Changepoint Detection over Graphs with the Spectral Scan Statistic

James Sharpnack
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
jsharpna@cs.cmu.edu

Alessandro Rinaldo
Statistics Department
Carnegie Mellon University
Pittsburgh, PA 15213
arinaldo@cmu.edu

Aarti Singh
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
aarti@cs.cmu.edu

Abstract

We consider the change-point detection problem of deciding, based on noisy measurements, whether an unknown signal over a given graph is constant or is instead piecewise constant over two induced subgraphs of relatively low cut size. We analyze the corresponding generalized likelihood ratio (GLR) statistic and relate it to the problem of finding a sparsest cut in a graph. We develop a tractable relaxation of the GLR statistic based on the combinatorial Laplacian of the graph, which we call the spectral scan statistic, and analyze its properties. We show how its performance as a testing procedure depends directly on the spectrum of the graph, and use this result to explicitly derive its asymptotic properties on few graph topologies. Finally, we demonstrate both theoretically and by simulations that the spectral scan statistic can outperform naive testing procedures based on edge thresholding and χ^2 testing.

1 Introduction

In this article we are concerned with the basic but fundamental task of deciding whether a given graph, over which a noisy signal is observed, contains a cluster of anomalous or activated nodes comprising an induced subgraph. Such a problem is highly relevant in a variety of scientific areas, such as surveillance, disease outbreak detection, biomedical imaging, sensor network detection, gene network analysis, environmental monitoring and malware detection over a computer network. Recent theoretical contributions in the statistical literature (see, e.g., [4, 3, 2, 1]) have detailed the inherent difficulty of such testing problems in relatively simplified settings and under specific conditions on the graph topology. A natural algorithm for detection of anomalous clusters of activity in graphs is the generalized likelihood ratio test (GLRT) or scan statistic, a computationally intensive procedure that entails scanning all clusters in our class for anomalous activation. Unfortunately, its performance over general graphs is not well understood, and little attention has been paid to determining alternative, computationally tractable, procedures.

In this article we assume that the class of clusters of constant signal consists of sub-graphs of small cut size. We believe this is a natural and realistic assumption which, as we demonstrate below, allows us to explicitly incorporate into the detection problem the properties of the graph topology through its spectrum. In particular, we show that the GLRT is an integer program with a term in the objective that corresponds to the sparsest cut in a graph, a known NP-hard problem [27]. With this in mind, we propose a relaxation of the GLRT, called the spectral scan statistic, which is based on the combinatorial Laplacian of the graph and, importantly, is a computationally efficient program. As our main result, we derive theoretical guarantees for the performance of the spectral scan statistic, that hold for any graph and are based on the spectrum of the combinatorial Laplacian. For comparison purposes, we derive theoretical guarantees for two simple estimators, the edge thresholding and the χ^2 test. We conclude our study by applying the main result to balanced binary trees, the lattice, and Kronecker graphs, giving precise asymptotic results. Simulations for these models verify that the spectral scan statistic dominates the simple estimators. Before we elaborate on the statistical setup, we will examine two real-world examples of graph structured signals with low cut size.

Disease Detection in Human Networks. Many

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

common experimental techniques in virology report various indicators of a virus, such as antibody protein concentrations (western blot, enzyme-linked immunosorbent assay) or measuring virus concentrations directly (the plaque assay). One popular method, the western blot [8], reports concentrations by the shade of bands from an x-ray film darkened by a luminescent compound. Infectious diseases diffuse within human networks, so we can exploit this network structure in the detection of infectious diseases, then we may be able to detect and localize an incipient infection under low signal-to-noise ratios (very light bands in the western blot).

Sensor Networks. Sensor networks might be deployed for detecting nuclear substances, water contaminants, or activity in video surveillance. Water supply contamination is a common cause for outbreaks of cholera, gastroenteritis, E. coli, and polio. The design of sensor networks for water supply was the subject of an engineering challenge in [31]. Because of the potential for large scale health problems, it is of interest to detect contaminated water under low signal-to-noise regimes. As we will see, by exploiting the graph structure (in this case, the pipe network for the water supply), one can detect activity in networks when the activity is very faint. Furthermore, the graph structure provides a versatile framework for modeling environmental constraints.

Contributions. Our contributions are as follows. (1) We define a new class of signals based on the notion of small cut size that reflects in a natural way the topological properties of the graph. (2) We analyze the corresponding GLR statistic and show that it is, in fact, related to the problem of finding the sparsest cut. We then develop a computationally efficient relaxation of the GLR statistic, called the spectral scan statistic and analyze its properties. In our main theoretical result, we show that the performance of the spectral scan statistic depends explicitly on the spectral properties of the graph. (3) Using such results we are able to characterize in a very explicit form the performance of the spectral scan statistic on a few notable graph topologies and demonstrate its superiority over naive detectors, such as the edge thresholding and the χ^2 test. (4) Finally, we have formulated the detection problem under more general and realistic scenarios, which involve composite null and alternative hypotheses as opposed to simple hypotheses as is customary in the theoretical statistical literature on this subject.

Related Work. Normal means testing in high-dimensions is a well established and fundamental problem in statistics (see, e.g., [19]). A significant portion of the recent work in this area, [4, 3, 2, 1], has focused on incorporating structural assumptions on the signal, as a way to mitigate the effect of high-dimensionality and also because many real-life problems can be represented as instances of the normal means problem with graph-structured signals (see, for an example, [20]). These contributions have considered the GLRT when the alternative hypothesis takes on the form of a combinatorial space. However, the performance of such test has been analyzed only for certain types of graphs, and it is unclear to what extent those analyses extend to general graph topologies. Moreover, while much is known about the theoretical performance of the GLRT, little attention is paid to its computational feasibility. Another line of research relevant to our problem is the optimal fail detection with nuisance parameters and matched subspace detection in the signal processing literature (see, e.g. [34, 6, 16, 15]). Though our problem can be cast as a special case of the more general problem of optimal testing of a linear subspace under nuisance parameters considered in that line of work, the focus on a graph-structured signal, as well as the type of analysis based on the interplay between the scan statistic and the spectral properties of the graph contained in our work, are novel.

