)}{\theta}}\right)\alpha\left\|p^{(i)}\right\|^2 \\
&\geq \left(1 + \alpha^2 - 2\alpha + \frac{\theta}{4}\alpha - 4\sqrt{\frac{\Psi(\ell)}{\theta}}\alpha\right)\left\|p^{(i)}\right\|^2 \\
&\geq \left(\frac{\theta}{4} - 4\sqrt{\frac{\Psi(\ell)}{\theta}}\right) \cdot \alpha \cdot \frac{\theta - 4\sqrt{\Psi(\ell)}}{\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}} \\
&\geq \left(\frac{\theta}{16} - \sqrt{\frac{\Psi(\ell)}{\theta}}\right)\left(\theta - 4\sqrt{\Psi}\right) \cdot \frac{1}{\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}} \\
&\geq \left(\frac{\theta^2}{16} - \frac{5}{4}\sqrt{\theta\Psi(\ell)}\right) \cdot \frac{1}{\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}} \\
&= \frac{\theta^2 - 20\sqrt{\theta\Psi(\ell)}}{16\min\left\{\mathrm{vol}(S_i), \mathrm{vol}(S_j)\right\}}
\end{aligned}$$

which completes the proof. □

With these lemmas stated above, we analyse the performance of spectral clustering when fewer eigenvectors are employed to construct the embedding.

Lemma C.5. *Let $\{A_i\}_{i=1}^k$ be the output of spectral clustering, and σ and $M_{\sigma,i}$ be defined as in (4) and (5). Assuming $\Psi(\ell) \leq \theta^3/1600$, it holds that*

$$\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 64 \cdot (1 + \mathsf{APT}) \cdot \frac{\Psi(\ell)}{\theta^2}.$$

Proof. Let us define $B_{ij} = A_i \cap S_j$ to be the vertices in A_i which belong to the true cluster S_j. Then, we have that

$$\begin{aligned}\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} &= \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq \sigma(i)}}^{k} \mathrm{vol}(B_{ij}) \left(\frac{1}{\mathrm{vol}(S_{\sigma(i)})} + \frac{1}{\mathrm{vol}(S_j)} \right) \\ &\leq 2 \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq \sigma(i)}}^{k} \frac{\mathrm{vol}(B_{ij})}{\min\{\mathrm{vol}(S_{\sigma(i)}), \mathrm{vol}(S_j)\}}, \end{aligned} \tag{14}$$

and that

$$\begin{aligned}\mathrm{COST}(A_1, \dots A_k) &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} d(u) \left\| F(u) - c_i \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} d(u) \left(\frac{\left\| p^{(j)} - c_i \right\|^2}{2} - \left\| p^{(j)} - F(u) \right\|^2 \right) \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} \frac{d(u) \left\| p^{(j)} - p^{(\sigma(i))} \right\|^2}{8} - \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \sum_{u \in B_{ij}} d(u) \left\| p^{(j)} - F(u) \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \mathrm{vol}(B_{ij}) \frac{\left\| p^{(j)} - p^{(\sigma(i))} \right\|^2}{8} - \sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| p^{(i)} - F(u) \right\|^2 \\ &\geq \sum_{i=1}^{k} \sum_{\substack{1 \leq j \leq k \\ j \neq \sigma(i)}} \frac{\mathrm{vol}(B_{ij})}{16 \cdot \min\{\mathrm{vol}(S_{\sigma(i)}), \mathrm{vol}(S_j)\}} \left(\theta^2 - 20\sqrt{\theta \cdot \Psi(\ell)} \right) - \sum_{i=1}^{k} \sum_{u \in S_i} d(u) \left\| p^{(i)} - F(u) \right\|^2 \\ &\geq \frac{1}{32} \cdot \left(\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \right) \left(\theta^2 - 20\sqrt{\theta \cdot \Psi(\ell)} \right) - \Psi(\ell), \end{aligned}$$

where the second inequality follows by the inequality of

$$\|a - b\|^2 \geq \frac{\|b - c\|^2}{2} - \|a - c\|^2,$$

the third inequality follows since c_i is closer to $p^{(\sigma(i))}$ than $p^{(j)}$, the fifth one follows from Lemma 5.3, and the last one follows by (14).

On the other hand, since $\mathsf{COST}(A_1, \dots, A_k) \leq \mathsf{APT} \cdot \Psi(\ell)$ by Lemma 5.2, we have that

$$\sum_{i=1}^{k} \frac{\mathrm{vol}(M_{\sigma,i} \triangle S_i)}{\mathrm{vol}(S_i)} \leq 32 \cdot (1 + \mathsf{APT}) \cdot \Psi(\ell) \cdot \left(\theta^2 - 20\sqrt{\theta \cdot \Psi(\ell)} \right)^{-1} \leq 64 \cdot (1 + \mathsf{APT}) \cdot \Psi(\ell),$$

where the last inequality follows by the assumption that $\Psi(\ell) \leq \theta^3/1600$. Therefore, the statement follows. □

Proof of Theorem 4. This result can be obtained by using the same technique as the one used in the proof of Theorem 2, but applying Lemma 5.2 instead of Lemma 4.1 and Lemma C.5 instead of Lemma 4.3 in the analysis. □

D. Omitted Details from Section 6

In this section we provide additional details on our experimental setup and provide some supplementary results on the BSDS dataset. We implement the spectral clustering algorithm in Python, using the `scipy` library for computing eigenvectors, and the k-means algorithm from the `sklearn` library. Our experiments on synthetic data are performed on a desktop computer with an Intel(R) Core(TM) i5-8500 CPU @ 3.00GHz processor and 16 GB RAM. The experiments on the BSDS, MNIST, and USPS datasets are performed on a compute server with 64 AMD EPYC 7302 16-Core Processors.

