Noise Thresholds for Spectral Clustering

Sivaraman Balakrishnan **Min Xu** **Akshay Krishnamurthy** **Aarti Singh**

School of Computer Science, Carnegie Mellon University
{sbalakri,minx,akshaykr,aarti}@cs.cmu.edu

Abstract

Although spectral clustering has enjoyed considerable empirical success in machine learning, its theoretical properties are not yet fully developed. We analyze the performance of a spectral algorithm for hierarchical clustering and show that on a class of hierarchically structured similarity matrices, this algorithm can tolerate noise that grows with the number of data points while still perfectly recovering the hierarchical clusters with high probability. We additionally improve upon previous results for k-way spectral clustering to derive conditions under which spectral clustering makes no mistakes. Further, using minimax analysis, we derive tight upper and lower bounds for the clustering problem and compare the performance of spectral clustering to these information theoretic limits. We also present experiments on simulated and real world data illustrating our results.

1 Introduction

Clustering, a fundamental and ubiquitous problem in machine learning, is the task of organizing data points into homogenous groups using a given measure of similarity. Two popular forms of clustering are *k-way*, where an algorithm directly partitions the data into k disjoint sets, and *hierarchical*, where the algorithm organizes the data into a hierarchy of groups. Popular algorithms for the k-way problem include k-means, spectral clustering, and density-based clustering, while *agglomerative* methods that merge clusters from the bottom up are popular for the latter problem.

Spectral clustering algorithms embed the data points by projection onto a few eigenvectors of (some form of) the graph Laplacian matrix and use this spectral embedding to find a clustering. This technique has been shown to work on various arbitrarily shaped clusters and, in addition to being straightforward to implement, often outperforms traditional clustering algorithms such as the k-means algorithm.

Real world data is inevitably corrupted by noise and it is of interest to study the robustness of spectral clustering algorithms. This is the focus of our paper.

Our main contributions are:

- We leverage results from perturbation theory in a novel analysis of a spectral algorithm for hierarchical clustering to understand its behavior in the presence of noise. We provide strong guarantees on its correctness; in particular, we show that the amount of noise spectral clustering tolerates can grow rapidly with the size of the smallest cluster we want to resolve.
- We sharpen existing results on k-way spectral clustering. In contrast with earlier work, we provide precise error bounds through a careful characterization of a k-means style algorithm run on the spectral embedding of the data.
- We also address the issue of optimal noise thresholds via the use of minimax theory. In particular, we establish tight information-theoretic upper and lower bounds for cluster resolvability.

2 Related Work and Definitions

There are several high-level justifications for the success of spectral clustering. The algorithm has deep connections to various graph-cut problems, random walks on graphs, electric network theory, and via the graph Laplacian to the Laplace-Beltrami operator. See [16] for an overview.

Several authors (see von Luxburg et. al. [17] and references therein) have shown various forms of asymptotic convergence for the Laplacian of a graph constructed from random samples drawn from a distribution on or near a manifold. These results however often do not easily translate into precise guarantees for successful recovery of clusters, which is the emphasis of our work.

There has also been some theoretical work on spectral algorithms for cluster recovery in random graph models. McSherry [9] studies the "cluster-structured" random graph model in which the probability of adding an edge can vary depending on the clusters the edge connects. He considers a specialization of this model, the planted partition model, which specifies only two probabilities, one for inter-cluster edges and another for intra-cluster edges. In this case, we can view the observed adjacency matrix as a random perturbation of a low rank "expected" adjacency matrix which encodes the cluster membership. McSherry shows that one can recover the clusters from a low rank approximation of the observed (noisy) adjacency matrix. These results show that low-rank matrices have spectra that are robust to noise. Our results however, show that we can obtain similar insensitivity (to noise) guarantees for a class of interesting structured *full-rank* matrices, indicating that this robustness extends to a much broader class of matrices.

More recently, Rohe et al [11] analyze spectral clustering in the stochastic block model (SBM), which is an example of a structured random graph. They consider the *high-dimensional* scenario where the number of clusters k grows with the number of data points n and show that under certain assumptions the *average* number of mistakes made by spectral clustering $\to 0$ with increasing n. Our work on hierarchical clustering also has the same high-dimensional flavor since the number of clusters we resolve grows with n. However, in the hierarchical clustering setting, errors made at the bottom level propogate up the tree and we need to make precise arguments to ensure that the *total* number of errors $\to 0$ with increasing n (see Theorem 1).

Since Rohe et al [11] and McSherry [9] consider random graph models, the "noise" on each entry has *bounded* variance. We consider more general noise models and study the relation between errors in clustering and noise variance. Another related line of work is on the problem of spectrally separating mixtures of Gaussians [1, 2, 8].

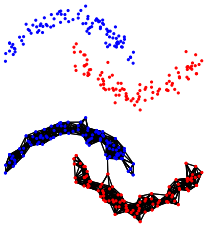
Ng et al. [10] study k-way clustering and show that the eigenvectors of the graph Laplacian are stable in 2-norm under small perturbations. This justifies the use of k-means in the perturbed subspace since ideally without noise, the spectral embedding by the top k eigenvectors of the graph Laplacian reflects the true cluster memberships, However, closeness in 2-norm does not translate into a strong bound on the *total number* of errors made by spectral clustering.

Huang et al. [7] study the misclustering rate of spectral clustering under the somewhat unnatural assumption that every coordinate of the Laplacian's eigenvectors are perturbed by independent and identically distributed noise. In contrast, we specify our noise model as an additive perturbation to the similarity matrix, making no direct assumptions on how this affects the spectrum of the Laplacian. We show that the eigenvectors are stable in ∞-norm and use this result to precisely bound the misclustering rate of our algorithm.

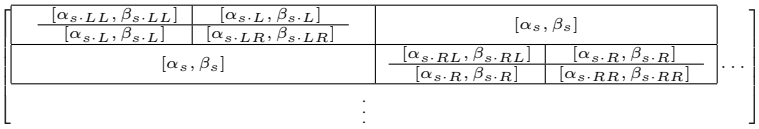
2.1 Definitions

The clustering problem can be defined as follows: Given an $(n \times n)$ similarity matrix on n data points, find a set $\mathcal{C}$ of subsets of the points such that points belonging to the same subset have high similarity and points in different subsets have low similarity. Our first results focus on *binary* hierarchical clustering, which is formally defined as follows:

Definition 1 *A **hierarchical clustering** $\mathcal{T}$ on data points $\{X_i\}_{i=1}^n$ is a collection of clusters (subsets of the points) such that $C_0 := \{X_i\}_{i=1}^n \in \mathcal{T}$ and for any $C_i, C_j \in \mathcal{T}$, either $C_i \subset C_j$, $C_j \subset C_i$, or $C_i \cap C_j = \emptyset$. A **binary hierarchical clustering** $\mathcal{T}$ is a hierarchical clustering such that for each non-atomic $C_k \in \mathcal{T}$, there exists two proper subsets $C_i, C_j \in \mathcal{T}$ with $C_i \cap C_j = \emptyset$ and $C_i \cup C_j = C_k$. We label each cluster by a sequence s of Ls and Rs so that $C_{s \cdot L}$ and $C_{s \cdot R}$ partitions C_s, $C_{s \cdot LL}$ and $C_{s \cdot LR}$ partititons $C_{s \cdot L}$, and so on.*



(a)



(b)

Figure 1: (a): Two moons data set (Top). For a similarity function defined on the ϵ-neighborhood graph (Bottom), this data set forms an ideal matrix. (b) An ideal matrix for the hierarchical problem.

Ideally, we would like that at all levels of the hierarchy, points within a cluster are more similar to each other than to points outside of the cluster. For a suitably chosen similarity function, a data set consisting of clusters that lie on arbitrary manifolds with complex shapes can result in this ideal case. As an example, in the two-moons data set in Figure 1(a), the popular technique of constructing a nearest neighbor graph and defining the distance between two points as the length of the *longest* edge on the *shortest* path between them results in an ideal similarity matrix. Other non-Euclidean similarity metrics (for instance density based similarity metrics [12]) can also allow for non-parametric cluster shapes.

For such ideal similarity matrices, we can show that the spectral clustering algorithm will deterministically recover all clusters in the hierarchy (see Theorem 5 in the appendix). However, since this ideal case does not hold in general, we focus on similarity matrices that can be decomposed into an ideal matrix and a high-variance noise term.

Definition 2 *A similarity matrix W is a **noisy hierarchical block matrix** (noisy HBM) if $W \triangleq A+R$ where A is ideal and R is a perturbation matrix, defined as follows:*

- *An **ideal similarity matrix**, shown in Figure 1(b), is characterized by ranges of off-block-diagonal similarity values $[\alpha_s, \beta_s]$ for each cluster C_s such that if $x \in C_{s\cdot L}$ and $y \in C_{s\cdot R}$ then $\alpha_s \leq A_{xy} \leq \beta_s$. Additionally, $\min\{\alpha_{s\cdot R}, \alpha_{s\cdot L}\} > \beta_s$.*
- *A symmetric $(n \times n)$ matrix R is a **perturbation matrix** with parameter σ if (a) $\mathbb{E}(R_{ij}) = 0$, (b) the entries of R are subgaussian, that is $\mathbb{E}(\exp(tR_{ij})) \leq \exp(\frac{\sigma^2 t^2}{2})$ and (c) for each row i, $R_{i1}, \ldots, R_{in}$ are independent.*

The perturbations we consider are quite general and can accommodate bounded (with σ upper bounded by the range), Gaussian (where σ is the standard deviation), and several other common distributions. This model is well-suited to noise that arises from the direct measurement of similarities. It is also possible to assume instead that the measurements of individual data points are noisy though we do not focus on this case in our paper.

In the k-way case, we consider the following similarity matrix which is studied by Ng et. al [10].

Definition 3 *W is a **noisy k-Block Diagonal** matrix if $W \triangleq A + R$ where R is a perturbation matrix and A is an ideal matrix for the k-way problem. An ideal matrix for the k-way problem has within-cluster similarities larger than $\beta_0 > 0$ and between cluster similarities 0.*

Finally, we define the combinatorial Laplacian matrix, which will be the focus of our spectral algorithm and our subsequent analysis.

Definition 4 *The **combinatorial Laplacian** L of a matrix W is defined as $L \triangleq D - W$ where D is a diagonal matrix with $D_{ii} \triangleq \sum_{j=1}^{n} W_{ij}$.*

We note that other analyses of spectral clustering have studied other Laplacian matrices, particularly, the *normalized Laplacians* defined as $L_n \triangleq D^{-1}L$ and $L_n \triangleq D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. However as we show in Appendix E, the normalized Laplacian can mis-cluster points even for an ideal noiseless similarity matrix.

Algorithm 1 HS

input (noisy) $n \times n$ similarity matrix W
  Compute Laplacian $L = D - W$
  $v_2 \leftarrow$ smallest non-constant eigenvector of L
  $C_1 \leftarrow \{i : v_2(i) \geq 0\}$, $C_2 \leftarrow \{j : v_2(j) < 0\}$
  $\mathcal{C} \leftarrow \{C_1, C_2\} \cup$ HS $(W_{C_1}) \cup$ HS (W_{C_2})
output $\mathcal{C}$

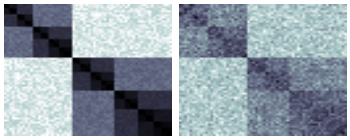


Figure 2: An ideal matrix and a noisy HBM. Clusters at finer granularity are masked by noise.

Algorithm 2 K-WAY SPECTRAL

input (noisy) $n \times n$ similarity matrix W, number of clusters k
  Compute Laplacian $L = D - W$
  $V \leftarrow (n \times k)$ matrix with columns $v_1, ..., v_k$, where $v_i \triangleq i$th smallest eigenvector of L
  $c_1 \leftarrow V_1$ (the first row of V).
  For $i = 2 \ldots k$ let $c_i \leftarrow \text{argmax}_{j \in \{1 \ldots n\}} \min_{l \in \{1,\ldots,i-1\}} ||V_j - V_{c_l}||_2$.
  For $i = 1 \ldots n$ set $c(i) = \text{argmin}_{j \in \{1 \ldots k\}} ||V_i - V_{c_j}||_2$
output $\mathcal{C} \triangleq \{\{j \in \{1 \ldots n\} : c(j) = i\}\}_{i=1}^{k}$

3 Algorithms and Main Results

In our analysis we study the algorithms for hierarchical and k-way clustering, outlined in Algorithms 1 and 2. Both of these algorithms take a similarity matrix W and compute the eigenvectors corresponding to the smallest eigenvalues of the Laplacian of W. The algorithms then run simple procedures to recover the clustering from the spectral embedding of the data points by these eigenvectors. Our Algorithm 2 deviates slightly from the standard practice of running k-means in the perturbed subspace. We instead use the optimal algorithm for the k-center problem (Hochbaum-Shmoys [6]) because of its amenability to theoretical analysis. We will in this section outline our main results; we sketch the proofs in the next section and defer full proofs to the Appendix.

We first state the following general assumptions, which we place on the *ideal* similarity matrix A:

Assumption 1 *For all i, j, $0 < A_{ij} \leq \beta^*$ for some constant β^*.*

Assumption 2 *(Balanced clusters) There is a constant $\eta \geq 1$ such that at every split of the hierarchy $\frac{|C_{\max}|}{|C_{\min}|} \leq \eta$, where $|C_{\max}|, |C_{\min}|$ are the sizes of the biggest and smallest clusters respectively.*

Assumption 3 *(Range Restriction) For every cluster s, $\min\{\alpha_{s \cdot L}, \alpha_{s \cdot R}\} - \beta_s > \eta(\beta_s - \alpha_s)$.*

It is important to note that these assumptions are placed *only* on the ideal matrices. The noisy HBMs can with high probability violate these assumptions.

We assume that the entries of A are strictly greater than 0 for technical reasons; we believe, as confirmed empirically, that this restriction is not necessary for our results to hold. Assumption 2 says that at every level the largest cluster is only a constant fraction larger than the smallest. This can be relaxed albeit at the cost of a worse rate. For the ideal matrix, the Assumption 3 ensures that at every level of the hierarchy, the gap between the within-cluster similarities and between-cluster similarities is larger than the range of between-cluster similarities. Earlier papers [9, 11] assume that the ideal similarities are constant within a block in which case the assumption is trivially satisfied by the definition of the ideal matrix. However, more generally this assumption is necessary to show that the entries of the eigenvector are safely bounded away from zero. If this assumption is violated by the ideal matrix, then the eigenvector entries can decay as fast as $O(1/n)$ (see Appendix E for more details), and our analysis shows that such matrices will no longer be robust to noise.

Other analyses of spectral clustering often directly make less interpretable assumptions about the spectrum. For instance, Ng et al. [10] assume conditions on the eigengap of the normalized Laplacian and this assumption implicitly creates constraints on the entries of the ideal matrix A that can be hard to make explicit.

To state our theorems concisely we will define an additional quantity $\gamma^*_{\mathcal{S}}$. Intuitively, $\gamma^*_{\mathcal{S}}$ quantifies how close the ideal matrix comes to violating Assumption 3 over a set of clusters $\mathcal{S}$.