1.1 Problem Setup

In this section, we formalize the problem of detecting a change of signal from a single set of noisy observations recorded at the vertices of the graph. For a given connected, undirected, possibly weighted large graph $G = (V, E, \mathbf{W})$ on $|V| = n$ nodes, we observe *one* realization of the random vector

$$\mathbf{y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$, with σ^2 known. We will assume that there are two groups of constant signal for $\boldsymbol{\beta}$, namely that there exists a subset $C \subset V$ such that $\boldsymbol{\beta}$ is constant within both C and it complement $\bar{C} = V \backslash C$. We formalize this assumption by writing

$$\boldsymbol{\beta} = \mu \mathbf{1} + \delta \mathbf{1}_C, \tag{2}$$

where $\mu, \delta \in \mathbb{R}$ are unknown parameters, $\mathbf{1} \in \mathbb{R}^n$ is a n-dimensional vector of ones and $\mathbf{1}_C$ is the indicator function of the subset C. The parameter μ can be thought of as the magnitude of the background signal and is a nuisance parameter, while δ quantifies the the gap in signal between the two clusters. Setting $\bar{\boldsymbol{\beta}} = \mathbf{1}^\top \boldsymbol{\beta}/n$, we will use $\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|$ to measure the energy of the signal (note that this quantity is independent of μ) where $\|.\|$ always denotes the ℓ_2 norm. We will define the signal-to-noise ratio (SNR) to be

$$\frac{\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|}{\sigma} = \sqrt{\frac{|C||\bar{C}|}{n}} \frac{\delta}{\sigma}.$$

We will not assume any knowledge of the true clustering $(C, \bar{C})$, other than that it belongs to a given class $\mathcal{C}$ of bi-partitions $(C, \bar{C})$ of V such that C and $\bar{C}$ are both large and have low cut size. Formally, we define, for some $\rho > 0$,

$$\mathcal{C} = \mathcal{C}(\rho) = \left\{ C \subset V, C \neq \emptyset \colon \frac{|\partial C|}{|C||\bar{C}|} \leq \frac{\rho}{n} \right\}, \qquad (3)$$

where $\partial C = \{(i,j) \in E : i \in C, j \in \bar{C}\}$ is the boundary of C. Note that $\mathcal{C}$ is a symmetric class in the sense that $C \in \mathcal{C}$ if and only if $\bar{C} \in \mathcal{C}$. We are interested in the problem of testing whether the gap parameter δ in equation (2) is zero (i.e. the signal $\boldsymbol{\beta}$ is constant) or it is non-zero for some $C \in \mathcal{C}$, regardless of the value of μ. Thus, we can naturally cast our structured change-point detection problem as the following composite hypothesis testing problem:

$$H_0 \colon \boldsymbol{\beta} \in \Theta_0 \quad \text{vs} \quad H_1 \colon \boldsymbol{\beta} \in \Theta_1, \qquad (4)$$

where $\Theta_0 = \{\mu \mathbf{1}, \mu \in \mathbb{R}\}$ and $\Theta_1 = \{\mathbf{1}\mu + \mathbf{1}_C \delta, \mu \in \mathbb{R}, \delta \in \mathbb{R} \setminus \{0\}, C \in \mathcal{C}\}$. Notice that the alternative can be written as the union of alternatives of the form $H_1^C \colon \boldsymbol{\beta} \in \Theta_1^C := \{\mathbf{1}\mu + \mathbf{1}_C \delta, \mu \in \mathbb{R}, \delta \in \mathbb{R} \setminus \{0\}\}$, $C \in \mathcal{C}$. Notice that it is not required that C is a connected set of vertices.

To make our analysis meaningful, we measure the difficulty of the detection problem in terms of the energy parameter by assuming that, for some $\eta > 0$, $\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\| \geq \eta, \quad \forall \boldsymbol{\beta} \in \Theta_1$. Thus, we can think of η as the minimal degree of separation between the null and alternative hypotheses. Below we will analyze asymptotic conditions under which the hypothesis testing problem described above is feasible, in a sense made precise in the next definition, when the size of the graph n increases. To this end, we will further assume that the relevant parameters of the model, η, σ, δ and ρ change with n as well, even though we will not make such dependence explicit in our notation for ease of readability. Our results establish conditions for asymptotic disinguishability as a function of the SNR η/σ and ρ and the spectrum of the graph G.

Definition 1. *Let P_β denote the distribution of $\mathbf{y}$ induced by the model (1), where $\beta \in \Theta_0 \cup \Theta_1$. For a given statistic $S(\mathbf{y})$ and threshold $\tau \in \mathbb{R}$, let $T = T(\mathbf{y})$ be 1 if $S(\mathbf{y}) > \tau$ and 0 otherwise. We say that the hypotheses H_0 and H_1 are* **asymptotically distinguished by the test** *T if*

$$\sup_{\beta \in H_0} \mathbb{P}_\beta\{T = 1\} \to 0 \quad \text{and} \quad \sup_{\beta \in H_1} \mathbb{P}_\beta\{T = 0\} \to 0, \qquad (5)$$

where the limit is taken as $n \to \infty$. We say that H_0 and H_1 are **asymptotically indistinguishable** *if there does not exist any test for which the above limits hold.*

Notation. We will need some mathematical terminology from algebraic graph theory ([17]). A central object to our analysis is the *combinatorial Laplacian* matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{W}$ is the weight matrix of the graph G and $\mathbf{D} = \text{diag}\{d_v\}_{v \in V}$ is the diagonal matrix of node degrees, $d_v = \sum_{w \in V} W_{v,w}$, $v \in V$. If the graph is weighted then $W_{v,w}$ reflects this. We will denote the eigenvalues of $\mathbf{L}$ with $\{\lambda_i\}_{i=1}^n$, which we will always take in increasing order. Since G is connected, the smaller eigenvalue $\lambda_1 = 0$, with corresponding eigenvector, $\mathbf{1}$. λ_2 is known as the *algebraic connectivity*, which is known to provide bounds for the minimum cut sparsity via Cheeger's inequality. Throughout this study we use Bachmann-Landau notation for asymptotic statements: if $a_n/b_n \to 0$ then $a_n = o(b_n)$ and $b_n = \omega(a_n)$. If $a_n/b_n \to c$ for some $c > 0$ then we write $a_n \asymp b_n$. When $\mathbf{y} \in \mathbb{R}^n$ is a vector then $\bar{\mathbf{y}} = \frac{\mathbf{1}}{n} \sum_{i=1}^n y_i$, but for a set $C \subseteq V$ then we define $\bar{C} = V \backslash C$.