D.1. Omitted Detail for the BSDS Experiment

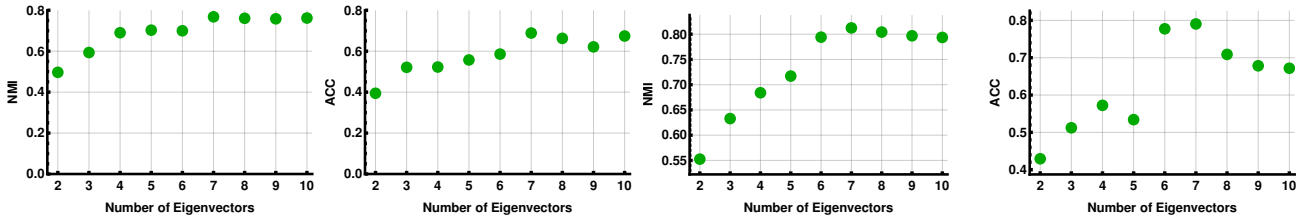
Ground-truth segmentations. The BSDS dataset provides several human-generated ground truth segmentations for each image. Since there are different numbers of ground truth clusterings associated with each image, in our experiments we take the target number of clusters for a given image to be the one closest to the median.

Constructing the similarity graph. Given a particular image in the dataset, we first downsample the image to have at most $20,000$ pixels. Then, we represent each pixel by the point $(r, g, b, x, y) \in \mathbb{R}^5$ where $r, g, b \in [1, 255]$ are the RGB values of the pixel and x and y are the coordinates of the pixel in the downsampled image. We construct the similarity graph by taking each pixel to be a vertex in the graph, and for every pair of pixels $u, v \in \mathbb{R}^5$, we add an edge with weight $\exp(-\|u - v\|^2 / 2\sigma^2)$ where $\sigma = 20$.

Evaluation. We evaluate each segmentation produced with spectral clustering using the Rand Index as implemented in the benchmarking code provided along with the BSDS dataset. For each image, this computes the average Rand Index across all of the provided ground-truth segmentations for the image. Figures 7 and 8 give some additional examples of our results from the BSDS dataset.

D.2. Omitted Detail for the MNIST and USPS Experiments

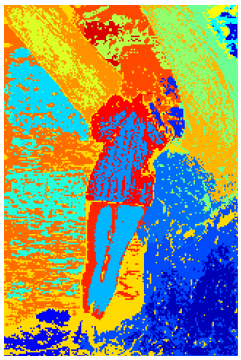
In both the MNIST and USPS datasets, each image is represented as an array of grayscale pixels with values between 0 and 255. The MNIST dataset has $60,000$ images with dimensions 28×28 and the USPS dataset has $7,291$ images with dimensions 16×16. In each case, we consider each image to be a single data point in $\mathbb{R}^{(d^2)}$ where d is the dimension of the images and construct the k-nearest neighbour graph for each dataset. For the MNIST dataset, this gives a graph with $60,000$ vertices and $138,563$ edges; for the USPS dataset, this gives a graph with $7,291$ vertices and $16,715$ edges. Figure 9 shows the accuracy (ACC) (Rand, 1971) and Normalised Mutual Information (NMI) (Lancichinetti et al., 2009) metrics for clustering with different numbers of eigenvectors; these results are consistent with the one based on ARI shown in Section 6.



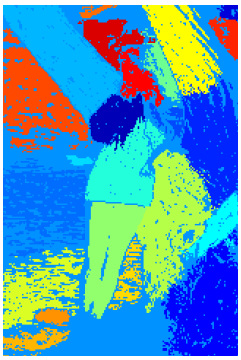
(a) MNIST NMI (b) MNIST ACC (c) USPS NMI (d) USPS ACC

Figure 9. The clustering accuracy (ACC) and Normalised Mutual Information (NMI) when clustering MNIST and USPS with different numbers of eigenvectors. Spectral clustering with 7 eigenvectors gives the best result across the two metrics.

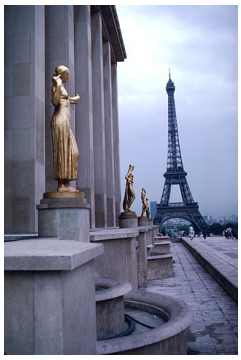
(a) Original Image



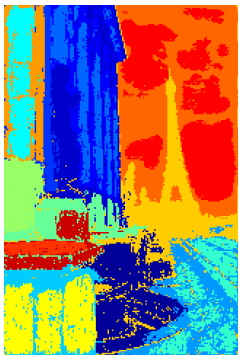
(b) Segmentation into 24 clusters with 8 eigenvectors; Rand Index 0.82.