Definition 5 *For a set of clusters $\mathcal{S}$, define $\gamma^*_{\mathcal{S}} \triangleq \min_{s\in\mathcal{S}} \min\{\alpha_{s\cdot L}, \alpha_{s\cdot R}\} - \beta_s - \eta(\beta_s - \alpha_s)$.*

We, as well as previous works [10, 11], rely on results from perturbation theory to bound the error in the observed eigenvectors in 2-norm. Using this approach, the straightforward way to analyze the number of errors is pessimistic since it assumes the difference between the two eigenvectors is concentrated on a few entries. However, we show that the perturbation is in fact generated by a random process and thus unlikely to be adversarially concentrated. We formalize this intuition to *uniformly* bound the perturbations on every entry and get a stronger guarantee.

We are now ready to state our main result for hierarchical spectral clustering. At a high level, this result gives conditions on the noise scale factor σ under which Algorithm HS will recover all clusters $s \in \mathcal{S}_m$, where $\mathcal{S}_m$ is the set of all clusters of size at least m.

Theorem 1 *Suppose that $W = A + R$ is an $(n \times n)$ noisy HBM where A satisfies Assumptions 1, 2, and 3. Suppose that the scale factor of R increases at $\sigma = o\left(\min\left(\kappa^{\star 5}\sqrt{\frac{m}{\log n}}, \kappa^{\star 4}\sqrt[4]{\frac{m}{\log n}}\right)\right)$ where $\kappa^\star = \min\left(\alpha_0, \frac{\gamma^\star_{\mathcal{S}m}}{1+\eta}\right)$, $m > 0$ and $m = \omega(\log n)$ [1]. Then for all n large enough, with probability at least $1 - 6/n$,* HS *, on input M, will* exactly *recover all clusters of size at least m.*

A few remarks are in order:

1. It is impossible to resolve the entire hierarchy, since small clusters can be irrecoverably buried in noise. The amount of noise that algorithm HS can tolerate is directly dependent on the size of the smallest cluster we want to resolve.
2. As a consequence of our proof, we show that to resolve only the first level of the hierarchy, the amount of noise we can tolerate is (pessimistically) $o(\kappa^{\star 5}\sqrt[4]{n/\log n})$ which grows rapidly with n.
3. Under this scaling between n and σ, it can be shown that popular agglomerative algorithms such as single linkage will fail with high probability. We verify this negative result through experiments (see Section 5).
4. Since we assume that β^* does not grow with n, both the range $(\beta_s - \alpha_s)$ and the gap $(\min\{\alpha_{s\cdot L}, \alpha_{s\cdot R}\} - \beta_s)$ must decrease with n and hence that $\gamma^*_{\mathcal{S}_m}$ must decrease as well. For example, if we have uniform ranges and gaps across all levels, then $\gamma^*_{\mathcal{S}_m} = \Theta(1/\log n)$. For constant α_0, for n large enough $\kappa^\star = \frac{\gamma^\star_{\mathcal{S}m}}{1+\eta}$. We see that in our analysis $\gamma^\star_{\mathcal{S}m}$ is a crucial determinant of the noise tolerance of spectral clustering.

We extend the intuition behind Theorem 1 to the k-way setting. Some arguments are more subtle since spectral clustering uses the *subspace* spanned by the k smallest eigenvectors of the Laplacian. We improve the results of Ng et. al. [10] to provide a coordinate-wise bound on the perturbation of the subspace, and use this to make precise guarantees for Algorithm K-WAY SPECTRAL.

Theorem 2 *Suppose that $W = A+R$ is an $(n\times n)$* ***noisy k-Block Diagonal*** *matrix where A satisfies Assumptions 1 and 2. Suppose that the scale factor of R increases at rate $\sigma = o(\frac{\beta_0}{k}(\frac{n}{k\log n})^{1/4})$. Then with probability $1 - 8/n$, for all n large enough,* K-WAY SPECTRAL*will exactly recover the k clusters.*

3.1 Information-Theoretic Limits

Having introduced our analysis for spectral clustering a pertinent question remains. *Is the algorithm optimal in its dependence on the various parameters of the problem?*

We establish the minimax rate in the simplest setting of a single binary split and compare it to our own results on spectral clustering. With the necessary machinery in place, the minimax rate for the k-way problem follows easily. We derive lower bounds on the problem of correctly identifying two clusters under the assumption that the clusters are balanced. In particular, we derive conditions on (n, σ, γ), i.e. the number of objects, the noise variance and the gap between inter and intra-cluster similarities, under which *any* method will make an error in identifying the correct clusters.

[1] Recall $a_n = o(b_n)$ and $b_n = \omega(a_n)$ if $lim_{n\to\infty}\frac{a_n}{b_n} = 0$

Theorem 3 *There exists a constant $\alpha \in (0, 1/8)$ such that if, $\sigma \geq \gamma \sqrt{\frac{n}{\alpha \log\left(\frac{n}{2}\right)}}$ the probability of failure of* any *estimator of the clustering remains bounded away from 0 as $n \to \infty$.*

Under the conditions of this Theorem γ and $\kappa^\star$ coincide, provided the inter-cluster similarities remain bounded away from 0 by at least a constant. As a direct consequence of Theorem 1, spectral clustering requires $\sigma \leq \min\left(\gamma^5 \sqrt{\frac{n}{C \log\left(\frac{n}{2}\right)}}, \gamma^4 \sqrt[4]{\frac{n}{C \log\left(\frac{n}{2}\right)}}\right)$ (for a large enough constant C).

Thus, the noise threshold for spectral clustering does not match the lower bound. To establish that this lower bound is indeed tight, we need to demonstrate a (not necessarily computationally efficient) procedure that achieves this rate. We analyze a combinatorial procedure that solves the NP-hard problem of finding the minimum cut of size exactly $n/2$ by searching over all subsets. This algorithm is strongly related to spectral clustering with the combinatorial Laplacian, which solves a *relaxation* of the balanced minimum cut problem. We prove the following theorem in the appendix.

Theorem 4 *There exists a constant C such that if $\sigma < \gamma \sqrt{\frac{n}{C \log\left(\frac{n}{2}\right)}}$ the combinatorial procedure described above succeeds with probability at least $1 - \frac{1}{n}$ which goes to 0 as $n \to \infty$.*

This theorem and the lower bound together establish the minimax rate. It however, remains an open problem to tighten the analysis of spectral clustering in this paper to match this rate. In the Appendix we modify the analysis of [9] to show that under the added restriction of block constant ideal similarities there is an efficient algorithm that achieves the minimax rate.

4 Proof Outlines

Here, we present proof sketches of our main theorems, deferring the details to the Appendix.

Outline of proof of Theorem 1

Let us first restrict our attention toward finding the first split in the hierarchical clustering. Once we prove that we can recover the first split correctly, we can then recursively apply the same arguments along with some delicate union bounds to prove that we will recover all large-enough splits of the hierarchy. To make presentation clearer, we will only focus here on the scaling between σ^2 and n. Of course, when we analyze deeper splits, n becomes the size of the sub-cluster.

Let $W = A + R$ be the $n \times n$ noisy HBM. One can readily verify that the Laplacian of W, L_W, can be decomposed as $L_A + L_R$. Let $v^{(2)}, u^{(2)}$ be the second eigenvector of L_A, L_W respectively.

We first show that the unperturbed $v^{(2)}$ can *clearly* distinguish the two outermost clusters and that λ_1, λ_2, and λ_3 (the first, second, and third smallest eigenvalues of L_W respectively), are far away from each other. More precisely we show $|v_i^{(2)}| = \Theta(\frac{1}{\sqrt{n}})$ for all $i = 1, ..., n$ and its sign corresponds to the cluster identity of point i. Further the eigen-gap, $\lambda_2 - \lambda_1 = \lambda_2 = \Theta(n)$, and $\lambda_3 - \lambda_2 = \Theta(n)$. Now, using the well-known Davis-Kahan perturbation theorem, we can show that

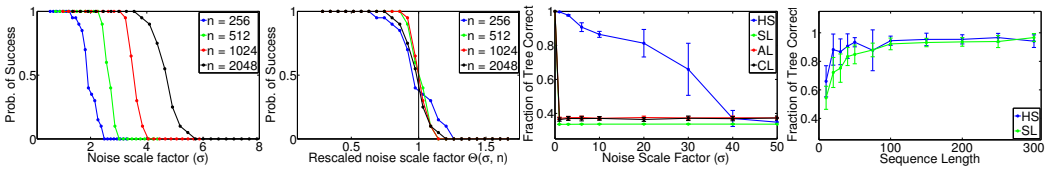
$$||v^{(2)} - u^{(2)}||_2 = O\left(\sigma \frac{\sqrt{n \log n}}{\min(\lambda_2, \lambda_3 - \lambda_2)}\right) = O\left(\sigma \sqrt{\frac{\log n}{n}}\right)$$

The most straightforward way of turning this l_2-norm bound into uniform-entry-wise l_∞ bound is to assume that only one coordinate has large perturbation and comprises all of the l_2-perturbation. We perform a much more careful analysis to show that all coordinates uniformly have low perturbation. Specifically, we show that if $\sigma = O(\sqrt[4]{\frac{\log n}{n}})$, then with high probability, $||v_i^{(2)} - u_i^{(2)}||_\infty = O(\sqrt{\frac{1}{n}})$.

Combining this and the fact that $|v_i^{(2)}| = \Theta(\frac{1}{\sqrt{n}})$, and performing careful comparison with the leading constants, we can conclude that spectral clustering will correctly recover the first split.

Outline of proof of Theorem 2

Leveraging our analysis of Theorem 1 we derive an ℓ_∞ bound on the bottom k-eigenvectors. One potential complication we need to resolve is that the k-Block Diagonal matrix has repeated eigenvalues and more careful *subspace* perturbation arguments are warranted.



(a) (b) (c) (d)

Figure 3: (a),(b): Threshold curves for the first split in HBMs. Comparison of clustering algorithms with $n = 512, m = 9$ (c), and on simulated phylogeny data (d).

We further propose a *different* algorithm, K-WAY SPECTRAL, from the standard k-means. The algorithm carefully chooses cluster centers and then simply assigns each point to its nearest center. The ℓ_∞ bound we derive is much stronger than ℓ_2 bounds prevalent in the literature and in a straightforward way provides a no-error guarantee on K-WAY SPECTRAL.

Outline of proof of Theorem 3

As is typically the case with minimax analysis, we begin by restricting our attention to a small (but hard to distinguish) class of models, and follow this by the application of Fano's inequality. Models are indexed by $\Theta(n, \sigma, \gamma, I_1)$, where I_1 denotes the indices of the rows (and columns) in the first cluster. For simplicity, we'll focus only on models with $|I_1| = n/2$.

Since we are interested in the worst case we can make two further simplifications. The ideal (noiseless) matrix can be taken to be block-constant since the worst case is when the diagonal blocks are at their lower bound (which we call p) and the off diagonal blocks are at their upper bound (q). We consider matrices $W = A + R$, which are $(n \times n)$ matrices, with $R_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Given the true parameter θ_0 we choose the following "hard" subset $\{\theta_1, \ldots, \theta_M\}$. We will select models which mis-cluster only the last object in I_1, there are exactly $n/2$ such models. Our proof is an application of Fano's inequality, using the Hamming distance and the KL-divergence between the true model I_1 and the estimated model $\hat{I}_1$. See the appendix for calculations and proof details.

The proof of Theorem 4 follows from a careful union bound argument to show that even amongst the combinatorially large number of balanced cuts of the graph, the true cut has the lowest weight.

5 Experiments

We evaluate our algorithms and theoretical guarantees on simulated matrices, synthetic phylogenies, and finally on two real biological datasets. Our experiments focus on the effect of noise on spectral clustering in comparison with agglomerative methods such as single, average, and complete linkage.

5.1 Threshold Behavior

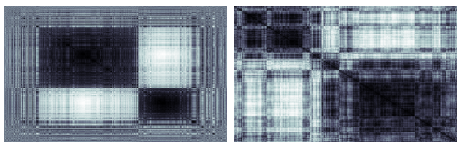
One of our primary interests is to empirically validate the relation between the scale factor σ and the sample size n derived in our theorems. For a range of scale factors and noisy HBMs of varying size, we empirically compute the probability with which spectral clustering recovers the first split of the hierarchy. From the probability of success curves (Figure 3(a)), we can conclude that spectral clustering can tolerate noise that grows with the size of the clusters.

We further verify the dependence between σ and n for recovering the first split. For the first split we observe that when we rescale the x-axis of the curves in Figure 3(a) by $\sqrt{\log(n)/n}$ the curves line up for different n. This shows that empirically, at least for the first split, spectral clustering appears to achieve the minimax rate for the problem.

5.2 Simulations

We compare spectral clustering to several agglomerative methods on two forms of synthetic data: noisy HBMs and simulated phylogenetic data. In these simulations, we exploit knowledge of the true *reference tree* to quantitatively evaluate each algorithm's output as the fraction of triplets of leaves for which the most similar pair in the output tree matches that of the reference tree. One can verify that a tree has a score of 1 if and only if it is identical to the reference tree.

Initially, we explore how HS compares to agglomerative algorithms on large noisy HBMs. In Figure 3(c), we compare performance, as measured by the triplets metric, of four clustering algorithms (HS , and single, average, and complete linkage) with $n = 512$ and $m = 9$. We also evaluate



(a)

Dataset	HS	Agglomerative
Gene (n = 2048)	0.0775	0.0203
Gene (n = 1024)	0.1006	0.0312
Gene (n = 512)	0.0785	0.0280
Phylogeny (l = 100)	0.0067	0.0063
Phylogeny (l = 200)	0.0066	0.0069
Phylogeny (l = 300)	0.0066	0.0060

(b)

Figure 4: Experiments with real world data. (a): Heatmaps of single linkage (left) and HS (right) on gene expression data with $n = 2048$. (b) Δ-entropy scores on real world data sets.

HS and single linkage as applied to reconstructing phylogenetic trees from genetic sequences. In Figure 3(d), we plot accuracy, again measured using the triplets metric, of the two algorithms as a function of sequence length (for sequences generated from the `phyclust` R package [3]), which is inversely correlated with noise (i.e. short sequences amount to noisy similarities). From these experiments, it is clear that HS consistently outperforms agglomerative methods, with tremendous improvements in the high-noise setting where it recovers a significant amount of the tree structure while agglomerative methods do not.

5.3 Real-World Data

We apply hierarchical clustering methods to a yeast gene expression data set and one phylogenetic data set from the PFAM database [5]. To evaluate our methods, we use a Δ-entropy metric defined as follows: Given a permutation π and a similarity matrix W, we compute the rate of decay off of the diagonal as $\hat{s}_d \triangleq \frac{1}{n-d}\sum_{i=1}^{n-d} W_{\pi(i),\pi(i+d)}$, for $d \in \{1, ..., n-1\}$. Next, we compute the entropy $\hat{E}(\pi) \triangleq -\sum_{i=1}^{n-1} \hat{p}_\pi(i) \log \hat{p}_\pi(i)$ where $\hat{p}_\pi(i) \triangleq (\sum_{d=1}^{n} \hat{s}_d)^{-1}\hat{s}_i$. Finally, we compute Δ-entropy as $\hat{E}_\Delta(\pi) = \hat{E}(\pi_{random}) - \hat{E}(\pi)$. A good clustering will have a large amount of the probability mass concentrated at a few of the $\hat{p}_\pi(i)$s, thus yielding a high $\hat{E}_\Delta(\pi)$. On the other hand, poor clusterings will specify a more uniform distribution and will have lower Δ-entropy.