2 Methods

The hypothesis testing problem at hand presents two challenges: (1) the model contains a nuisance parameter $\mu \in \mathbb{R}$ and (2) the alternative hypothesis is comprised of a union of hypotheses indexed by $C \in \mathcal{C}$. The existence of the nuisance paramter sets our problem further apart from existing work of structured normal means problems (see, e.g. [4, 3, 2, 1]), which relies on a simplified framework consisting of a simple null hypothesis and a composite hypothesis consisting of unions of simple alternatives. We will eliminate the interference caused by the nuisance parameter by considering test procedures that are independent of μ. The formal justification for this choice is based on the theory of optimal invariant hypothesis testing (see, e.g. [23]) and of uniformly best constant power tests (see [39]). Due to space limitations we will not provide the details and refer the reader to [15, 16, 14, 13, 34, 6] and references therein for in depth-treatments of these issues related to the model at hand.

For the simpler problem of testing H_0 versus H_1^C for some $C \subset V$, the optimal test is based on the likelihood ratio (LR) statistic (see the proof of Lemma 2 below for a derivation)

$$\begin{aligned} 2 \log \Lambda_C(\mathbf{y}) &= 2 \log \left(\frac{\sup_{\boldsymbol{\beta} \in \Theta_1} f_{\boldsymbol{\beta}}(\mathbf{y})}{\sup_{\boldsymbol{\beta} \in \Theta_0} f_{\boldsymbol{\beta}}(\mathbf{y})} \right) \\ &= \frac{1}{\sigma^2} \frac{|V|}{|C||\bar{C}|} \left(\sum_{v \in C} \tilde{\mathbf{y}}_v \right)^2, \end{aligned} \qquad (6)$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}}$ and $f_{\boldsymbol{\beta}}$ is the Lebesgue density of $P_{\boldsymbol{\beta}}$. This test rejects H_0 for large values of $\Lambda_C(\mathbf{y})$. Optimality follows from the fact that the statistical

model we consider has the monotone likelihood ratio property.

When testing against composite alternatives, like in our case, it is customary to consider instead the generalized likelihood ratio (GLR) or scan statistic, which in our case reduces to

$$\hat{g} = \max_{C \in \mathcal{C}(\rho)} 2\sigma^2 \log \Lambda_C(\mathbf{y}).$$

Through manipulations of the likelihoods, we find that the GLR statistic has a very convenient form which is tied to the spectral properties of the graph G via its Laplacian.

Lemma 2. *Let* $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{1}(\frac{1}{n}\sum_{v\in V}\mathbf{y}_v)$ *and* $\mathbf{K} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. *Then the GLR statistic is*

$$\hat{g} = \max_{\mathbf{x}\in\{0,1\}^n} \frac{\mathbf{x}^\top \tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{K}\mathbf{x}} \; s.t. \; \frac{\mathbf{x}^\top \mathbf{L}\mathbf{x}}{\mathbf{x}^\top \mathbf{K}\mathbf{x}} \leq \rho, \qquad (7)$$

where $\mathbf{L}$ *is the combinatorial Laplacian of the graph* G.

The proof is provided in the appendix. The savvy reader will notice the connection between the graph constrained scan statistic (7) and the graph sparsest cut program. By Lagrangian duality, we see that the program (7) is equivalent to (for some Lagrangian parameter ν)

$$\min_{C\subseteq V} \frac{|\partial C|}{|C||\bar{C}|} - \nu \frac{(\sum_{i\in C}\tilde{y}_i)^2}{|C||\bar{C}|}$$

the first term of which is precisely the *sparsest cut* objective, and the second term drives the solution C to have positive within cluster empirical correlations. The sparsest cut program is known to be NP-hard, with poly-time algorithms known for trees and planar graphs [27]. Because of this fact, approximate algorithms have been proposed over the past two decades, most notably the uniform multicommodity flow approach of [24, 37] and the semi-definite relaxation of the cut metric [5]. [18] observed that the minimum cut sparsity is bounded by the algebraic connectivity (λ_2), suggesting the Fiedler vector (i.e. the second eignenvector of $\mathbf{L}$) to be an appropriate relaxation of the characteristic vector of the cut. Moreover, the well known Cheeger inequality shows that the minimum cut sparsity (in a regular graph) is bounded by the algebraic connectivity (see [9]). We will follow the tradition of bounding sparsity with the algebraic connectivity, and provide a surrogate estimator to the scan statistic based on this simple spectral relaxation.

Proposition 3. *Define the Spectral Scan Statistic (SSS) as*

$$\hat{s} = \sup_{\mathbf{x}\in\mathbb{R}^n} (\mathbf{x}^\top \tilde{\mathbf{y}})^2 \; s.t. \; \mathbf{x}^\top \mathbf{L}\mathbf{x} \leq \rho, \|\mathbf{x}\| \leq 1, \mathbf{x}^\top \mathbf{1} = 0.$$

Then the GLR statistic is bounded by the SSS: $\hat{g} \leq \hat{s}$.

Proof. First let us notice that $\mathbf{K} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the projection onto the subspace orthogonal to $\mathbf{1}$. Because $\mathbf{K}$ is thus idempotent, $\tilde{\mathbf{y}}\mathbf{1} = 0$, and since $\mathbf{L}\mathbf{1} = 0$ we can rewrite

$$\hat{g} = \max_{\mathbf{x}\in\{0,1\}^n\setminus\{\mathbf{0},\mathbf{1}\}} \frac{(\mathbf{K}\mathbf{x})^\top \tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top (\mathbf{K}\mathbf{x})}{(\mathbf{K}\mathbf{x})^\top (\mathbf{K}\mathbf{x})} \quad \text{s.t.} \; \frac{(\mathbf{K}\mathbf{x})^\top \mathbf{L}(\mathbf{K}\mathbf{x})}{(\mathbf{K}\mathbf{x})^\top (\mathbf{K}\mathbf{x})} \leq \rho$$

So, we have the following relaxation,

$$\hat{g} \leq \max_{\mathbf{x}\neq 0, \mathbf{x}^\top\mathbf{1}=0} \frac{\mathbf{x}^\top \tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \text{ s.t. } \frac{\mathbf{x}^\top \mathbf{L}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq \rho = \hat{s}$$

□

Notice that because the domain $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{L}\mathbf{x} \leq \rho, \|\mathbf{x}\| \leq 1, \mathbf{x}^\top \mathbf{1} = 0\}$ is symmetric around the origin, this is precisely the square of the solution to

$$\sqrt{\hat{s}} = \sup_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^\top \mathbf{y} \text{ s.t. } \mathbf{x}^\top \mathbf{L}\mathbf{x} \leq \rho, \|\mathbf{x}\| \leq 1, \mathbf{x}^\top \mathbf{1} = 0, \qquad (8)$$

where we have used the fact that $\mathbf{x}^\top \tilde{\mathbf{y}} = ((\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{x})^\top \mathbf{y} = \mathbf{x}^\top \mathbf{y}$ because $\mathbf{x}^\top \mathbf{1} = 0$ within $\mathcal{X}$.