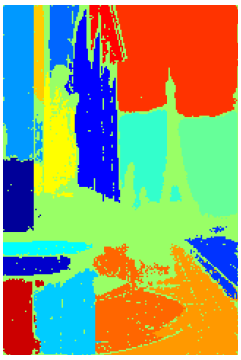
(c) Segmentation into 24 clusters with 24 eigenvectors; Rand Index 0.77.



(d) Original Image

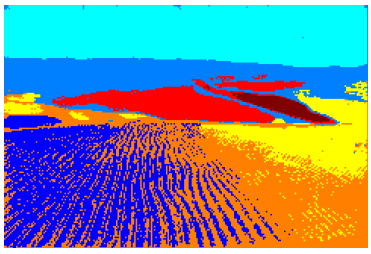


(e) Segmentation into 18 clusters with 6 eigenvectors; Rand Index 0.77.

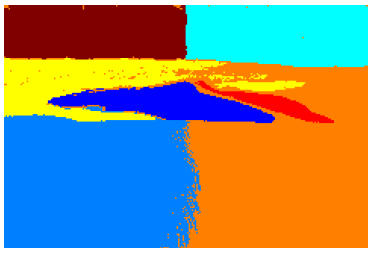


(f) Segmentation into 18 clusters with 18 eigenvectors; Rand Index 0.74.

(g) Original Image

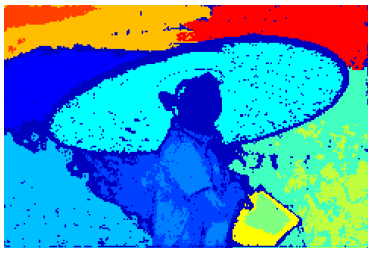


(h) Segmentation into 7 clusters with 3 eigenvectors; Rand Index 0.76.

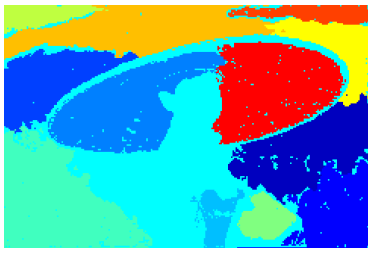


(i) Segmentation into 7 clusters with 7 eigenvectors; Rand Index 0.74.

(j) Original Image



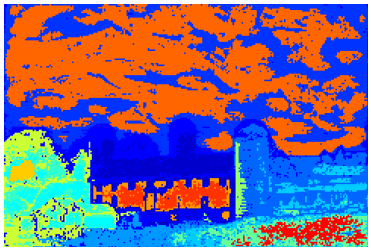
(k) Segmentation into 13 clusters with 8 eigenvectors; Rand Index 0.80.



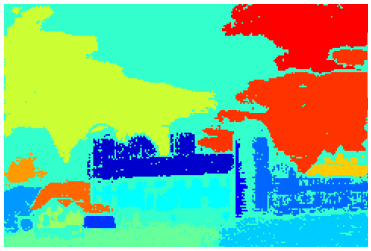
(l) Segmentation into 13 clusters with 13 eigenvectors; Rand Index 0.77.

Figure 7. Examples of the segmentations produced with spectral clustering on the BSDS dataset.

(a) Original Image

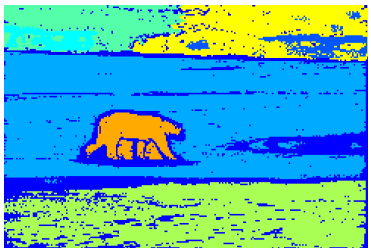


(b) Segmentation into 16 clusters with 4 eigenvectors; Rand Index 0.78.



(c) Segmentation into 16 clusters with 16 eigenvectors; Rand Index 0.65.

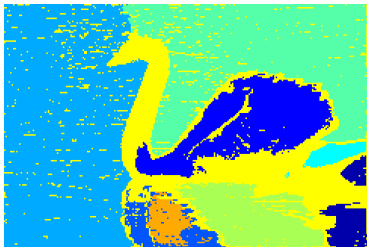
(d) Original Image



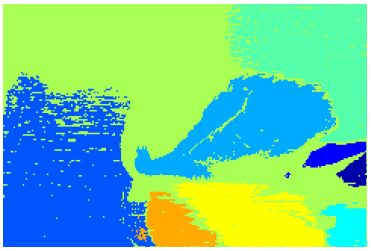
(e) Segmentation into 8 clusters with 5 eigenvectors; Rand Index 0.86.

(f) Segmentation into 8 clusters with 8 eigenvectors; Rand Index 0.79.

(g) Original Image



(h) Segmentation into 9 clusters with 7 eigenvectors; Rand Index 0.69.



(i) Segmentation into 9 clusters with 9 eigenvectors; Rand Index 0.61.

Figure 8. Examples of the segmentations produced with spectral clustering on the BSDS dataset.