We first compare HS to single linkage on yeast gene expression data from DeRisi et al [4]. This dataset consists of 7 expression profiles, which we use to generate Pearson correlations that we use as similarities. We sampled gene subsets of size $n = 512, 1024$, and 2048 and ran both algorithms on the reduced similarity matrix. We report Δ-entropy scores in Table 4(b). These scores quantitatively demonstrate that HS outperfoms single linkage and additionally, we believe the clustering produced by HS (Figure 4(a)) is qualitatively better than that of single linkage.

Finally, we run HS on real phylogeny data, specifically, a subset of the PDZ domain (PFAM Id: PF00595). We consider this family because it is a highly-studied domain of evolutionarily well-represented protein binding motifs. Using alignments of varying length, we generated similarity matrices and computed Δ-entropy of clusterings produced by both HS and Single Linkage. The results for three sequence lengths (Table 4(b)) show that HS and Single Linkage are comparable.

6 Discussion

In this paper we have presented a new analysis of spectral clustering in the presence of noise and established tight information theoretic upper and lower bounds. As our analysis of spectral clustering does not show that it is minimax-optimal it remains an open problem to further tighten, or establish the tightness of, our analysis, and to find a computationally efficient minimax procedure in the general case when similarities are not block constant. Identifying conditions under which one can guarantee correctness for other forms of spectral clustering is another interesting direction. Finally, our results apply only for binary hierarchical clusterings, yet k-way hierarchies are common in practice. A future challenge is to extend our results to k-way hierarchies.

7 Acknowledgements

This research is supported in part by AFOSR under grant FA9550-10-1-0382 and NSF under grant IIS-1116458. AK is supported in part by a NSF Graduate Research Fellowship. SB would like to thank Jaime Carbonell and Srivatsan Narayanan for several fruitful discussions.

References

[1] Dimitris Achlioptas and Frank Mcsherry. On spectral learning of mixtures of distributions. In *Computational Learning Theory*, pages 458–469, 2005.

[2] S. Charles Brubaker and Santosh Vempala. Isotropic pca and affine-invariant clustering. In *FOCS*, pages 551–560, 2008.

[3] Wei-Chen Chen. *Phylogenetic Clustering with R package phyclust*, 2010.

[4] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, 278(5338):680–686, 1997.

[5] Robert D. Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, Prasad Gunesekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L. Sonnhammer, Sean R. Eddy, and Alex Bateman. The Pfam Protein Families Database. *Nucleic Acids Research*, 2010.

[6] Dorit S. Hochbaum and David B. Shmoys. A Best Possible Heuristic for the K-Center Problem. *Mathematics of Operations Research*, 10:180–184, 1985.

[7] Ling Huang, Donghui Yan, Michael I. Jordan, and Nina Taft. Spectral Clustering with Perturbed Data. In *Advances in Neural Inforation Processing Systems*, 2009.

[8] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *18th Annual Conference on Learning Theory (COLT*, pages 444–457, 2005.

[9] Frank McSherry. Spectral partitioning of random graphs. In *IEEE Symposium on Foundations of Computer Science*, page 529, 2001.

[10] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.

[11] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral Clustering and the High-Dimensional Stochastic Block Model. *Technical Report 791, Statistics Department, UC Berkeley*, 2010.

[12] Sajama and Alon Orlitsky. Estimating and Computing Density Based Distance Metrics. In *ICML05, 22nd International Conference on Machine Learning*, 2005.

[13] Dan Spielman. *Lecture Notes on Spectral Graph Theory*, 2009.

[14] Terence Tao. Course notes on random Matrix Theory, 2010.

[15] Alexandre B. Tsybakov. *Introduction a lÉstimation Non-paramÃl'trique*. Springer, 2004.

[16] Ulrike von Luxburg. A Tutorial on Spectral Clustering. Technical Report 149, Max Planck Institute for Biological Cybernetics, August 2006.

[17] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of Spectral Clustering. In *The Annals of Statistics*, pages 857–864. MIT Press, 2004.

A Details of Experiments

The Δ-entropy metric is defined as follows: Given a permutation π and a similarity matrix W, we compute the rate of decay off of the diagonal as $\hat{s}_d \triangleq \frac{1}{n-d}\sum_{i=1}^{n-d} W_{\pi(i),\pi(i+d)}$, for $d \in \{1, ..., n-1\}$. Next, we compute the entropy $\hat{E}(\pi) \triangleq -\sum_{i=1}^{n-1} \hat{p}_\pi(i) \log \hat{p}_\pi(i)$ where $\hat{p}_\pi(i) \triangleq (\sum_{d=1}^{n} \hat{s}_d)^{-1} \hat{s}_i$. Finally, we compute Δ-entropy as $\hat{E}_\Delta(\pi) = \hat{E}(\pi_{random}) - \hat{E}(\pi)$. The distribution defined by the $\hat{p}_\pi$s of a good clustering have a large amount of the probability mass concentrated at a few of the $\hat{p}_\pi(i)$s, and thus it will have a high $\hat{E}_\Delta(\pi)$. On the other hand, poor clusterings will specify a more uniform distribution and will have lower Δ-entropy.

B Proof of Theorem 1

To ease the presentation of our proof, we compartmentalize into several sections outlined as follows:

1. Noiseless Spectral Clustering: We show that Algorithm HS will perfectly cluster a noiseless Hierarchical Block Matrix (HBM).
2. Derive spectral properties of noiseless matrices: We study the spectral properties of a related matrix, the Constant Block Matrix (CBM), and use it to understand the spectral properties of the HBM. This analysis is entirely deterministic.
3. Bound spectral norm of noise matrices: We analyze the noise matrices and show that, with high probability, they have small spectral norm uniformly across all levels of the hierarchy.
4. Davis-Kahan For Laplacians: We next derive a variant of the well-known Davis-Kahan sin theorem that we can apply to Laplacians. This allows us to bound the ℓ_2-norm deviation between the eigenvectors of the HBM and the noisy HBM in terms of the spectral norm of the noise matrices.
5. ℓ_∞-norm deviation bounds: We observe that due to the independence and randomness of the noise, it is unlikely that the eigenvector of the noisy HBM is spiked in just one or a few coordinates. We formalize this notion by deriving ℓ_∞-norm deviation bounds between the eigenvectors of the HBM and the noisy HBM.
6. Final steps: we conclude that for sufficiently large n, every entry of the second eigenvectors (across all calls to Algorithm HS) correctly clusters the data.

B.1 Noiseless Spectral Clustering

We first show that in the absence of noise, Algorithm HS will correctly cluster the data.

Theorem 5 *Given an ideal noiseless Hierarchical Block Matrix W (i.e. $R = 0$) satisfying Assumption 1,* HS *will recover the true hierarchical clustering.*

Note that this theorem would not hold if Algorithm HS used either the normalized Laplacian or the similarity matrix directly. In fact, in Appendix E, we show several examples that demonstrate the shortcomings of these approaches. In addition, note that we do not require Assumptions 2 and 3 for Theorem 5.

Our proof strategy is to first show that HS will correctly output the first split the hierarchical clustering in Lemma 6. Repeated application of this lemma concludes the proof. Recall that the ideal matrix has within cluster similarity greater than all between cluster similarities; this motivates the statement of Lemma 6.

Lemma 6 *Let W be a $(p+q) \times (p+q)$ matrix with the* Large-Small *block structure of* $\left(\begin{array}{c|c} W_L & W_S \\ \hline W_S^\intercal & W'_L \end{array}\right)$ *such that W_L is a $p \times p$ block, W'_L is a $q \times q$ block and*

$$\min_{1 \le i,j \le p} (W_L)_{ij} > \max_{1 \le i \le p < j \le p+q} (W_S)_{ij} > 0$$

$$\min_{p+1\le i,j\le p+q}(W'_L)_{ij} > \max_{1\le i\le p<j\le p+q}(W_S)_{ij} > 0$$

Let D be the diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. Let v be the smallest non-constant eigenvector of the graph-Laplacian $L = D - W$, then v has either the sign pattern of $\begin{pmatrix} v_+ \\ v_- \end{pmatrix}$ where v_+, the first p elements of v, are strictly positive and v_-, the other q elements of v, are strictly negative or the reverse sign pattern.

Proof (of Lemma 6)
Step 1: First, we will show that if a $(p+q)\times(p+q)$ symmetric matrix B has the *Positive-Negative* block structure of $\left(\begin{array}{c|c} B_+ & B_- \\ \hline B_-^\mathsf{T} & B'_+ \end{array}\right)$, where every *non-diagonal* element in the $p\times p$ block B_+ and the $q\times q$ block B'_+ is strictly positive and every element in the $p\times q$ block B_- is strictly negative, then the first eigenvector of B, call it v, either has the sign pattern of $\begin{pmatrix} v_+ \\ v_- \end{pmatrix}$ where v_+, the first p elements of v, are strictly positive and v_-, other q elements of v, are strictly negative or has the reverse sign pattern.

Let $v = \begin{pmatrix} v_+ \\ v_- \end{pmatrix}$ be the largest eigenvector of B where v_+ are the first p elements and v_- are the other q elements. Let I_+, I_- be index sets of positive and negative elements in v_+, and I the index of all elements in v_+. Let J_+, J_- be index sets of positive and negative elements in v_-, and J the index of all elements in v_-. Then

$$v^\mathsf{T}Bv = \underbrace{v_+^\mathsf{T}B_+v_+}_{\text{term 1}} + \underbrace{v_+^\mathsf{T}B_-v_-}_{\text{term 2}} + \underbrace{v_-^\mathsf{T}B_-^\mathsf{T}v_+}_{\text{term 3}} + \underbrace{v_-^\mathsf{T}B'_+v_-}_{\text{term 4}}$$

Let us form a new vector w by changing the signs of all elements in I_- and all elements in J_+. We now proceed to compare $w^\mathsf{T}Bw$ with $v^\mathsf{T}Bv$ term by term, noting that $||w||_2 = ||v||_2 = 1$

Term 1 is $v_+^\mathsf{T}B_+v_+ = \sum_{i,j\in I} v_iB_{ij}v_j$. Since $B_{ij} > 0$, $w_iB_{ij}w_j \ge v_iB_{ij}v_j$ for all i,j, we notice that we have strictly increased term 1, provided that I_-, I_+ are non-empty. An analogous argument reveals that we do not decrease term 4 by changing v to w. Furthermore, we strictly increased term 4 if J_-, J_+ are non-empty.

Term 2 is $v_+^\mathsf{T}B_-v_-^\mathsf{T} = \sum_{i\in I, j\in J} v_iB_{ij}v_j$. Since $B_{ij} < 0$, we see that $w_iB_{ij}w_j = -|v_i|B_{ij}|v_j| \ge v_iB_{ij}v_j$ for all i,j with strict inequality whenever $i\in I_-, j\in J_-$ or $i\in I_+, j\in J_+$. Thus we have strictly increased term 2 (and 3 by analogous argument) provided that the index sets are non-empty.

We see then that unless I_-, J_+ are empty or I_+, J_- are empty, $w^\mathsf{T}Bw > v^\mathsf{T}Bv$. However, v is assumed to be largest eigenvector and hence maximize $v^\mathsf{T}Bv$ among all unit-norm vectors. We reached a contradiction and thus, all of v_+ must have same sign and be opposite of v_-.

Now suppose $v_i = 0$, then $B_i^\mathsf{T}v = 0$ where B_i is the $i-th$ row of B. However, since v cannot be all zero, we see then that $B_i^\mathsf{T}v > 0$. Thus, v_i cannot be zero for all i and v_+ is all positive and v_- is all negative.

Step 2: Now we prove the claim of the theorem. Let $\underline{1}$ be a vector of all ones. Since the W satisfy the Large-Small block structure there exist $c\in\mathbb{R}$ such that the matrix $B \triangleq c\underline{1}\underline{1}^\mathsf{T} - L = c\underline{1}\underline{1}^\mathsf{T} - D + W$ has the *Positive-Negative* block structure of $\left(\begin{array}{c|c} B_+ & B_- \\ \hline B_-^\mathsf{T} & B'_+ \end{array}\right)$ except on the diagonals.

Let $\{v^{(i)}\}$ be the eigenvectors of L with corresponding eigenvalue $\{\lambda_i\}$. Since we know that $\underline{1}$ is an un-normalized eigenvector of L with eigenvalue 0, let $v^{(1)} = \underline{1}$ and $\lambda_1 = 0$. All other eigenvectors of L must be orthogonal to $\underline{1}$ and hence, $\{v^{(i)}\}$ are also eigenvectors of B. Furthermore, for B, $\{v^{(i)}\}$ have the corresponding eigenvalues of $\{-\lambda_i\}$ except for $\{v^{(1)}\}$, which has the eigenvalue of $\{c\}$.

We know thus that the v, the largest eigenvector of B, is also the smallest non-constant eigenvector of L. By step 1, we know that v has the sign pattern of $v = (v_+v_-)^\mathsf{T}$. □

B.2 Spectral Properties of Noiseless Matrices

As will become evident later, it will also be important to establish bounds on certain spectral quantities of noiseless HBMs. We use several results from spectral graph theory to obtain these bounds in this section. To derive these bounds, we first must study a more structured matrix, which we call the Constant Block Matrix (CBM). The CBM has the same cluster structure as the HBM only it has constant off-block-diagonal similarities rather than ranges as with the HBM.

Definition 6 *A similarity matrix A is a **Constant Block Matrix** if A is an ideal matrix with $\epsilon_s \triangleq \alpha_s = \beta_s$ for all clusters s.*

Lemma 7 *(Spectrum of CBM) Consider an $(n \times n)$ Constant-Block Matrix A characterized by an ϵ_s for each level s, with $\min\{\epsilon_{s\cdot L}, \epsilon_{s\cdot R}\} > \epsilon_s$ and with balance factor η. Then the laplacian L_A has the following eigenvalues ($\lambda_1 \leq \lambda_2, \leq \ldots \leq \lambda_n$) and eigenvectors ($v_1, \ldots v_n$):*

1. $v^{(1)} = \frac{1}{\sqrt{n}}\underline{1}$ *with* $\lambda_1 = 0$.
2. $\sqrt{\frac{1}{n\eta}} \leq |v^{(2)}(i)| \leq \sqrt{\frac{\eta}{n}}$ *with* $\lambda_2 = n\epsilon_0$.
3. $\frac{n}{1+\eta}(\eta\epsilon_0 + \min\{\epsilon_L, \epsilon_R\}) \leq \lambda_3 \leq \frac{n}{1+\eta}(\epsilon_0 + \eta\max\{\epsilon_L, \epsilon_R\})$.

Proof (of Lemma 7) The first claim is true simply because L_A is a Laplacian Matrix.

We prove the remaining claims by induction on number of levels l in the Constant-Block Matrix A. Suppose that A is an $n \times n$ constant matrix of only one level, that is, $A_{ij} = \epsilon_0$ for all i, j. It is then easy to verify that every vector orthogonal to $\underline{1}$ is an eigenvector of L_A with eigenvalue of $n\epsilon_0$.

Suppose now that A is an $n \times n$ CBM with entries ϵ_s as defined in the lemma. Let $|C_L|, |C_R|$ be the sizes of the two first-level clusters and let $L = \{1, ..., |C_L|\}$ and $R = \{|C_L| + 1, ..., |C_L| + |C_R|\}$ be the index sets of the two first-level clusters.