Remark 4. *Through a reparametrization we can show that the program* (8) *has a linear objective and only quadratic constraints. After forming the Lagrangian we can show that this is equivalent to*

$$\inf_{\nu_0,\nu_1\geq 0} \nu_0\rho + \nu_1 + \frac{1}{4}\tilde{\mathbf{y}}^\top [\nu_0\mathbf{L} + \nu_1\mathbf{I}]^{-1}\tilde{\mathbf{y}}$$

which can be solved by first order interior point methods over the parameters ν_0, ν_1 *where the gradient calculation requires the solution to a linear system. Furthermore, the linear systems are semidefinite, diagonally dominant, hence by the recent work of [21], has a running time of* $O(|E|\log n)$ *modulo logarithmic precision factors.*

The formulation in (8) shows that the SSS is related to the supremum of a Gaussian process over $\mathcal{X}$. This fact will turn out to be extremely convenient, as we show next.

3 Theoretical Analysis

We first derive a simple condition for asymptotic indistinguishability based on testing the null versus a single component in the alternative. A more refined analysis of the lower bound for the general hypothesis (4) is beyond the scope of this article. Recall that, under alternative hypothesis, $\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\| \geq \eta$ uniformly over Θ_1.

Theorem 5. *(1) H_0 and H_1 are asymptotically indistinguishable if $\eta/\sigma = o(1)$.*
(2) Suppose that there is a subset of clusters $\mathcal{C}' \subseteq 2^V$ such that all the elements of $\mathcal{C}'$ are disjoint, of the same size ($|C| = c$ for all $C \in \mathcal{C}'$), and

$$\forall C \in \mathcal{C}', \quad \frac{n|\partial C|}{|C||\bar{C}|} \leq \frac{\rho}{2}$$

i.e., elements of $\mathcal{C}'$ belong to the alternative hypothesis with $\rho/2$ cut sparsity. Furthermore assume that $\frac{c|\mathcal{C}'|}{n} \to 1$. Consider the observation model (1), and the testing problem given by (4). Then H_0 and H_1 are asymptotically indistinguishable if

$$\frac{\eta}{\sigma} = o(|\mathcal{C}'|^{1/4})$$

The proof is in the appendix. We will analyze the performance of the SSS statistic by relying on its representation (8) as the square of the supremum of a Gaussian process. We draw heavily on the theory of the generic chaining, perfected in [38], which essentially reduces the problem of computing bounds on the expected supremum of Gaussian processes to geometric properties of its index space.

Theorem 6. *The following hold with probability at least $1 - \delta$. Under the null H_0,*

$$\hat{s} \leq \left(\sqrt{2\sigma^2 \sum_{i>1} \min\{1, \rho\lambda_i^{-1}\}} + \sqrt{2\sigma^2 \log \frac{2}{\delta}} \right)^2,$$

while under the alternative H_1,

$$\hat{s} \geq \left(\eta - \sqrt{2\sigma^2 \log \frac{2}{\delta}} \right)^2.$$

Proof. For a detailed proof, please see the appendix. We use generic chaining to control the process $\{\mathbf{x}^\top \mathbf{y}\}_{\mathbf{x} \in \mathcal{X}}$ appearing in the SSS. First, we notice that the index set $\mathcal{X}$ is the intersection of an ellipsoid and the unit ball, which is the intuition behind the following lemma.

Lemma 7. *Let $\mathbf{L}$ have spectrum $\{\lambda_i\}_{i=1}^n$. Then under H_0,*

$$\mathbb{E} \sup_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \mathbf{y} \leq \sqrt{2\sigma^2 \sum_{i>1} \min\{1, \rho\lambda_i^{-1}\}}.$$

The proof is provided in the appendix and is a direct result of Lemma 14 from [22]. We can then use the well known phenomena, that the supremum of a Gaussian process concentrates around it's expectation. Hence, by Lemma 14 the first statement in Theorem 6 holds. The second statement follows by applying standard concentration results to the univariate Gaussian $\frac{\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}}{\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|} \mathbf{y}$ and noticing that $\frac{\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}}{\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|} \in \mathcal{X}$ and $\mathbb{E} \frac{(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top}{\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|} \mathbf{y} = \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\| \geq \eta$ under H_1. □

As a corollary we will provide sufficient conditions for asymptotic distinguishability that depend on the spectrum of the Laplacian $\mathbf{L}$. As we will show in the next section, these conditions can be applied to a number of graph topologies whose spectral properties are known.

Corollary 8. *The null and alternative, as described in Thm. 6, are asymptotically distinguished by the SSS, $\hat{s}$, and the GLRT, $\hat{g}$, if*

$$\frac{\eta}{\sigma} = \omega \left(\sqrt{\sum_{i>1} \min\{1, \rho\lambda_i^{-1}\}} \right) \tag{9}$$

Other stronger sufficient conditions are

$$\frac{\eta}{\sigma} = \omega \left(\sqrt{k + \frac{(n-k)\rho}{\lambda_{k+1}}} \right) \tag{10}$$

if k is large enough that $\lambda_{k+1} > \rho$.

Proof. To see equation (9) we note that, due to Theorem 6, if

$$\sqrt{2\sigma^2 \sum_{i>1} \min\{1, \rho\lambda_i^{-1}\}} + \sqrt{2\sigma^2 \log \frac{2}{\delta}} = o\left(\eta - \sqrt{2\sigma^2 \log \frac{2}{\delta}} \right)$$

then we attain asymptotic distinguishability by choosing any threshold τ between, and sufficiently far from, the left and right hand side of the previous display. To show equation (10) we note that by choosing k such that $\lambda_{k+1} > \rho$ we see that

$$\sum_{1<i\leq k} \min\{1, \rho\lambda_i^{-1}\} \leq k$$
$$\Rightarrow \sum_{i>k} \min\{1, \rho\lambda_i^{-1}\} \leq (n-k)\frac{\rho}{\lambda_{k+1}}.$$

□

Interestingly, there are no logarithmic terms in (9) that usually accompany uniform bounds of this type, which is attributed to the generic chaining.