Let v be a vector such that $v(i) = \sqrt{\frac{|C_R|}{n|C_L|}}$ for $i \in L$ and $v(i) = -\sqrt{\frac{|C_L|}{n|C_R|}}$ for $i \in R$. Then it is easy to verify that v is an eigenvector of L_A with eigenvalue $n\epsilon_0$. We will show that $n\epsilon_0$ is the second smallest eigenvalue.

Note the upper left block A_{LL} and the lower right block A_{RR} are Hierarchical Constant-Block matrices. Hence, the upper left block of the Laplacian is $(L_A)_{LL} = L_{(A_{LL})} + |C_R|\epsilon_0 I$ and the lower right block of the Laplacian is $(L_A)_{RR} = L_{(A_{RR})} + |C_L|\epsilon_0 I$.

From the inductive hypothesis, the second smallest eigenvalue of $L_{(H_{LL})}$ is $|C_L|\epsilon_L$. Let v_{C_L} be its corresponding eigenvector of $L_{(A_{LL})}$, we can see that by filling all additional coordinates of v_{C_L} with zero, v_{C_L} becomes an eigenvector of L_A with eigenvalue $|C_L|\epsilon_L + |C_R|\epsilon_0 > n\epsilon_0$. Hence, we know that at least $|C_L| - 1$ eigenvalues of L_A are larger than $n\epsilon_0$. Apply the same argument to $L_{(A_{RR})}$ and we see that at least $n-2$ eigenvalues of L_A are larger than $n\epsilon_0$. Since 0 is an eigenvalue of L_A, we conclude that $n\epsilon_0$ is the second smallest eigenvalue of L_A.

Since $\frac{1}{\eta} \leq \frac{|C_R|}{|C_L|} \leq \eta$, we have proved claim 2. Note that in proving claim 2, we have also shown that the third smallest eigenvalue of L_A is $\min(|C_L|\epsilon_L + |C_R|\epsilon_0, |C_R|\epsilon_R + |C_L|\epsilon_0)$. Apply the definition of η and we see that the third claim holds true as well. □

Equipped with these properties of the CBM, we now turn our attention to the more general Hierarchical Block Matrix. The following result extends Lemma 7 to the more general HBM and shows that under Assumption 3, the spectrum is well behaved, i.e. the eigenvalues and eigenvector elements vary with n at the same rate as those of the CBM.

Lemma 8 *(Spectrum of HBMs) Consider an $(n \times n)$ ideal Hierarchical Block Matrix $A = \begin{pmatrix} A_L & A_S \\ A_S^\intercal & A'_L \end{pmatrix}$ such that all values in off-diagonal blocks A_S are in $[\alpha_0, \beta_0]$ and all values in the diagonal blocks A_L, A'_L are in $[\alpha_1, \beta^*]$ (here we take $\alpha_1 = \min\{\alpha_L, \alpha_R\}$).*

Suppose A satisfies Assumptions 1 and 2 with balance factor η. Suppose also that A satisfies Assumption 3. Then:

1. *Let $\lambda_1, \lambda_2, \lambda_3$ be the first, second and third smallest eigenvalue of L_A respectively ($\lambda_1 = 0$), then the eigengap $\delta \triangleq \min(|\lambda_2 - \lambda_1|, |\lambda_3 - \lambda_2|) \geq \min(n\alpha_0, \frac{n}{\eta+1}(\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0) = \Theta(n)$*
2. *Let $v^{(2)}$ be the second eigenvector of L_A, then every entry of $v^{(2)}$ satisfies $\sqrt{\frac{1}{K_\eta n}} \leq |v^{(2)}(i)| \leq \sqrt{\frac{K_\eta}{n}}$ where*

$$K_\eta = \left(\frac{(\beta^* - \alpha_0)}{(\alpha_1 - \beta_0)} \frac{\beta_0 - \alpha_0 + \eta(\beta^* - \alpha_0)}{\alpha_1 - \beta_0 - \eta(\beta_0 - \alpha_0)} \right)^2$$

Note that once we prove this Lemma, we can recursively apply it on sub-matrices that represent the similarity matrix of sub-clusters to characterize the eigenvectors and eigenvalues at every split of the hierarchical clustering. One complication with recursively applying Lemma 8 is that at different level i, we would get a different K_η. To succinctly present the final rates, we define K_η^* as the maximum over all K_η for all levels i:

$$K_\eta^* = \max_{s \in \mathcal{S}_m} \left(\frac{(\beta^* - \alpha_s)}{(\min\{\alpha_{s\cdot L}, \alpha_{s\cdot R}\} - \beta_s)} \frac{\beta_s - \alpha_s + \eta(\beta^* - \alpha_s)}{\min\{\alpha_{s\cdot L}, \alpha_{s\cdot R}\} - \beta_s - \eta(\beta_s - \alpha_s)} \right)^2$$

where β^* is the largest entry in the entire ideal HBM A and L is the number of levels we recover.

Our proof will construct two ideal Constant-Block Matrices, show that eigenvalues and eigenvectors of the HBM A are constrained by the two CBMs, and then leverage Lemma 7 to get the final result. Before we proceed to the proof, we state two well-known results in Spectral Graph Theory that we will use:

Lemma 9 *[13] If L_G and L_H are two graph Laplacians such that $L_G \succeq cL_H$, then $\lambda_k(G) \geq c\lambda_k(H)$. (where we say PSD matrices $A \succeq B$ if $A - B \succeq 0$)*

Lemma 10 *[13] Let $G = (V, E, w)$ and $H = (V, E, z)$ be two graphs that differ only in edge weights. Then $L_G \succeq \min_{e \in E} \frac{w(e)}{z(e)} L_H$.*

Proof (of Lemma 8): Let H_α be a two level ideal Constant-Block matrix with the same block structure as A. Let all entries of the diagonal blocks of H_α have value $\alpha_1 \triangleq \min\{\alpha_L, \alpha_R\}$ and let all entries of the off-diagonal blocks of H_α have value α_0. Define another constant-block matrix H_β similarly, the diagonal blocks are β^* while the off-diagonal blocks are β^0.

Lemma 7 characterizes the spectrum of H_α and H_β. Using this characterization, along with Lemmas 9 and 10, we have that $n\alpha_0 \leq \lambda_2(L_A) \leq n\beta_0$ and that $\frac{n}{1+\eta}(\eta\alpha_0 + \alpha_1) \leq \lambda_3(L_A) \leq \frac{n}{1+\eta}(\beta_0 + \eta\beta^*)$.

Combined with the fact that $\lambda_1 = 0$ for any Laplacian, we get that $\delta \geq \min(n\alpha_0, \frac{n}{\eta+1}(\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0))$. Under Range Restriction Assumption 3, we see that $(\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0) > 0$ and hence $\delta = \Theta(n)$.

To establish bounds on entries of $v^{(2)}$, we consider a single coordinate of $v^{(2)}$; using the definition of eigenvector we get that

$$v^{(2)}(i) = \frac{A_i v^{(2)}}{d_i - \lambda_2}$$

Where A_i is the i-th row of A. From theorem 5, we can assume without loss of generality that $v^{(2)}(i)$ is all strictly positive for one cluster and strictly negative for other. From the fact that $\underline{1}$ is an

eigenvector of L_A, we get that $\sum_{i:v^{(2)}(i)>0} |v^{(2)}(i)| = \sum_{i:v^{(2)}(i)<0} |v^{(2)}(i)|$. Hence:

$$J(\alpha_1 - \beta_0) \leq A_i v^{(2)} \leq J(\beta^* - \alpha_0)$$

where $J = \frac{1}{2}\sum_i |v^{(2)}(i)|$. We can similarly derive an upper and lower bound for $d_i - \lambda_2$:

$$\begin{aligned} n\frac{1}{1+\eta}\alpha_1 + n\frac{\eta}{1+\eta}\alpha_0 - n\beta_0 \\ \leq d_i - \lambda_2 \leq n\frac{1}{1+\eta}\beta_0 + n\frac{\eta}{1+\eta}\beta^* - n\alpha_0 \end{aligned}$$

Note that with the Range Restriction, the lower bound of $d_i - \lambda_i$ is positive and is $\Theta(n)$. Combining these two results, we get

$$\begin{aligned} \frac{Jc_1}{n} &\leq |v^{(2)}(i)| \leq \frac{Jc_2}{n} \\ c_1 &= \frac{(\alpha_1 - \beta_0)(\eta+1)}{\beta_0 + \eta\beta^* - (1+\eta)\alpha_0} \\ c_2 &= \frac{(\beta^* - \alpha_0)(\eta+1)}{\alpha_1 + \eta\alpha_0 - (1+\eta)\beta_0} \end{aligned}$$

Since $v^{(2)}$ must be a unit vector, we can bound J and get that

$$\frac{c_1}{c_2}\frac{1}{\sqrt{n}} \leq |v^{(2)}(i)| \leq \frac{c_2}{c_1}\frac{1}{\sqrt{n}}$$

Set $K_\eta = (\frac{c_2}{c_1})^2$ and we get the desired result. □

B.3 Bounds on the Noise

We now analyze the noise matrices. We begin by stating results bounding subgaussian random variables, then we turn our attention to bounding the spectral norm of the random matrices.

Lemma 11 *(Max of Subgaussian) Let $X_1, ...X_n$ be identically distributed subgaussian random variables with scale σ. With probability $1 - \delta$*

$$\max_{i=1,...n} |X_i| \leq \sigma\sqrt{2\log n + 2\log\frac{2}{\delta}}$$

Proof It is well known that for a single subgaussian random variable X with scale factor σ,

$$\mathbb{P}(|X| \geq x) \leq 2\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Thus, by union bound,

$$\mathbb{P}\left(\max_i |X_i| \geq \sigma\sqrt{2\log n + 2\log\frac{2}{\delta}}\right) \leq n\mathbb{P}\left(|X_i| \geq \sigma\sqrt{2\log n + 2\log\frac{2}{\delta}}\right) \leq \delta$$

□

Lemma 12 *(Sums of Subgaussians) Suppose $X_1, \ldots, X_n$ are independent subgaussian random variables, each with $\mathbb{E}(e^{tX_i}) \leq e^{\frac{\sigma_i^2 t^2}{2}}$. For any scalars $a_1, \ldots, a_n$ independent of $X_1, \ldots, X_n$ we have, $\sum_{i=1}^n a_i X_i$ is a subgaussian random variable with $\mathbb{E}(e^{t\sum_{i=1}^n a_i X_i}) \leq e^{\frac{t^2\sum_{i=1}^n a_i^2\sigma_i^2}{2}}$.*

Proof The proof is by simple mathematical expansions. We note that $\mathbb{E}(e^{(t\sum_{i=1}^{n} a_i X_i)}) = \prod_{i=1}^{n} \mathbb{E}(e^{ta_i X_i}) \leq \prod_{i=1}^{n} e^{\frac{t^2 a_i^2 \sigma_i^2}{2}}$ and multiplying the exponentials gives the result. □

Now we turn to the study of the random matrices. We start by stating a well known result from Random Matrix Theory:

Lemma 13 *[14] The operator norm of R is $O(\sqrt{n})$ and is concentrated as*

$$P(||R||_2 \geq A\sigma\sqrt{n}) \leq C\exp(-cAn)$$

for absolute constants c, C and for all $A \geq C$.

The matrix we will ultimately have to bound is L_R, we derive this bound next:

Lemma 14 *(Noise-Laplacian) Let R be a perturbation matrix, let $L_R = D_R - R$. For all $n \geq n_0$, we have that with probability at least $1 - 4/n$,*

$$||L_R||_2 \leq 4\sigma\sqrt{n\log n}$$

where n_0 is an absolute constant.

Proof

$$||L_R||_2 = ||D_R - R||_2 \leq ||D_R||_2 + ||R||_2$$

D_R is diagonal and $||D_R||_2$ is the largest (in absolute value) diagonal element. Since every diagonal element of D_R is subgaussian with scale factor $\leq \sqrt{n}\sigma$, we can apply Lemma 11 and get that $||D_R||_2 \leq \sigma\sqrt{n}\sqrt{2\log n + 2\log\frac{4}{\delta}}$ with probability at least $1 - \delta/2$. Setting $\delta = 4/n$ we have $||D_R||_2 \leq 2\sigma\sqrt{n\log n}$.

Using Lemma 13, we know that with probability $1 - 8/n$, for n large enough (depending on the absolute constants c and C), $||R||_2 = C\sigma\sqrt{n}$. Hence, for n large enough, $||D_R||_2 \geq ||R||_2$ and $||L_R||_2 \leq 2||D_R||_2$ and we get the desired result. □

In order to guarantee recovery of all clusters of size at least m, it is not sufficient to bound $||L_R||$ at just the top-most level of the hierarchy. We must ensure that the noise matrices for all of the subclusters we hope to recover have uniformly bounded spectral norm (where the specific bound could be different for different submatrices). The following lemmas establish the desired uniform bound.

Before we present the lemmas, we specify our notation. For each level $l \in \{0, \ldots, n\}$ in the hierarchy, denote the set of clusters at level l by $\{C_{li} : i \in \{1 \ldots, 2^l\}\}$ and let $m_{li} = |C_{li}|$. For any subcluster C_{li} we write the corresponding noise degree matrix as $D_R^{C_{li}}$ and the corresponding noise matrix as $R^{C_{li}}$.

Lemma 15 *(Hierarchical Noise Degree Bound) Let R be the noise matrix associated with a $n \times n$ noisy Hierarchical Block Matrix satisfying Assumptions 1 and 2. Then with probability $1 - 2/n$, for all sub-clusters C_{li} in the true hierarchical clustering, the corresonding noise degree matrix $D_R^{C_{li}}$ will have operator norm bounded by*

$$||D_R^{C_{li}}||_2 \leq \sigma\sqrt{6m_{li}\log n}$$

Proof We first bound the number of levels in the tree. l is bounded by $\log n$ in the balanced binary case, but bounded by n in the worst case irrespective of η.

Now, at each level we bound at most n random draws from various sub-Gaussians. For instance, consider the first level. We need to bound the operator norm of a diagonal degree matrix, and each diagonal entry is a draw from a sub-Gaussian with scale factor at most $\sqrt{n}\sigma$, and there are at most n diagonal entries. On the second level we will have two matrices but still n degree random variables we will need to bound. Over l levels there are at most nl random variables to bound.

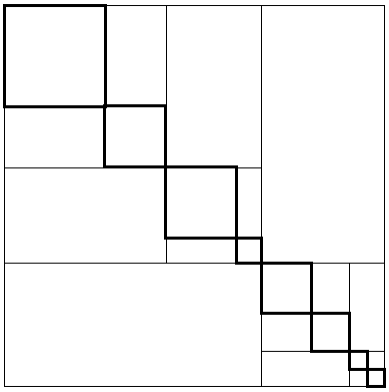


Figure 5: All sub-matrices corresponding to sub-clusters at level 3

Let $0 < \delta < 1$ and let $\Gamma_{n,\delta} \triangleq \sqrt{2\log n + \log\frac{2}{\delta}}$, then we have, by union bound:

$$\mathbb{P}(\exists \text{ cluster } C_{li}, ||D_R^{C_{li}}||_2 \geq \sigma\sqrt{2m_{li}}\Gamma_{n,\delta}) \leq \sum_{l=1}^{n} \mathbb{P}(\exists \text{ cluster } C_{li} \text{ at level } l, ||D_R^{C_{li}}||_2 \geq \sigma\sqrt{2m_{li}}\Gamma_{n,\delta})$$

We emphasize again that $m_{li} = |C_{li}|$ and is dependent on C_{li}.

Note that at any fixed level l, we can bound the spectral norm of all diagonal matrices $D_R^{C_{li}}$ by bounding all the diagonal elements of all the matrices $D_R^{C_{li}}$. The collection of all diagonal elements is just a collection of n sub-Gaussian random variables:

$$\{(D_R^{C_{li}})_{jj} : i \in \{1, \ldots 2^l\}, j \in C_{li}\}$$

where each $(D_R^{C_{li}})_{jj}$ is the jth diagonal entry of $D_R^{C_{li}}$ and hence it is sum of m_{li} entries from R_{li} and has scale factor at most $\sqrt{m_{li}}\sigma$. Then:

$$\begin{aligned}
&\mathbb{P}(\exists \text{ cluster } C_{li}, ||D_R^{C_{li}}||_2 \geq \sigma\sqrt{2m_{li}}\Gamma_{n,\delta}) \\
&\leq \sum_{l=1}^{n}\sum_{i=1}^{2^l}\sum_{j\in C_{li}} \mathbb{P}(|(D_R^{C_{li}})_{jj}| \geq \sigma\sqrt{2m_{li}}\Gamma_{n,\delta}) \\
&\leq \sum_{l=1}^{n}\sum_{i=1}^{2^l}\sum_{j\in C_{li}} 2\exp\left(-\frac{2m_{li}\sigma^2\Gamma_{n,\delta}^2}{2m_{li}\sigma^2}\right) \\
&= \sum_{l=1}^{n}\sum_{i=1}^{2^l}\sum_{j\in C_{li}} \frac{1}{n}\left(\frac{1}{n}\right)\delta \\
&\leq \sum_{l=1}^{n}\left(\frac{1}{n}\right)\delta \\
&\leq \delta
\end{aligned}$$

where the second inequality follows from the fact that if X is a subgaussian random variable with scale factor σ, then $\mathbb{P}(|X| \geq x) \leq 2\exp(-\frac{x^2}{2\sigma^2})$, the third and four inequality follow from the union bound and from the fact for a fixed level, there are at most n subgaussian random variables. Plugging in the desired value for δ we arrive at the result. □

Lemma 16 *(Hierarchical Laplacian Spectral Bound) Let R be the noise matrix associated with an $n \times n$ noisy Hierarchical Block Matrix satisfying Assumptions 1 and 2.*

Then with probability $1 - 4/n$, for all large enough sub-clusters C_{li} (with size $m_{li} = \omega(\log n)$) in the true hierarchical clustering, the corresponding noise Laplacian matrix $L_R^{C_{li}}$ will have operator norm bounded by

$$||L_R^{C_{li}}||_2 \leq 2\sigma\sqrt{6m_{li}\log n}$$

for all $n \geq n_0(\eta)$, where $n_0(\eta)$ is a constant depending on η.

Proof We will argue that if n is large enough and that $m = \omega(\log n)$, then for every sub-cluster C_{li} with $m_{li} \geq m$, $||D_R^{C_{li}}||_2$ will be larger than spectral norm of just the noise sub-matrix $||R^{C_{li}}||_2$.

Let us now bound the number of levels in the tree. We will need to be more careful than in Lemma 15 where bounding l by n did not affect the rate. When the clusters are imbalanced with a balance factor η we have

$$l \leq \frac{1}{\log(\frac{1+\eta}{\eta})}\log n = C_\eta \log n$$

with $C_\eta = \frac{1}{\log(1+1/\eta)}$. To see this note that at each split the larger cluster is of size at most $\frac{\eta}{1+\eta}n$. After l levels the cluster size is at most 1, i.e.

$$\left(\frac{\eta}{1+\eta}\right)^l n = 1$$

We can solve this to obtain that $l \leq C_\eta \log n$.

Returning to the proof, we note that we need to bound the norm of at most $2^{l+1} - 2 \leq e^{2l}$, sub-gaussian matrices of varying sizes.

From Lemma 13 we know also that for each C_{li}, $||R^{C_{li}}||_2 \leq B_{li}\sigma\sqrt{m_{li}}$ holds with probability at least $\exp(-cB_{li}m_{li})$, where $B_{li} \geq C$ for some absolute constant C.

By letting $B_{li} = \max(\frac{2C_\eta \log n + \log\frac{2}{\delta}}{cm_{li}}, C)$, we can take union bound over all $2^{l+1} - 2$ noise sub-matrices and get that with probability at least $1 - \frac{\delta}{2}$, for all sub-clusters C_{li}, $||R^{C_{li}}||_2 \leq \max\left(\sigma\frac{2C_\eta \log n + \log\frac{2}{\delta}}{c\sqrt{m_{li}}}, C\sigma\sqrt{m_{li}}\right)$. Taking $\delta = 4/n$ we have $||R_{li}||_2 \leq \max\left(\sigma\frac{3C_\eta \log n}{c\sqrt{m_{li}}}, C\sigma\sqrt{m_{li}}\right)$.

Now, for $m_{li} = \omega(\log n)$ the second term dominates and we have for $n \geq n_0(\eta)$, $||R^{C_{li}}||_2 \leq C\sigma\sqrt{m_{li}}$ where $n_0(\eta)$ is a constant depending on η.

From Lemma 15, we know that with probability $1 - 2/n$, for every C_{li}, $||D_R^{C_{li}}||_2 \leq \sigma\sqrt{6m_{li}\log n}$.

As before we see that if $m = \omega(\log n)$, with probability at least 1 - 4/n, $||D_R^{C_{li}}||_2$ dominates $||R^{C_{li}}||_2$ and we have for $n \geq n_0(\eta)$ $||L_R^{C_{li}}||_2 \leq 2||D_R^{C_{li}}||_2$.

Hence, with probability at least $1 - 4/n$, for every sub-cluster C_{li}, $||L_R^{C_{li}}||_2 \leq 2\sigma\sqrt{6m_{li}\log n}$. □

We stress that at this point, we have dealt with all of the randomness involved in recovering the clusters, across all levels. Specifically, we now know that with probability at least $1 - 4/n$, every noise Laplacian of size m_{li} will have spectral norm bounded by $O(\sigma\sqrt{m_{li}\log n})$.

B.4 Davis-Kahan for Laplacians and ℓ_2 Deviation Bounds

We now derive some results related to perturbation theory that will be useful in our final proof. The first is a variant of the Davis-Kahan theorem that bounds the eigenvector deviation in ℓ_2-norm. We also state Weyl's inequality which bounds the deviation between eigenvalues. Let λ_i be the ordered eigenvectors and eigenvalues of L_A, and $u^{(i)}, \mu_i$ be the ordered eigenvectors and eigenvalues of L_W.

Lemma 17 *(Davis-Kahan Modified) With probability at least $1 - \delta$,*

$$||u^{(i)} - v^{(i)}||_2 \leq \frac{\sqrt{2}||L_R||}{\xi_i}$$

where ξ_i denotes the eigengap for the i^{th} eigenvalue of L_A, i.e. $\xi_i = \min_{i \neq j}|\lambda_i - \lambda_j|$.

Proof (of Lemma 17) Note that $L_R + L_A = L_W$.

From Davis-Kahan theorem, we know that

$$|\sin\theta_i| \leq \frac{||L_R||}{\min_{j\neq i}|\lambda_i - \lambda_j|}$$

where λ_i and λ_j are respectively the i-th and j-th smallest eigenvalue of L_A and θ_i is the angle between $v^{(i)}$ and $u^{(i)}$, i.e. $\cos\theta_i = v^{(i)\mathsf{T}}u^{(i)}$

Without loss of generality, we can orient vectors as desired and assume that $|\theta_i| \leq \frac{\pi}{2}$. Since $v^{(i)}$ and $u^{(i)}$ are unit vectors, we get that

$$\begin{aligned}||u^{(i)} - v^{(i)}||_2 &\leq \sqrt{||v^{(i)}||^2 + ||u^{(i)}||^2 - 2v^{(i)\mathsf{T}}u^{(i)}} \\ &= |2\sin\frac{\theta_i}{2}| \leq |2\sqrt{2}\sin\frac{\theta_i}{2}\cos\frac{\theta_i}{2}| = |\sqrt{2}\sin\theta_i|\end{aligned}$$

The second inequality follow because $\sqrt{2}\cos\frac{\theta_i}{2} \geq 1$ under assumption that $|\theta_i| \leq \frac{\pi}{2}$. Combining this with Davis-Kahan gives us the desired result. □

For ease of reference, we also state here a well-known result in perturbation theory that we use.

Lemma 18 *(Weyl's Inequality) Let L_W, L_A be $n \times n$ positive definite matrices and let $L_R = L_W - L_A$. Let $\lambda_1 \leq ... \leq \lambda_n$ and $\mu_1 \leq ... \leq \mu_n$ be the eigenvalues of L_A and L_W respectively. Then, for all i, $|\lambda_i - \mu_i| \leq ||L_R||_2$.*

Before we proceed, we remark here that Lemma 17, combined with Lemma 16, immediately gives us an ℓ_2 deviation between the eigenvectors of the noisy HBM and the ideal HBM. Specifically, if we additionally use Lemma 8 to lower bound ξ_i, we see that for the cluster C_{li}:

$$||u^{(2)} - v^{(2)}||_2 = O\left(\sigma\sqrt{\frac{\log n}{m_{li}}}\right)$$

Using the uniform spectral bounds in Lemma 16, we arrive at this ℓ_2-norm deviation bound for all clusters of size at least $m = \omega(\log(n))$ with probability $1 - 4/n$.

B.5 Uniform bounds on $u^{(2)} - v^{(2)}$

Note that the above result is not sufficient to guarantee that spectral clustering will make no mistakes as $u^{(2)}$ could be spiked even if it is close to $v^{(2)}$ in ℓ_2. To make this guarantee, we perform a more careful analysis and show that $u^{(2)}$ is uniformly close to $v^{(2)}$ in every coordinate.

In this analysis, let us focus on a cluster C_s of size m_s. For ease of notation, we will denote the adjacency matrix of C_s by A and the perturbation of C_s by R. We will further use D_{Ai} and D_{Ri} to denote the sum of the i^{th} row of A and R respectively. Repeated application for all of the clusters, using the fact that all of the noise laplacians can be bounded, will guarantee the correctness of our algorithm across all clusters. Let $k = u^{(2)} - v^{(2)}$. We will show that with high probability, $k(i)$, the element-wise perturbation is uniformly low.

$$\begin{aligned}k(i) &= \frac{(L_{Ai} + L_{Ri})u^{(2)}}{\mu_2} - v^{(2)}(i) \\ &= \frac{L_{Ai}(v^{(2)} + k)}{\mu_2} + \frac{L_{Ri}u^{(2)}}{\mu_2} - v^{(2)}(i) \\ &= \frac{\lambda_2 v^{(2)}(i)}{\mu_2} + \frac{L_{Ai}k}{\mu_2} + \frac{L_{Ri}u^{(2)}}{\mu_2} - v^{(2)}(i) \\ &= v^{(2)}(i)\frac{\lambda_2 - \mu_2}{\mu_2} + \frac{L_{Ai}k}{\mu_2} + \frac{L_{Ri}u^{(2)}}{\mu_2}\end{aligned}$$

where L_{Ai} is the i-th row of L_A. Since the diagonal elements are much larger than the off-diagonals, we will have to treat them differently.

Consider, only the second term.

$$\frac{L_{Ai}k}{\mu_2} = \frac{D_{Ai}k(i) - A_ik}{\mu_2}$$

Similarly, for the third term.

$$\frac{L_{Ri}u^{(2)}}{\mu_2} = \frac{L_{Ri}v^{(2)}}{\mu_2} + \frac{D_{Ri}k(i)}{\mu_2} - \frac{R_ik}{\mu_2}$$

Using this we get,

$$k(i) = \frac{1}{c_i}\left(v^{(2)}(i)(\lambda_2 - \mu_2) - A_ik + L_{Ri}v^{(2)} - R_ik\right)$$

where $c_i = \mu_2 - D_{Ai} - D_{Ri}$.

We are interested in the absolute difference. Using triangle inequality we get,

$$|k(i)| \leq \frac{\overbrace{|v^{(2)}(i)(\lambda_2 - \mu_2)|}^{T_1} + \overbrace{|A_ik|}^{T_2} + \overbrace{|L_{Ri}v^{(2)}|}^{T_3} + \overbrace{|R_ik|}^{T_4}}{\underbrace{|c_i|}_{T_5}}$$

Call the numerator terms T_1, T_2, T_3 and T_4 and the denominator term T_5. We will bound each of these terms seperately.

Bound on T_1: To bound T_1 we will use simple results we have already derived.

$$T_1 = |v^{(2)}(i)(\lambda_2 - \mu_2)| = |v^{(2)}(i)||\lambda_2 - \mu_2| \leq \sqrt{\frac{K_\eta^\star}{m_s}}||L_R||_2$$

Here we use Lemma 8 along with Weyl's Inequality (Lemma 18). Using our uniform bound on $||L_R||_2$ (Lemma 16), we see that with probability $1 - 4/n$

$$T_1 \leq 2\sigma\sqrt{6K_\eta^\star \log n}$$

Bound on T_2: Remember that, $\kappa^\star = \min(\alpha_0, \frac{\gamma_S^\star}{1+\eta})$.

$$T_2 = |A_ik| \leq ||A_i||_2||k||_2 \leq \frac{\sqrt{m_s}\beta^*\sqrt{2}||L_R||_2}{\xi_2} \leq \frac{4\sigma\beta^*}{\kappa^\star}\sqrt{3\log n}$$

where ξ_2 is the eigengap corresponding to the second eigenvector. The first inequality is Cauchy-Schwarz while the second follows from Lemma 17. The third inequality uses Lemma 8 to bound the eigengap ξ_2 which is at least $m_s\kappa^\star$, and Lemma 16 to bound $||L_R||_2$. This inequality holds under the same $1 - 4/n$ probability event used in the T_1 bound.

Bound on T_3 The terms T_3 and T_4 are the main "noise" terms.

$$T_3 = |L_{Ri}v^{(2)}| = |D_{Ri}v^{(2)}(i) - R_iv^{(2)}|$$

Since each entry R_{ij} is subgaussian with scale factor σ, since $R_{ij}, R_{ij'}$ are independent for all $j \neq j'$, and since $||v^{(2)}||_2 = 1$, we conclude that $R_iv^{(2)}$ is distributed as a subgaussian with scale factor σ by Lemma 12.

$D_{Ri}v^{(2)}(i)$ is a subgaussian random variable with scale factor $\leq \sqrt{K_\eta^\star}\sigma$ since $v^{(2)}(i) \leq \sqrt{\frac{K_\eta^\star}{m_s}}$ and each entry D_{Ri} is subgaussian with scale factor $\sqrt{m_s}\sigma$. Since $\sigma^2 \leq K_\eta^\star\sigma^2$, T_3 is a draw from a subgaussian with scale factor $\leq \sqrt{2K_\eta^\star}\sigma$.