For comparison, we consider the performance of two naive procedures for detection: the energy detector, which reject H_0 if $\|\tilde{\mathbf{y}}\|^2$ is too large and the edge thresholding detector, which reject H_0 if $\max_{(v,w)\in E} |\mathbf{y}_v - \mathbf{y}_w|$ is large. The following is a classical result that can be found in [19].

Theorem 9. *H_0 and H_1 are asymptotically distinguished by $\|\tilde{\mathbf{y}}\|$ if*

$$\frac{\eta}{\sigma} = \omega(n^{1/4}).$$

while $\|\tilde{\mathbf{y}}\|$ fails to asymptotically distinguish H_0 from H_1 if

$$\frac{\eta}{\sigma} = o(n^{1/4})$$

In [35] the authors examined the problem of exact recovery of cluster boundaries in the graph-structured normal means problem by taking differences between observations corresponding to adjacent nodes. The following result stems from Theorem 2.1 of [35], and the fact that $|C||\bar{C}|/n$ scales like $\min\{|C|, |\bar{C}|\}$ up to a factor of 2.

Theorem 10. *H_0 and H_1 are asymptotically distinguished by $\max_{(v,w)\in E} |\mathbf{y}_v - \mathbf{y}_w|$ if*

$$\frac{\eta}{\sigma} = \omega\left(\sqrt{\max_{C\in\mathcal{C}, |C|\leq n/2} |C| \log n}\right).$$

Hence, if $\max_{C\in\mathcal{C},|C|\leq n/2} |C|$ is large then the edge thresholding statistic may be dominated by the SSS, because the bound in Corollary 8 is always smaller than $\sqrt{n}$.

4 Specific Graph Models

In this section we demonstrate the power and flexibility of Theorem 6 by analyzing in detail the performance of the spectral scan statistic over three graph topologies: balanced binary trees, the 2 dimensional lattice and the Kronecker graphs (see [26, 25]).

4.1 Balanced Binary Trees

We begin the analysis of the spectral scan statistic by applying it to the balanced binary tree (BBT) of depth ℓ. The class of signals that we will consider have clusters of constant signal which are subtrees of size at least cn^α for $0 < c \leq 1/2, 0 < \alpha \leq 1$. Hence, the cut size of the signals are 1 and $\rho = [cn^\alpha(1 - cn^{\alpha-1})]^{-1}$.

Corollary 11. *Let G be the balanced binary tree with n vertices, and $\rho = n[cn^\alpha(n - cn^\alpha)]^{-1}$.*
(a) The spectral scan statistic can asymptotically distinguish H_0 from H_1 if the SNR satisfies

$$\frac{\eta}{\sigma} = \omega(n^{\frac{1-\alpha}{2}} \log n).$$

(b) H_0 and H_1 are asymptotically indistinguishable if

$$\frac{\eta}{\sigma} = o(n^{\frac{1-\alpha}{4}}).$$

The lower bound in part (b) is a direct result of Theorem 5 (b). This result shows that when α is near to 1 then there is little gap between the upper bound of the SSS and the lower bound. To illustrate our claim, we simulate the probability of correct discovery of changepoints (rejecting H_0 when the truth is H_1) versus the probability of false alarm (falsely rejecting H_0). We compare the following estimators: the energy statistic, edge differencing, the SSS, and the unconstrained GLRT. The unconstrained GLRT is formed by choosing the cluster C without the constraint $C \in \mathcal{C}$, which is formed by merely ordering the elements of $\mathbf{y}$ and greedily adding the components to C until the RHS of (7) is maximized. These are given for the four estimators in Figure 1 and for the SSS as $n = 2^{\ell+1} - 1$ increases. In these simulations a subtree at level 2 (of size $n/4$) was chosen as C, the gap-to-noise ratio is fixed at $\delta/\sigma = 0.8$, and $\rho = 4/n$. We see that even in the low n regime, exploiting the graph structure is essential to improve the power of testing H_0 against H_1. As n increases with δ/σ fixed the performance of the SSS dramatically increases.

4.2 Lattice

We will analyze the performance guarantees of the SSS over the 2-dimensional lattice graph with p vertices along each dimension ($n = p^2$).

Corollary 12. *Let G be the $p \times p$ square lattice ($n = p^2$), and let $\rho = Cn^{-(1-\alpha)/2}$ for $\alpha \in [0, 1)$.*
(a) The spectral scan statistic can asymptotically distinguish H_0 from H_1 if the SNR satisfies

$$\frac{\eta}{\sigma} = \omega(n^{\frac{1+\alpha}{4}} \sqrt{\log n})$$

(b) H_0 and H_1 are asymptotically indistinguishable if the SNR is weaker than

$$\frac{\eta}{\sigma} = o(n^{\frac{\alpha}{4}})$$

The proof of (a) is in the appendix. Unfortunately, the upper bound in (a) is larger than that provided for the energy statistic in Theorem 9. Our experiments (Figure 1) suggest though that these upper bounds can be greatly improved. The lower bound, Corollary 12 (b), holds because we can form $\mathcal{C}'$ of Theorem 5 (b) from disjoint squares of size a constant multiple of $n^{1-\alpha}$ making $|\mathcal{C}'| \asymp n^\alpha$. We demonstrate the improvement of the SSS over competing tests in Figure 1. In these simulations a $\sqrt{n}/2 \times \sqrt{n}/2$ square was chosen to be C with $\rho = 4/\sqrt{n}$. Despite the weaker guarantee in Corollary 12 the SSS demonstrates the importance of exploiting the graph structure.