To ensure that T_3 is uniformly low for all i, we take a union bound and use Lemma 11. Note that this union bound is across all levels of the hierarchy, so there are $nl \leq n^2$ subgaussians that we must bound. We get that with probability at least $1 - 2/n$,

$$T_3 \quad \leq \quad 4\sigma\sqrt{K_\eta^\star 3\log n}$$

Bound on T_4

$$T_4 = |R_i k| \leq ||R_i||_2||k||_2 \leq ||R||_2||k||_2$$

From the proof of Lemma 16, we see that for $m_s = \omega(\log n)$, and for n large enough, under the $1 - 4/n$ probability event described in T_1,

$$||R||_2 \leq C\sigma\sqrt{m_s}$$

for some absolute constant C. So we have,

$$T_4 \leq C\sigma\sqrt{m_s}\frac{\sqrt{2}||L_R||_2}{\xi_2} \leq \frac{4C\sigma^2\sqrt{3\log n}}{\kappa^\star}$$

Bound on T_5 The term T_5 appears in the denominator and here, we establish a lower bound on it.

$$T_5 = |\mu_2 - D_{Ai} - D_{Ri}| = |D_{Ai} + D_{Ri} - \mu_2|$$

Note that $D_{Ai} \geq \frac{m_s}{1+\eta}(\eta\alpha_s + \alpha_{s+})$ where $\alpha_{s+} \triangleq \min\{\alpha_{s\circ L}, \alpha_{s\circ R}\}$ and that $\mu_2 \leq \lambda_2 + ||L_R||_2 \leq m_s\beta_s + ||L_R||_2$. Hence:

$$\begin{aligned} T_5 \quad &\geq \quad \left|\frac{m_s}{1+\eta}(\eta\alpha_s + \alpha_{s+}) + D_{Ri} - m_s\beta_s - ||L_R||_2\right| \\ &\geq \quad \frac{m_s}{1+\eta}\left|\alpha_{s+} + \eta\alpha_s - (1+\eta)\beta_s - \frac{1+\eta}{m_s}(2||D_R||_2 + ||R||_2)\right| \end{aligned}$$

Where the inequalities only hold provided that the term inside the absolute value is ≥ 0. Note that $\alpha_{s+} + \eta\alpha_s - (1+\eta)\beta_s$ is just γ. We will show that for large enough n, this is indeed true. Under the $1 - 4/n$ probability event described in the T_1 bound, we have:

$$2||D_R||_2 + ||R||_2 \leq 3\sigma\sqrt{6m_s\log n}$$

Now, provided that $\sigma = o(\gamma\sqrt{\frac{m_s}{\log n}})$, and using the definition of γ, we have that then $\frac{1+\eta}{m_s}(2||D_R||_2 + ||R||_2) = o(\gamma)$, and for large enough n, we can conclude that $\gamma - \frac{1+\eta}{m_s}(2||D_R||_2 + ||R||_2) \geq \frac{\gamma}{2} \geq \frac{\gamma_S^\star}{2}$. From the statement of the theorem we have $\sigma = o(\min(\kappa^{\star 5}\sqrt{\frac{m_s}{\log n}}, \kappa^{\star 4}\sqrt[4]{\frac{m_s}{\log n}})) = o(\gamma\sqrt{\frac{m_s}{\log n}})$. Therefore:

$$T_5 \geq \frac{m_s}{1+\eta}\frac{\gamma_S^\star}{2} \geq \frac{m_s\kappa^\star}{2}$$

Putting Everything Together Combining all of the terms together, we see that with probability $1 - \frac{6}{n}$

$$||k||_\infty \leq \frac{2\sigma\sqrt{\log n}}{m_s\kappa^\star}\left[s\sqrt{6K_\eta^\star} + \frac{4\sqrt{3}\beta^*}{\kappa^\star} + 4\sqrt{3K_\eta^\star} + \frac{4C\sigma\sqrt{3}}{\kappa^\star}\right]$$

To arrive at our final rate, we must characterize the dependence of $K_\eta^\star$ on κ^*. Note in the expression for $K_\eta^\star$ that $\min\{\alpha_{s\cdot L}, \alpha_{s\cdot R}\} - \beta_s \geq \gamma^*$, and that the terms in the numerator are all bounded by a constant depending on η and β^* which is the bound on the entries of the similarity matrix. Thus, we get $K_\eta^\star \leq \frac{C_{\eta,\beta^*}}{\gamma^{\star 4}} \leq \frac{C_{\eta,\beta^*}}{\kappa^{\star 4}}$.

We now see that if $\sigma = o(\min(\kappa^{\star 5}\sqrt{\frac{m_s}{\log n}}, \kappa^{\star 4}\sqrt[4]{\frac{m_s}{\log n}}))$ and $m = \omega(\log n)$, then for large enough n we have $||k||_\infty \leq \sqrt{\frac{1}{K_\eta m}}$ and our algorithm makes no mistakes in resolving all clusters of size at least m_s.

C Proof of Theorem 2 (k-way)

The proof will be very similar to that of Theorem 1.

The difficulty here is that the spectral embedding of each point is not just a single number, but rather a k-dimensional vector. To make matters worse, because L_A has a k-dimensional eigenspace associated with eigenvalue 0 (in other words, eigenvalue 0 has geometric multiplicity k), there are many different possible spectral embeddings of each point–one for each set of basis of the eigenspace.

Let $u^{(1)}, ..., u^{(k)}$ be perturbed eigenvectors of L_W. The set of $u^{(j)}$'s cannot be close to all sets of lowest k eigenvectors of L_A because there are infinite number of sets of lowest k eigenvectors of L_A due to geometric multiplicity. Thus, the best we can say is that there exist at least one set of lowest k eigenvectors of L_A that is close to $u^{(1)}, ..., u^{(k)}$. Lemma 19, 20 formalize these concepts.

The following Lemmas extend Davis-Kahan theorem to describe perturbation of subspaces:

Lemma 19 *Let W be a matrix with eigenvalues $\mu_1 \leq \mu_2, ... \leq \mu_n$ (possibly with multiplicity) and corresponding eigenvectors $u_1, u_2, ...u_n$. Let A be a matrix with eigenvalues $\lambda_1 \leq \lambda_2, ... \leq \lambda_n$ (possibly with multiplicity) and corresponding eigenvectors $v_1, v_2, ...v_n$. Let $R \equiv W - A$.*

Let $U = span\{u_i\}_{i \in I}$ where I is some index set. Let $V = span\{v_i\}_{i \in I}$. Then we have, for all unit-normed $u \in U$:

$$||P_{V^\perp} u||_2 \leq \frac{2||R||_2}{\delta}\sqrt{k}$$

where $k \equiv \dim U = \dim V$, $P_{V^\perp}$ is the orthogonal projection onto $V^\perp$, $\delta \equiv \min_{i \in I} \delta_i$ and $\delta_i \equiv \min_{j \notin I} |\lambda_i - \lambda_j|$.

Intuitively, U, an eigen-subspace of W must be close to V, the corresponding eigen-subspace of A. We simply quantified "close" as the projection of U onto $V^\perp$.

Proof Let U, V be eigen-subspaces of W, A as defined in theorem. Fix $i \in I$ and let μ_i be an eigen-value that correspond to $u_i \in U$. Define $\bar{A} = A - \lambda_i I$ and $\bar{W} = W - \lambda_i I$.

Recall that u_i is the eigenvector of W that correspond to μ_i; we can expand u_i in the eigenbasis of A and get $u_i = \sum_j c_j v_j$.

$$\begin{aligned}
||\bar{A}u_i||_2^2 &= ||\bar{A}\sum_j c_j v_j||_2^2 \\
&= \sum_j c_j^2 (\lambda_j - \lambda_i)^2 \\
&\geq \sum_{j \notin I} c_j^2 (\lambda_j - \lambda_i)^2 \\
&\geq \delta_i^2 \sum_{j \notin I} c_j^2 \\
&= \delta_i^2 ||P_{V^\perp} u_i||_2^2
\end{aligned}$$

By using Weyl's Inequality, we can upper bound $||\bar{A}u_i||$ as such:

$$||\bar{A}u_i||_2 \leq ||\bar{W}u_i||_2 + ||R||_2 \leq |\mu_i - \lambda_i| + ||R||_2 \leq 2||R||_2$$

Combine the two results, we get:

$$||P_{V^\perp} u_i||_2 \leq \frac{2||R||_2}{\delta_i}$$

Let $u \in U$ and let $||u||_2 = 1$, then $u = \sum_{j \in I} c_j u_j$. We will now upper bound $||P_{V^\perp} u||_2$

$$\begin{aligned} ||P_{V^\perp} u||_2^2 &= ||\sum_{j \in I} c_j P_{V^\perp} u_j||_2^2 \\ &= \sum_{j \in I} c_j^2 ||P_{V^\perp} u_j||_2^2 + \sum_{j \neq i, \in I} c_j c_i \langle P_{V^\perp} u_j, P_{V^\perp} u_i \rangle \end{aligned}$$

We already have that $||P_{V^\perp} u_i||_2^2 \leq \frac{4||R||_2^2}{\delta_i^2}$. Define $\delta = \min_i \delta_i$, then we have $||P_{V^\perp} u_i||_2^2 \leq \frac{4||R||_2^2}{\delta^2}$.

By Cauchy-Schwartz, we get $\langle P_{V^\perp} u_j, P_{V^\perp} u_i \rangle \leq ||P_{V^\perp} u_j||_2 ||P_{V^\perp} u_i||_2 \leq \frac{4||R||_2^2}{\delta^2}$.

Combine the two above bounds, we can now continue upper bounding $||P_{V^\perp} u||$:

$$\begin{aligned} ||P_{V^\perp} u||_2^2 &\leq \frac{4||R||_2^2}{\delta^2} (\sum_{j \in I} c_j^2 + \sum_{j \neq i, \in I} |c_i||c_j|) \\ &\leq \frac{4||R||_2^2}{\delta^2} (\sum_{j \in I} |c_j|)^2 \\ &\leq \frac{4||R||_2^2}{\delta^2} k \sum_{j \in I} |c_j|^2 \\ &\leq \frac{4||R||_2^2}{\delta^2} k \end{aligned}$$

Thus, we get $||P_{V^\perp} u||_2 \leq \frac{2||R||_2}{\delta} \sqrt{k}$ as desired □

Lemma 20 *(Eigenspace-Perturbation) Let $U = span\{u_i\}_{i \in I}$ and $V = span\{v_i\}_{i \in I}$ be eigen-subspaces of matrices W, A respectively.*

Assume $\frac{2||R||_2}{\delta} \sqrt{k} \leq 1/2$, then there exist a V-invariant isometry (orthonormal matrix) Θ such that for all i

$$||\Theta v_i - u_i||_2 \leq \frac{6||R||_2}{\delta} \sqrt{k}$$

We say that Θ is V-invariant if for all $v \in V$, $\Theta v \in V$.

The difficulty of proving Lemma 20 comes from the fact that $P_V u_i$ and $P_V u_j$ need not be orthogonal even if u_i and u_j are orthogonal. We use the next PSD Deviation Lemma to address this difficulty.

Lemma 21 *(PSD Deviation) Let K be a positive definite matrix with some eigenvectors that span V. Let $0 \leq \theta < 1$ and let all eigenvalues of K be between $1 + \theta$ and $1 - \theta$.*

Then $||Kv - v||_2 \leq \theta ||v||_2$ for all $v \in V$.

Proof (of Lemma 21) Let $w_1, ...w_k$ be the eigenvectors of K that span V with corresponding eigenvalues $\lambda_1, ...\lambda_k$.

Then $u = \sum_k c_k w_k$ and we get:

$$\begin{aligned} ||Kv - v||_2 &= ||\sum_k c_k K w_k - \sum_k c_k w_k||_2 \\ &= ||\sum_k c_k \lambda_k w_k - \sum_k c_k w_k||_2 \\ &= ||\sum_k c_k (\lambda_k - 1) w_k||_2 \\ &\leq (\max_i |\lambda_i - 1|) ||\sum_k c_k w_k||_2 \\ &= \theta ||v||_2 \end{aligned}$$

□

Now we can prove the Eigenspace Perturbation lemma:

Proof (of Lemma 20)

Define $v_i' = P_V u_i$ for $i \in I$. The collection of vectors $\{v_i'\}_{i \in I}$ need not be orthogonal, but we claim they are independent. To see this, suppose that there exist coefficients c_i such that $\sum_{i \in I} c_i v_i' = 0$. Then

$$\sum_{i \in I} c_i v_i' = \sum_{i \in I} c_i P_V(u_i) = P_V(\sum_{i \in I} c_i u_i) = 0$$

The vector $\sum_{i \in I} c_i u_i$ is in U and non-zero. Hence, by Lemma 19 and the assumption that $\frac{2||R||_2}{\delta}\sqrt{k} \leq 1/2$, $||P_V(\sum_{i \in I} c_i u_i)||_2 \geq \frac{1}{2}||\sum_{i \in I} c_i u_i||_2 > 0$. This is a contradiction.

Because v_i''s are independent, there exist a basis-transform linear operator K such that $K v_i' = v_i$ for all i and $Kw = w$ for all $w \notin V$. Note that K is V-invariant since $\{v_i'\}_{i \in I}$ spans V.

Let $K = \Psi K^*$ be the V-invariant polar decomposition of K, that is, Ψ is an isometry, K^* is positive semidefinite, and K^* and Ψ are both V-invariant. Since Ψ is an isometry and hence preserves inner product, we get that the collection of vectors $\{K^* v_i'\}_{i \in I}$ must be orthogonal.

Also, since Ψ is an isometry and hence preserves norm, we get that $||K^* v_i'||_2 = 1$ for all $i \in I$ and $K^* \circ P_V$ is an isometry when restricted to subspace U. Since the eigenvalues of P_V restricted to U are bounded between 1 and $1 - \frac{2||R||_2}{\delta}\sqrt{k}$, we get that the eigenvalues of K^* restricted to range$(P_V) = V$ must be bounded between 1 and $1/(1 - \dfrac{2||R||_2}{\delta}\sqrt{k})$.

By assumption from theorem, we can bound, by using the fact that $\frac{1}{1-a} \leq 1 + 2a$ for $0 \leq a \leq 1/2$, the eigenvalues of K^* between 1 and $1 + 4\frac{||R||_2}{\delta}\sqrt{k}$. Hence, by Lemma 21, we get that for all $v \in V$, $||K^* v - v||_2 \leq 4\frac{||R||_2}{\delta}\sqrt{k}||v||_2$. Thus, we get:

$$\begin{aligned} ||u_i - K^* P_V u_i||_2 &\leq ||u_i - P_V u_i||_2 + ||K^* P_V u_i - P_V u_i||_2 \\ &\leq 2\frac{||R||_2}{\delta} + 4\frac{||R||}{\delta}\sqrt{k}||P_V u_i||_2 \\ &\leq 6\frac{||R||_2}{\delta}\sqrt{k} \end{aligned}$$

Where we used the fact that $||u_i - P_V u_i||_2 = ||P_{V^\perp} u_i||_2$, and Lemma 19 for the second inequality and the trivial upper bound that $||P_V u_i||_2 \leq 1$ for the third inequality.

Since $v_i = K P_V u_i = \Psi K^* P_V u_i$, $\Psi^{-1} v_i = K^* P_V u_i$. Hence, we have proven the theorem with Ψ^{-1} as the isometry. □

The next lemma describes the spectrum of the Laplacian of a k-Block Diagonal similarity matrix in a manner similar to Lemma 7 and Lemma 8.

Lemma 22 *Let A be a k-Block Diagonal Matrix with blocks $A^{(1)}, ...A^{(k)}$ such that all entries in $A^{(1)}, ...A^{(k)}$ are between β_1 and β_0 where $0 < \beta_0 \leq \beta^*$ and all remaining entries of A are 0, i.e.*

$$W = \begin{bmatrix} A^{(1)} & & ... & 0 \\ & A^{(2)} & & \\ ... & & ... & \\ 0 & & & A^{(k)} \end{bmatrix}$$

Let $0 < \nu < 1$ be such that νn is the size of the largest cluster. Then:

1. *$\lambda_1, ..., \lambda_k$, the lowest k eigenvalues of L_A, are 0 with corresponding eigenvectors*

$$v^{(1)} = \frac{1}{\sqrt{|C_1|}}(\underline{1}_{C_1}, \underline{0}_{C_2}, ..., \underline{0}_{C_k})$$
$$v^{(2)} = \frac{1}{\sqrt{|C_2|}}(\underline{0}_{C_1}, \underline{1}_{C_2}, ...\underline{0}_{C_k})$$
$$...$$
$$v^{(k)} = \frac{1}{\sqrt{|C_k|}}(\underline{0}_{C_1}, \underline{0}_{C_2}, ...\underline{1}_{C_k})$$

where $\underline{0}_{C_1}$ is an all-zero vector of length $|C_1|$.

2. *$\lambda_{k+1} \geq \frac{\nu n}{\eta}\beta_0$ (note that $\frac{\nu n}{\eta}$ lower bounds size of the smallest cluster)*

Proof The first claim follows because L_A is also block-diagonal and the diagonal blocks $(L_A)^{(i)} = L_{A^{(i)}}$.

To prove the second claim, we construct a block-diagonal matrix S with the same block structure as A and furthermore, the diagonals $S^{(1)}, ..., S^{(k)}$ all have constant value of β_0. The claim then follows by Lemma 9 and Lemma 10. □

Now we proceed to the proof of Theorem 2:

Proof (of Theorem 2)

Let $j \in \{1, ..., k\}$, define $v'^{(j)} = \Theta v^{(j)}$. Since Θ is V-invariant, we know that $L_A v'^{(j)} = 0$.

Let let $h^{(j)} = u^{(j)} - v'^{(j)}$.

$$\begin{aligned}
h^{(j)}(i) &= u^{(j)}(i) - v'^{(j)}(i) \\
&= \frac{(L_{A_i} + L_{R_i})u^{(j)}}{\mu_j} - v'^{(j)}(i) \\
&= \frac{L_{A_i}(v^{(j)} + h^{(j)})}{\mu_j} + \frac{L_{R_i}u^{(j)}}{\mu_j} - v'^{(j)}(i) \\
&= \frac{L_{A_i}h^{(j)}}{\mu_j} + \frac{L_{R_i}u^{(j)}}{\mu_j} - v'^{(j)}(i) \\
&= \frac{D_{A_i}h^{(j)}(i) - A_i h^{(j)}}{\mu_j} + \\
&\left(\frac{L_{R_i}v'^{(j)}}{\mu_j} + \frac{D_{R_i}h^{(j)}(i)}{\mu_j} - \frac{R_i h^{(j)}}{\mu_j}\right) - v'^{(j)}(i)
\end{aligned}$$

We will collect the terms containing $h(i)$ and get

$$\begin{aligned}
&\mu_j h^{(j)}(i) - D_{A_i}h^{(j)}(i) - D_{R_i}h^{(j)}(i) \\
&= -A_i h^{(j)} + L_{R_i}v'^{(j)} - R_i h^{(j)} - v'^{(j)}(i)\mu_j
\end{aligned}$$

and hence

$$h^{(j)}(i) = \underbrace{\frac{1}{|\mu_j - D_{A_i} - D_{R_i}|}}_{T_5}\left(\underbrace{|v'^{(j)}(i)|}_{T_1} + \underbrace{|A_i h^{(j)}|}_{T_2} + \underbrace{|L_{R_i}v'^{(j)}|}_{T_3} + \underbrace{|R_i h^{(j)}|}_{T_4}\right)$$

Call the numerator terms T_1, T_2, T_3, T_4 and call the denominator term T_5. We will bound each of these terms uniformly across all clusters $j = 1, ..., k$ and across all elements $h^{(j)}(i)$, $i = 1, ...n$.

Bound for T_1: Since Θ is V-invariant, we know that $v'^{(j)} = \sum_{t=1}^{k} \alpha_t v^{(t)}$ and hence, $v'^{(j)}$ has vector-structure of $(\frac{1}{\sqrt{|C_1|}}\underline{\alpha_1}, \frac{1}{\sqrt{|C_2|}}\underline{\alpha_2}, \frac{1}{\sqrt{|C_3|}}\underline{\alpha_3}, ...)$ where $\underline{\alpha_1}$ is sub-vector of length $|C_1|$ etc.

Because $\alpha_t \leq 1$ for all j, we know that $|v'^{(j)}(i)| \leq \sqrt{\frac{\eta}{\nu n}}$.

We can bound $|\mu_j| \leq ||L_R||_2 + |\lambda_j| = ||L_R||_2$ by Weyl's Inequality. By Lemma 14, $||L_R||_2 \leq 4\sigma\sqrt{n \log n}$ with probability at least $1 - \frac{4}{n}$. Hence, T_1 is upper bounded by $4\sigma\sqrt{\frac{\eta}{\nu}}\sqrt{\log n}$.

Bound for T_2: $|A_i h^{(j)}| \leq ||A_i||_2 ||h^{(j)}||_2 \leq \sqrt{\nu n}\beta^* \frac{6\sqrt{k}||L_R||_2}{\xi}$ where the bound on $||h^{(j)}||_2$ comes from Lemma 20.

Also, by Lemma 22, $\xi \equiv \lambda_{k+1} - \lambda_k = \lambda_{k+1} \geq \frac{\nu n}{\eta}\beta_0$. Thus, $|A_i h^{(j)}| \leq 6\frac{\beta^*}{\beta_0}\eta\sqrt{\frac{k}{\nu n}}||L_R||_2$.

In the $1 - \frac{4}{n}$ probability event described in Lemma 14, we get that

$$|A_i h^{(j)}| \leq 6\frac{\beta^*}{\beta_0}\eta\sqrt{\frac{k}{\nu}}4\sigma\sqrt{\log n}$$

Note that in order to invoke Lemma 20, we need to satisfy the condition that $\frac{6\sqrt{k}||L_R||_2}{\xi} \leq \frac{1}{2}$. Since

$$\begin{aligned}\frac{6\sqrt{k}||L_R||_2}{\xi} &\leq \frac{6\sigma\eta\sqrt{k}4\sqrt{\log n}}{\nu\beta_0\sqrt{n}} \\ &\leq \frac{6\sigma\eta}{\nu\beta_0}4\sqrt{\frac{k\log n}{n}}\end{aligned}$$

and since $\sigma = o\left(\frac{\beta_0}{k}(\frac{n}{k\log n})^{1/4}\right)$ under assumption of the theorem, for large enough n, the condition of Lemma 20 will be satisfied.

Bound for T_3:

$$|L_{R_i} v'^{(j)}| \leq |D_{R_i} v'^{(j)}(i)| + |R_i v'^{(j)}|$$

We see that $|D_{R_i} v'^{(j)}(i)| \leq |D_{R_i}||v'^{(j)}(i)|$. We know that in the same $1 - \frac{4}{n}$ probability event described in $T1$ bound, $|D_{R_i}| \leq 4\sigma\sqrt{n\log n}$. Hence,

$$|D_{R_i}||v'^{(j)}(i)| \leq 4\sigma\sqrt{\frac{\eta}{nu}\log n}$$

The second term $|R_i v'^{(j)}|$ is trickier to bound. We first expand $v'^{(j)}$ in term of $v^{(1)}, ..., v^{(k)}$.

$$\begin{aligned}|R_i v'^{(j)}| &\leq |\sum_{t=1}^{k} \alpha_t R_i v^{(t)}| \\ &\leq (\sum_{t=1}^{k} |\alpha_t|) \max_{t=1,...,k} |R_i v^{(t)}| \\ &\leq \sqrt{k} \max_{t=1,...,k} |R_i v^{(t)}|\end{aligned}$$

We know

$$R_i v^{(t)} = \frac{1}{\sqrt{|C_t|}} \sum_{l=1}^{|C_t|} R_{il}$$

By Lemma 12, we get that $R_i v^{(t)}$ is subgaussian with scale factor σ. Hence, with probability at least $1 - \frac{2}{n}$, uniform across $i = 1, ..., n$, $max_{t=1,...,k}|R_i v^{(t)}| \leq \sigma\sqrt{6\log n}$.

Hence, T_3 can be bounded as

$$|D_{R_i}v^{(j)}| + |R_i v_t| \leq 4\sigma\sqrt{\frac{\eta}{\nu}\log n} + \sigma\sqrt{k6\log n}$$

Bound for $T4$:

$$\begin{aligned}
|R_i h^{(j)}| &\leq ||R_i||_2 ||h^{(j)}||_2 \\
&\leq C\sigma\sqrt{n}\frac{6\sqrt{k}||L_R||_2}{\xi} \\
&\leq (C\sigma\sqrt{n})\frac{12\sqrt{kn}\eta\sqrt{4\log n}}{\nu n\beta_0} \\
&\leq 12C\sigma^2\frac{\sqrt{k}}{\beta_0}\frac{\eta}{\nu}\sqrt{4\log n}
\end{aligned}$$

where we will assume the $1 - \frac{4}{n}$ probability event described in $T1$ bound.

Bound for $T5$: Recall that since $T5$ appears in the denominator, we need a lower bound for it as opposed to an upper bound.

$$\begin{aligned}
|\mu_j - D_{A_i} - D_{R_i}| &= |D_{A_i} + D_{R_i} - \mu_j| \\
&\geq |\frac{\nu n}{\eta}\beta_0 + D_{R_i} - ||L_R||_2| \\
&\geq |\frac{\nu n}{\eta}\beta_0 - 3||L_R||_2| \\
&\geq \frac{\nu n}{\eta}|\beta_0 - \underbrace{\sigma\frac{\eta}{\nu n}4\sqrt{n\log n}}_{\text{decaying term}}|
\end{aligned}$$

Where the third inequality occurs under the same $1 - \frac{4}{n}$ probability event described in $T1$ bound.

Recall that we assume $\sigma = o\left(\frac{\beta_0}{k}(\frac{n}{k\log n})^{1/4}\right)$ in the statement of the theorem and under this condition, for large enough n, the decaying term will be less than $\frac{\beta_0}{2}$.

$$|\mu_j - D_{A_i} - D_{R_i}| \geq \frac{\nu n}{\eta}\frac{\beta_0}{2}$$

Suppose that both the event described in $T1$ and the event described in $T3$ hold, which happens with probability $1 - \frac{8}{n}$ by union bound, the following bounds hold simultaneously for all $j = 1, ..., k$.

$$\begin{aligned}
T_1 &\leq 4\sigma\sqrt{\frac{\eta}{\nu}}\sqrt{\log n} \\
T_2 &\leq 4\sigma\frac{\beta_1}{\beta_0}\eta\sqrt{\frac{k}{\nu}}\sqrt{\log n} \\
T_3 &\leq 4\sigma\sqrt{\frac{\eta}{\nu}\log n} + \sigma\sqrt{6k\log n} \\
T_4 &\leq 12C\sigma^2\frac{\sqrt{k}}{\beta_0}\frac{\eta}{\nu}\sqrt{4\log n} \\
T_5 &\geq \frac{\nu n}{\eta}\frac{\beta_0}{2}
\end{aligned}$$

Combining everything together, we conclude that, uniformly across all $j = 1, ..., k$:

$$||h^{(j)}||_\infty \leq 12\sigma\sqrt{4\log n}\frac{2\eta}{\nu n\beta_0}\left[\sqrt{\frac{\eta}{\nu}} + \frac{\beta_1\eta}{\beta_0}\sqrt{\frac{k}{\nu}} + \sqrt{k} + C\sigma\frac{\sqrt{k}}{\beta_0}\frac{\eta}{\nu}\right]$$

Since we hold β_1 and η to be a constant and $\nu \leq 1$, we see that the last term of the sum dominates the entire sum. We also note that $\nu \geq \frac{1}{k}$ and thus $\frac{1}{\nu} \leq k$.

It is then straightforward to check that under the assumption that $\sigma^2 = o\left(\sqrt{\frac{n}{\log n}}\frac{\beta_0^2}{k^{5/2}}\right)$, then for large enough n, $||h^{(j)}||_\infty \leq \sqrt{\frac{1}{8\nu nk}}$

Embedding of each point onto basis $\{v^{(1)}, ...v^{(k)}\}$ is k-dimensional vector with exactly one non-zero coordinate. By the above definition, we can see that if points $p_1, p_2 \in \mathbb{R}^k$ are in the different clusters, then $||p_1 - p_2|| \geq \sqrt{\frac{2}{\nu n}}$.

Let $v'^{(j)} = \Theta v^{(j)}$ be the transformed orthonormal basis, we will show that the embeddings of points onto the transformed basis maintain the same pair-wise distance. We know that $v'^{(j)} = \sum_j \alpha_{jt} v^{(t)}$ and hence, $v'^{(j)}$ has vector-structure of $(\frac{1}{\sqrt{|C_1|}}\underline{\alpha_{j1}}, \frac{1}{\sqrt{|C_2|}}\underline{\alpha_{j2}}, \frac{1}{\sqrt{|C_3|}}\underline{\alpha_{j3}}, ...)$ where $\underline{\alpha_{j1}}$ is sub-vector of length $|C_1|$ whose every entry is α_{j1}.

Let $p_1, p_2 \in \mathbb{R}^k$ be two points in the transformed-basis-embedding. Let p_1 be in cluster a and p_2 be in cluster b, then $||p_1 - p_2||_2 = ||\frac{1}{\sqrt{|C_a|}}(\alpha_{1a}, ...\alpha_{ka}) - \frac{1}{\sqrt{|C_b|}}(\alpha_{1b}, ...\alpha_{kb})||$. Thus, if p_1, p_2 are in the same cluster, $||p_1 - p_2|| = 0$.

Let $\alpha^a \triangleq (\alpha_{1a}, ...\alpha_{ka})$ and $\alpha^b \triangleq (\alpha_{1b}, ...\alpha_{kb})$. Then

$$\begin{aligned}
&||\frac{1}{\sqrt{|C_a|}}\alpha^a - \frac{1}{\sqrt{|C_b|}}\alpha^b||^2 \\
&= \frac{1}{|C_a|}||\alpha^a||^2 - \frac{1}{\sqrt{|C_a||C_b|}}2\langle\alpha^a, \alpha^b\rangle + \frac{1}{|C_b|}||\alpha^b||^2
\end{aligned}$$

Define $k \times k$ matrix M such that $M_{jt} = \alpha_{jt}$. Hence, row j of M contains the linear coefficients of $v'^{(j)}$ in term of basis $\{v^{(1)}, ...v^{(k)}\}$. Since $v'^{(j)}$'s are orthonormal, it must be that rows of M are orthonormal and therefore, M must be an isometry and its columns are also orthonormal.

Thus, we get that $||\alpha^a|| = ||\alpha^b|| = 1$ and $\langle\alpha^a, \alpha^b\rangle = 0$ and that $||p_1 - p_2||^2 = \frac{1}{|C_a|} + \frac{1}{|C_b|} \geq \frac{2}{\nu n}$ and that if p_1, p_2 are in different clusters, then $||p_1 - p_2|| \geq \sqrt{\frac{2}{\nu n}}$.