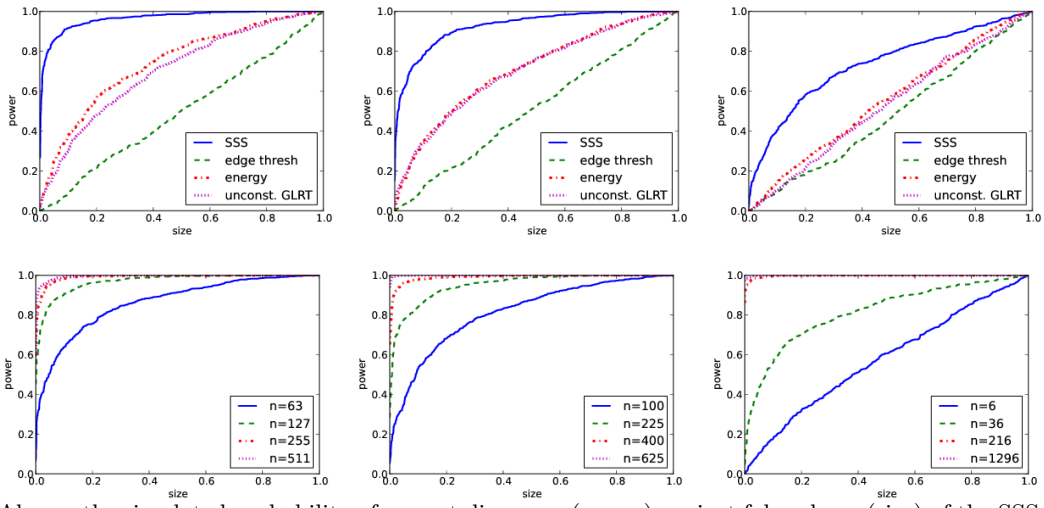

Figure 1: Above: the simulated probability of correct discovery (power) against false alarm (size) of the SSS compared to the energy detector, edge thresholding and the unconstrained GLRT of the BBT (left), Lattice (middle), and Kronecker graph (right). Below: the performance of the SSS as n increases.

4.3 Kronecker Graphs

Much of the research in complex networks has focused on observing statistical phenomena that are common across many data sources. Most notably, the degree distribution of real world graphs obey a power law [11] and networks are often found to have small diameter [29]. A class of graphs that satisfy these properties, while providing a simple modeling platform, are the Kronecker graphs (see [26, 25]). Let H_1 and H_2 be graphs on p vertices with Laplacians $\mathbf{L}_1, \mathbf{L}_2$ and edge sets E_1, E_2 respectively. The Kronecker product, $H_1 \otimes H_2$, is the graph over vertices $[p] \times [p]$ such that there is an edge $((i_1, i_2), (j_1, j_2))$ if $i_1 = j_1$ and $(i_2, j_2) \in E_2$ or $i_2 = j_2$ and $(i_1, j_1) \in E_1$. We will construct graphs that have a multi-scale topology using the Kronecker product. Let the multiplication of a graph by a scalar indicate that we multiply each edge weight by that scalar. First let H be a connected graph with p vertices. Then the graph G for $\ell > 0$ levels is defined as

$$\frac{1}{p^{\ell-1}} H \otimes \frac{1}{p^{\ell-2}} H \otimes ... \otimes \frac{1}{p} H \otimes H$$

The choice of multipliers ensures that it is easier to make cuts at the more coarse scale. Notice that all of the previous results have held for weighted graphs.

Corollary 13. *Let G be the Kronecker product graph described above with $n = p^\ell$ vertices, and consider only signals with cuts within the k coarsest scale ($\rho \propto p^{2k-\ell-1}$).*
(a) The spectral scan statistic can asymptotically distinguish H_0 from H_1 if the SNR satisfies

$$\frac{\eta}{\sigma} = \omega(p^2(\ell+2)n^{(2k+1)/\ell})$$

(b) H_0 and H_1 are asymptotically indistinguishable if

$$\frac{\eta}{\sigma} = o(n^{\frac{k}{4\ell}}/\sqrt{p})$$

The proof and an explanation of ρ is in the appendix. Again, we demonstrate the improvement of the SSS over competing tests in Figure 1. For these simulations the base graph H was chosen to be two triangles (K_3) connected by a single edge (p = 6). At the coarsest scale one of the K_3 subgraphs was chosen to be C with $\rho = 4/n$.

5 Discussion

We studied the problem of tractably detecting changepoints in networks under Gaussian noise. To this end we developed the spectral scan statistic as a computationally feasible alternative to the generalized likelihood ratio test. We completely characterized the performance of the SSS for any graph in terms of the spectrum of the combinatorial Laplacian. For comparison purposes, we developed theoretical guarantees for two simple estimators. We applied the main result to three graph models: binary balanced trees, the lattice and Kronecker graph. We see that not only is it statistically suboptimal to ignore graph structure, but for coarse cuts in the balanced binary tree and the Kronecker graph the SSS gives near optimal performance. This claim is backed by both simulation and theory.

Acknowledgements

This research is supported in part by AFOSR under grant FA9550-10-1-0382.

References

[1] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

[2] E. Arias-Castro, E. Candes, and A. Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

[3] E. Arias-Castro, E. Candes, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 36(4):1726–1757, 2008.

[4] E. Arias-Castro, D. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory*, 51(7):2402–2425, 2005.

[5] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009.

[6] B. Baygün and A. O. Hero. Optimal simultaneous detection and estimation under a false alarm constraint. *Signal Processing, IEEE Transactions on*, 41(3):688–703, 1995.

[7] C. Borell. The brunn-minkowski inequality in gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975.

[8] W. N. Burnette. western blotting: electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein a. *Analytical biochemistry*, 112(2):195–203, 1981.

[9] F. Chung. Discrete isoperimetric inequalities. *Surveys in Differential Geometry IX, International Press*, pages 53–82, 2004.

[10] B. Cirelson, I. Ibragimov, and V. Sudakov. Norms of gaussian sample functions. In *Proceedings of the Third JapanUSSR Symposium on Probability Theory*, pages 20–41. Springer, 1976.

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, volume 29, pages 251–262. ACM, 1999.

[12] M. Fiedler. Eigenvectors of acyclic matrices. *Czechoslovak Mathematical Journal*, 25(4):607–618, 1975.

[13] L. Fillatre. Asymptotically uniformly minimax detection and isolation in network monitoring. to appear in Signal Processing, IEEE Transactions on.

[14] L. Fillatre and I. Nikiforov. Non-bayesian detection and detectability of anomalies from a few noisy tomographic projections. *Signal Processing, IEEE Transactions on*, 55(2):401–413, 2007.

[15] M. Fouladirad, L. Freitag, and I. Nikiforov. Optimal fault detection with nuisance parameters and a general covariance matrix. *International Journal of Adaptive Control and Signal Processing*, 22(5):431–439, 2008.

[16] M. Fouladirad and I. Nikiforov. Optimal statistical fault detection with nuisance parameters. *Automatica*, 41(7):1157–1171, 2005.

[17] C. Godsil, G. Royle, and C. Godsil. *Algebraic graph theory*, volume 8. Springer New York, 2001.