Let q_1, q_2 be perturbed version of p_1, p_2, that is, the same points embedded in $(u^{(1)}, ..., u^{(k)})$-basis. Since each coordinate of the perturbed vector $u^{(j)}$ can change by at most $\sqrt{\frac{1}{8\nu nk}}$ from $v'^{(j)}$, we get that $||p_1 - q_1||_2 \leq \sqrt{\frac{1}{8\nu n}}$ and likewise for $||p_2 - q_2||_2$.

If q_1, q_2 are in the same cluster, $||q_1 - q_2||_2 \leq \sqrt{\frac{1}{2\nu n}}$ and if q_1, q_2 are in different clusters, $||q_1 - q_2||_2 \geq \sqrt{\frac{2}{\nu n}} - \sqrt{\frac{1}{2\nu n}} \geq \sqrt{\frac{1}{2\nu n}}$.

Since the maximum distance between two points in the same cluster is less than minimum distance between two points in different clusters, in our modified k-means procedure, the k chosen cluster centers will be in different clusters and the remaining points will be assigned to the correct clusters.

□

D Proofs of Theorem 3 and 4

First we state a version of Fano's Inequality from [15]:

Theorem 23 *Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ such that:*

1. $d(\theta_j, \theta_k) \geq 2s > 0, \ \forall 0 \leq j < j \leq M$
2. $P_j \ll P_0, \ \forall j = 1, \ldots, M$, *and*

$$\frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M$$

with $0 < \alpha < 1/8$ and $P_j = P_{\theta_j}, j = 0, 1, \ldots, M$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right) > 0$$

Recall that given a parameter θ_0 we choose the subset $\{\theta_1, \ldots, \theta_M\}$ in the following way. Consider, that there are $n/2$ entries in I_1, we will simply use the first $n/2 - 1$ of these indices and for each $j \in \{1, \ldots, n/2\}$ use a different index $\notin I_1$.

In order to apply Theorem 23. We need to calculate the Hamming distance between the estimate $\hat{I}_1$ and the true I_1 as d. It is clear that $d(\theta_j, \theta_k) \geq 1$ for $j \neq k$.

We also need the KL-divergence between the probability distributions induced by the parameters. Since each distribution is a gaussian, we can readily verify that $K(P_j, P_0) = \frac{n(p-q)^2}{\sigma^2} = \frac{n\gamma^2}{\sigma^2}$ for each j.

With this two calculations, we apply Theorem 23 and arrive at our lower bound.

As for the upper bound, our combinatorial procedure solves the minimum cut of size $n/2$. This is equivalent to the algorithm that outputs a set of coordinates of size $n/2$ (denoted by $\hat{I}$) so as to maximize the contrast between the two diagonal blocks (the $\hat{I}\hat{I}$ block and the $\hat{I}^c\hat{I}^c$ block) and the two off-diagonal blocks (the $\hat{I}\hat{I}^c$ block and the $\hat{I}^c\hat{I}$ block). Denote the true clusters by I and I^c.

Define,

$$S(W, I) = \sum_{i \in I, j \in I} W_{ij} + \sum_{i \in I^c, j \in I^c} W_{ij} - \sum_{i \in I, j \in I^c} W_{ij} - \sum_{i \in I^c, j \in I} W_{ij}$$

Our algorithm exactly minimizes $S(W, I)$ subject to $|I| = n/2$. To analyze this procedure it's useful to consider the random variable ζ defined as,

$$\zeta_{\hat{I}} = S(W, I) - S(W, \hat{I})$$

Further given a set $\hat{I}$ define the number of indices in which $\hat{I}$ and I agree to be s.

It is easy to see that for a given s,

$$\zeta_s \sim \mathcal{N}\left(8s\left(\frac{n}{2} - s\right)(p - q), 16s\left(\frac{n}{2} - s\right)\sigma^2\right)$$

The combinatorial procedure succeeds if w.h.p $\zeta_{\hat{I}} \geq 0$ for every $\hat{I} \neq I$.

This probability (by the application of a union bound) is bounded by

$$\begin{aligned}
\mathbb{P}_{\text{error}} &\leq \sum_{s=1}^{n/2-1} \binom{n/2}{s}\binom{n/2}{n/2-s}\mathbb{P}(\zeta_s \leq 0) \leq \sum_{s=1}^{n/2-1} \binom{n/2}{s}^2 \exp\left\{\frac{-C_1 s(n/2-s)(p-q)^2}{\sigma^2}\right\} \\
&\leq \sum_{s=1}^{\lfloor n/4 \rfloor} \binom{n/2}{s}^2 \exp\left\{\frac{-C_1 s(n/2-s)(p-q)^2}{\sigma^2}\right\} + \sum_{s=\lceil n/4 \rceil}^{n/2-1} \binom{n/2}{s}^2 \exp\left\{\frac{-C_1 s(n/2-s)(p-q)^2}{\sigma^2}\right\} \\
&\leq \sum_{s=1}^{\lfloor n/4 \rfloor} \exp\left\{C_2 s \log n - \frac{C_1 s(n/2-s)(p-q)^2}{\sigma^2}\right\} \\
&\quad + \sum_{s=\lceil n/4 \rceil}^{n/2-1} \exp\left\{C_3(n/2-s)\log n - \frac{C_1 s(n/2-s)(p-q)^2}{\sigma^2}\right\} \\
&\leq \max_{s\in\{1,\ldots,\lfloor n/4 \rfloor\}} \exp\left\{C_2' s\left[\log n - \frac{C_1' n(p-q)^2}{\sigma^2}\right]\right\} \\
&\quad + \max_{s\in\{\lceil n/4 \rceil,\ldots,n/2\}} \exp\left\{C_3'(n/2-s)\left[\log n - \frac{C_1'' n(p-q)^2}{\sigma^2}\right]\right\}
\end{aligned}$$

A simple calculation shows that if we select $n(p-q)^2 > C\sigma^2 \log\left(\frac{n}{2}\right)$ for a large enough constant C, both exponents are negative and this probability can be made smaller than any constant δ.

This establishes the minimax rate (i.e. the minimax scaling up to constants).

D.1 A minimax spectral algorithm

We show that for the restricted case of block constant similarities, under certain assumptions the Algorithm of [9] is minimax optimal for the k-way problem, for k constant. The algorithm and its analysis rely crucially on the fact that the noiseless matrix is of rank k and do not directly extend to the non constant block similarities we consider in this paper.

Consider the following algorithm:

1. Input: Noisy similarity matrix W, number of clusters k
2. Randomly divide the columns of W into two parts W_1 and W_2 of size $n/2$ and define $P_{W_1} = Q_{W_1}Q_{W_1}^T$, $P_{W_2} = Q_{W_2}Q_{W_2}^T$, where Q_{W_1} are the top k left singular vectors of W_1 and Q_{W_2} are the top k left singular vectors of W_2.
3. Compute $\hat{W} = [P_{W_2}W_1 | P_{W_1}W_2]$
4. Run the version of k-means described in our paper directly on the columns of the matrix $\hat{W}$ to recover the k clusters.

Analysis

The analysis of this algorithm closely follows [9]. Following McSherry we will analyze the algorithm under the assumption that each of the k clusters is exactly bisected in W_1 and W_2. It is straightforward to show that each cluster is approximately bisected with high probability for k constant and large n. Although it is possible to modify the analysis for the more realistic approximate bisection case (see [9] for a discussion), the assumption that the clusters are exactly bisected eases the analysis considerably. We derive a modified version of Theorem 12 of [9].

Theorem 24 *With probability at least $1-\delta$, we have for all u*

$$||A_u - \hat{W}_u||_2 \leq \gamma_1 + \gamma_2$$

where

$$\gamma_1 \leq C_1\sigma\sqrt{nk/s}$$

and

$$\gamma_2 \leq C_2\sigma\sqrt{k\log(n/\delta)}$$

where s is a lower bound on the cluster size.

First, we consider the implications of the theorem and then discuss its proof. When we have gap γ for any two points u, v in different clusters we have

$$||A_u - A_v||_2 \geq \sqrt{2s}\gamma$$

In particular, if we have

$$\sqrt{2s}\gamma \geq 4C_1\sigma\sqrt{nk/s} + 4C_2\sigma\sqrt{k\log(n/\delta)}$$

the algorithm described succeeds, since it is straightforward to see that every column in $\hat{W}$ is closer to every other column in its own cluster than *any* column in any other cluster.

Taking $s = \Theta(n/k)$, we get that if

$$\gamma \geq C_1\frac{\sigma k\sqrt{nk}}{n} + C_2\sigma k\sqrt{\frac{\log(n/\delta)}{n}}$$

for slightly modified constants C_1 and C_2, we succeed in recovering the clusters.

For constant k, the second term dominates and we recover the minimax rate, i.e.

$$\gamma \geq C\sigma\sqrt{\frac{\log(n/\delta)}{n}}$$

suffices. Note that the algorithm is not minimax in its dependence on k unlike the combinatorial procedure. Also, unlike the combinatorial algorithm and our own analysis of spectral clustering the analysis here relies crucially on the fact that the true matrix A is block constant (and thus rank-k).

We now prove the Theorem.

The proof will show for any u,

$$||P_{W_1}W_{2u} - A_{2u}||_2 \leq \gamma_1$$

and

$$||P_{W_1}(A_{2u} - W_{2u})||_2 \leq \gamma_2$$

where the subscript u denotes the u^{th} column of the matrix.

Combining, these two with the identical proof for the other partition, and using triangle inequality we will arrive at the final theorem.

Consider,

$$||(I-P_{W_1})A_2||_2 = ||(I-P_{W_1})A_1||_2 = ||(I-P_{W_1})W_1-(I-P_{W_1})(W_1-A_1)||_2 \leq 2||W_1-A_1||_2$$

The first equality follows because by our exact bisection assumption A_1 and A_2 can be taken to be identical. The inequality follows from two observations.

$$||(I - P_{W_1})W_1||_2 = ||W_1 - P_{W_1}W_1||_2 \leq ||W_1 - A_1||_2$$

which holds since the left side of the inequality is the $k + 1^{\text{th}}$ eigenvalue of W_1 and A_1 is a rank-k matrix. The second observation is that

$$||(I - P_{W_1})(W_1 - A_1)||_2 \leq ||I - P_{W_1}||_2||W_1 - A_1||_2 \leq ||W_1 - A_1||_2$$

since P_{W_1} is a projection matrix all of its eigenvalues are positive and bounded by 1.

Now, note that $(I - P_{W_1})A_2$ is of rank at most $2k$ and for any column u there are at least $s/2$ identical columns in $(I - P_{W_1})A_2$. From this we get that for any u,

$$||A_{2u} - P_{W_1}A_{2u}|| \leq \frac{||(I - P_{W_1})A_2||_F}{\sqrt{s/2}} \leq 4\sqrt{\frac{k}{s}}||W_1 - A_1||_2 \leq C_1\sigma\sqrt{\frac{nk}{s}} \equiv \gamma_1$$

with probability at least $\delta/2$ using the operator norm bound on $W_1 - A_1$.

Now,

$$||P_{W_1}(A_{2u} - W_{2u})||_2 = \sqrt{\sum_{j=1}^{k}\left((A_{2u} - W_{2u})^T P_{W_1 j}\right)^2}$$

Noting that $P_{W_1 j}$ is a unit vector independent of $(A_{2u} - W_{2u})$, each term in this sum is sub-Gaussian with scale factor at most σ. To make a guarantee for any u we will also combine this with a union bound.

From this, a calculation shows that with probability at least $1 - \delta$ we have,

$$||P_{W_1}(A_{2u} - W_{2u})||_2 \leq \sqrt{k}t$$

where $t = C_2\sigma\sqrt{\log(n/\delta)}$. This is just γ_2.

E Examples of worst case behavior

Here we demonstrate the undesireable spectral properties of both the combinatorial and normalized laplacians, in addition to the adjacency matrix. We use concrete examples of similarity matrices whose second eigenvector does not immediately produce the correct clustering. Additionally, we motivate our Range Restriction, by showing that if this condition is not satisfied, the entries of the eigenvector decay at $O(\frac{1}{n})$ instead of of $O(\sqrt{\frac{1}{n}})$.

First, we turn to the drawbacks of using the spectrum of the adjacency matrix. McSherry [9] shows that in the planted partition model, the eigenvectors of the adjacency matrix are enough to identify the clusters. However, in the more general HBM, this is not the case. Consider a matrix with small off-diagonal entries, larger entries on the diagonal blocks, and 2 very high entries in this block (See Figure 6(a)). This is an ideal matrix and the second eigenvector of the combinatorial Laplacian exactly identifies the true clustering, yet the eigenvector of the adjacency matrix fails to convey any meaningful information (See Figure 6(e)).

The normalized Laplacian can also fail to identify the clusters of an ideal hierarchical matrix. For example, on a similarity matrix like the one in Figure 6(b), the second eigenvector of the normalized laplacian identifies the clustering at the second level of the hierarchy rather than the first, as shown in Figure 6(f). We conjecture that different conditions will guarantee that correctness of a spectral method using the normalized laplacian, but we instead focus on the combinatorial Laplacian and our definition of ideal matrices.

The combinatorial Laplacian also has its shortcomings, most notably that it is highly influenced by outliers in the data. If even one data point disrupts the structure of the matrix, as in Figure 6(c), the second eigenvector of the combinatorial Laplacian becomes highly spiked and it can no longer tolerate even small perturbations (see Figure 6(g)).

A related example demonstrates the necessity of the Assumption 3. Consider the matrix shown in Figure 6(d), which is an ideal matrix that violates the range restriction. In this case, the eigenvector again becomes highly spiked (Figure 6(h)), and moreover, the entries decay at a rate of $O(1/n)$ (not shown), which is too sharp for our results to hold.

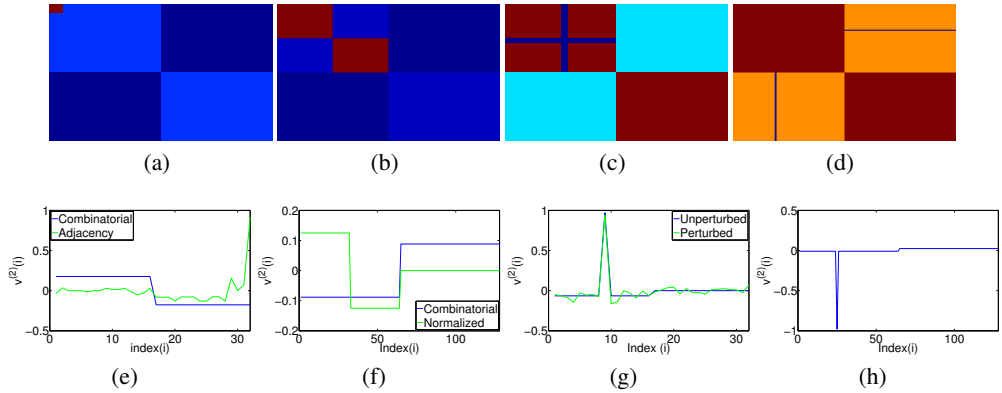


Figure 6: Example similarity matrices that result in undesireable behavior for Normalized Laplacians and Adjacency Matrices and Combinatorial Laplacians.