[18] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 11(9):1074–1085, 1992.

[19] Y. Ingster and I. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Verlag, 2003.

[20] L. Jacob, P. Neuvial, and S. Dudoit. Gains in power from structured two-sample tests of means on graphs. *Arxiv preprint arXiv:1009.5173*, 2010.

[21] I. Koutis, A. Levin, and R. Peng. Faster spectral sparsification and numerical algorithms for sdd matrices. *arXiv preprint arXiv:1209.5821*, 2012.

[22] M. Ledoux. *The concentration of measure phenomenon*, volume 89. Amer Mathematical Society, 2001.

[23] E. Lehmann and J. Romano. *Testing statistical hypotheses*. Springer Verlag, 2005.

[24] T. Leighton and S. Rao. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. In *Foundations of Computer Science, 1988., 29th Annual Symposium on*, pages 422–431. IEEE, 1988.

[25] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

[26] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *Proceedings of the 24th international conference on Machine learning*, pages 497–504. ACM, 2007.

[27] D. Matula and F. Shahrokhi. Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, 27(1):113–123, 1990.

[28] R. Merris. Laplacian graph eigenvectors. *Linear algebra and its applications*, 278(1):221–236, 1998.

[29] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[30] J. Molitierno, M. Neumann, and . SHADER. Tight bounds on the algebraic connectivity of a balanced binary tree. *Electronic Journal of Linear Algebra*, 6:62–71, 2000.

[31] A. Ostfeld, J. G. Uber, E. Salomons, J. W. Berry, W. E. Hart, C. A. Phillips, J.-P. Watson, G. Dorini, P. Jonkergouw, Z. Kapelan, et al. The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*, 134(6):556–568, 2008.

[32] O. Rojo. The spectrum of the laplacian matrix of a balanced binary tree. *Linear algebra and its applications*, 349(1):203–219, 2002.

[33] O. Rojo and R. Soto. The spectra of the adjacency matrix and laplacian matrix for some balanced trees. *Linear algebra and its applications*, 403:97–117, 2005.

[34] L. L. Scharf and B. Friedlander. Matched subspace detectors. *Signal Processing, IEEE Transactions on*, 42(8):2146–2157, 1994.

[35] J. Sharpnack, A. Rinaldo, and A. Singh. Sparsistency of the edge lasso over graphs. *AIStats (JMLR WCP)*, 22:1028–1036, 2012.

[36] J. Sharpnack and A. Singh. Identifying graph-structured activation patterns in networks. In *Proceedings of Neural Information Processing Systems, NIPS*, 2010.

[37] D. Shmoys. Cut problems and their application to divide-and-conquer. *Approximation algorithms for NP-hard problems*, pages 192–235, 1997.

[38] M. Talagrand. *The generic chaining*. Springer, 2005.

[39] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of American Mathematical Society*, 54:426–482, 1943.

A Appendix

A.1 Proofs in Section 2

Proof of Lemma 2. To expedite the proof, we express the LR statistics in terms of the sufficient statistics $\mathbf{y}_0 = \frac{1}{|C|}\sum_{i\in C}\mathbf{y}_i \sim N(\beta_0,\sigma_0^2)$ and $\mathbf{y}_1 = \frac{1}{|\bar{C}|}\sum_{i\in \bar{C}}\mathbf{y}_i \sim N(\beta_1,\sigma_1^2)$ for $\sigma_0 = \sigma/\sqrt{|C|}$ and $\sigma_1 = \sigma/\sqrt{|\bar{C}|}$. Then, we obtain

$$2\log\Lambda_C(\mathbf{y}) = \frac{1}{\sigma_0^2}(\mathbf{y}_0-\hat{\beta})^2 + \frac{1}{\sigma_1^2}(\mathbf{y}_1-\hat{\beta})^2$$

where $\hat{\beta} = \frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2}\mathbf{y}_0 + \frac{\sigma_0^2}{\sigma_0^2+\sigma_1^2}\mathbf{y}_1$ is the MLE under H_0. (The likelihood under the alternative balances with the normalizing constant of the null likelihood.) Thus,

$$\begin{aligned}
2\log\Lambda_C(\mathbf{y}) &= \frac{1}{\sigma_0^2}\left(\frac{\sigma_0^2}{\sigma_0^2+\sigma_1^2}(\mathbf{y}_0-\mathbf{y}_1)\right)^2 \\
&\quad + \frac{1}{\sigma_1^2}\left(\frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2}(\mathbf{y}_0-\mathbf{y}_1)\right)^2 \\
&= \frac{(\mathbf{y}_0-\mathbf{y}_1)^2}{\sigma_0^2+\sigma_1^2} = \frac{1}{\sigma^2}\frac{|C||\bar{C}|}{|V|}(\mathbf{y}_0-\mathbf{y}_1)^2 \\
&= \frac{1}{\sigma^2}\frac{|V|}{|C||\bar{C}|}\left(\frac{|\bar{C}|}{|V|}\sum_{v\in C}\mathbf{y}_v - \frac{|C|}{|V|}\sum_{v\in\bar{C}}\mathbf{y}_v\right)^2 \\
&= \frac{1}{\sigma^2}\frac{|V|}{|C||\bar{C}|}\left(\sum_{v\in C}\mathbf{y}_v - \frac{|C|}{|V|}\sum_{v\in V}\mathbf{y}_v\right)^2 \\
&= \frac{1}{\sigma^2}\frac{|V|}{|C||\bar{C}|}\left(\sum_{v\in C}\tilde{\mathbf{y}}_v\right)^2. \qquad (11)
\end{aligned}$$

Now we let $\mathbf{x} = \mathbf{1}_C$, making the statistic above

$$2\sigma^2\log\Lambda_C(\mathbf{y}) = \frac{\mathbf{x}^\top\tilde{\mathbf{y}}\tilde{\mathbf{y}}\mathbf{x}}{\mathbf{x}^\top\mathbf{K}\mathbf{x}} \text{ and } \frac{|\partial C||V|}{|C||\bar{C}|} = \frac{\mathbf{x}^\top\mathbf{L}\mathbf{x}}{\mathbf{x}^\top\mathbf{K}\mathbf{x}}.$$

The result now follows by considering all the indicator functions corresponding to the sets in $\mathcal{C}$. □

Proof of Remark 4. First we notice that (8) is equivalent to

$$\inf_{\mathbf{x}\in\mathbb{R}} -\mathbf{x}^\top\tilde{\mathbf{y}} \text{ s.t. } \mathbf{x}^\top\mathbf{L}\mathbf{x}\le\rho, \|\mathbf{x}\|\le 1$$

because $\mathbf{x}^\top\mathbf{L}\mathbf{x}$ and $\mathbf{x}^\top\tilde{\mathbf{y}}$ are invariant under changes in $\mathbf{1}^\top\mathbf{x}$. This admits the Lagrangian (for parameters $\nu_0,\nu_1>0$),

$$-\mathbf{x}^\top\tilde{\mathbf{y}} + \nu_0(\mathbf{x}^\top\mathbf{L}\mathbf{x}-\rho) + \nu_1(\mathbf{x}^\top\mathbf{x}-1)$$

which is minimized for fixed ν_0,ν_1 at $\mathbf{x} = -\frac{1}{2}[\nu_0\mathbf{L}+\nu_1\mathbf{I}]^{-1}\tilde{\mathbf{y}}$ (which confirms Slater's condition). Hence, the dual program is

$$\sup_{\nu_0,\nu_1\ge 0} -\nu_0\rho-\nu_1-\frac{1}{2}\tilde{\mathbf{y}}[\nu_0\mathbf{L}+\nu_1\mathbf{I}]^{-1}\tilde{\mathbf{y}}+\frac{1}{4}\tilde{\mathbf{y}}[\nu_0\mathbf{L}+\nu_1\mathbf{I}]^{-1}\tilde{\mathbf{y}}$$

□

A.2 Proofs in Section 3

Proof of Theorem 5 (1). Let the true $C\in\mathcal{C}$ be known. The performance of the optimal test with C known, which by the Neyman-Pearson Lemma is based on $2\log\Lambda_C(\mathbf{y})$, bounds the performance of that with C unknown. To this end, note that, under H_0, the LR statistic (6) has a χ_1^2, while under the alternative H_1^C it has a $\chi_1^2(\lambda)$ distribution with non-centrality parameter

$$\lambda = \frac{\delta^2}{\sigma^2}\frac{|C||\bar{C}|}{|V|} = \frac{\eta^2}{\sigma^2},$$

which is the square of the SNR. For fixed C, asymptotically indistinguishable of H_0 versus H_C^1 follows by considering any threshold and noticing that the associated type 1 and type 2 errors are non-vanishing under the SNR scaling assumed in the statement. Since the risk of testing H_0 versus H_1 is no smaller than the risk of testing H_0 versus H_C^1, the result follows. □

We remark that the proof of the previous result shows that when distinguishing H_0 from H_1^C, the power of the test is maximal when $|C| = |\bar{C}|$ for a fixed value of the SNR.

Proof of Theorem 5 (2). We will begin by constructing from our set, $\mathcal{C}'$, a new set, $\mathcal{S}$, of clusters which are difficult to distinguish in the sense that the Bayes risk for the uniform prior over those in the alternative is bounded away from 0. Enumerate $\mathcal{C}'$ such that $\mathcal{C}' = \{C_i\}_{i=1}^{|\mathcal{C}'|}$. We will build $\mathcal{S}$ by unioning k elements of $\mathcal{C}'$, then draw S, S' uniformly from $\mathcal{S}$. Specifically, let $k = \lfloor\sqrt{|\mathcal{C}'|}\rfloor$ (recall that $c = |C|, \forall C\in\mathcal{C}'$), and let K, K' be independent uniform samples without replacement of k elements from $\{1,\dots,|\mathcal{C}'|\}$. Then let $S = \cup_{i\in K}C_i$ and $S' = \cup_{i\in K'}C_i$. Notice that $kc = |S| \le n/2$ for n large enough.

$$\begin{aligned}
\frac{|\partial S|}{|S||\bar{S}|} &\le \frac{k\max_{C\in\mathcal{C}'}|\partial C|}{kc(n-kc)} \\
&\le \frac{n-c}{n-kc}\max_{C\in\mathcal{C}'}\frac{|\partial C|}{c(n-c)} \le 2\frac{\rho}{2} = \rho
\end{aligned}$$

Notice that the risk can be bounded by

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\Theta_0}\mathbb{E}_{\boldsymbol{\beta}}T(\mathbf{y}) + \sup_{\boldsymbol{\beta}\in\Theta_1}\mathbb{E}_{\boldsymbol{\beta}}[1-T(\mathbf{y})] \\
&\ge \mathbb{E}_{\boldsymbol{\beta}=\mathbf{0}}T(\mathbf{y}) + \frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}}\mathbb{E}_{\boldsymbol{\beta}^S}[1-T(\mathbf{y})] = R^*
\end{aligned}$$

where $\boldsymbol{\beta}^S = \eta\sqrt{\frac{n}{|S||\bar{S}|}}\mathbf{1}_S$ and $\mathcal{S} \subseteq \mathcal{C}$. Then by Proposition 3.2 in [1],

$$R^* \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}\exp\left\{\frac{\eta^2}{\sigma^2}Z\right\} - 1}$$

where

$$Z = \frac{n|S \cap S'|}{\sqrt{|S||\bar{S}||S'||\bar{S}'|}}$$

for S, S' drawn independently uniformly from $\mathcal{S}$. Notice that

$$\frac{n}{\sqrt{|\bar{S}'||\bar{S}|}} \leq 2$$

Hence,

$$Z \leq 2\frac{|S \cap S'|}{\sqrt{|S||S'|}} = 2\frac{|K \cap K'|}{\sqrt{|K||K'|}}$$

And we have that

$$R^* \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}e^{\frac{2\eta^2}{k\sigma^2}|K\cap K'|} - 1}$$

Hence, we can apply Proposition 3.4 from [1] (by substituting $\mu \leftarrow \eta\sqrt{2}/(\sigma\sqrt{k})$) and determine that $R^* > \delta$ if

$$\frac{\eta\sqrt{2}}{\sigma\sqrt{k}} \leq \sqrt{\log\left(1 + \frac{|\mathcal{C}'|\log(1+4(1-\delta)^2)}{k^2}\right)}$$

Because $k^2 \asymp |\mathcal{C}'|$ we have asymptotic indistinguishability if $\eta/\sigma = o(\sqrt{k}) = o(|\mathcal{C}'|^{1/4})$. For some explanation for the choice of k the term $k\log(1 + |\mathcal{C}'|/k^2)$ is largest when $k^2 \asymp |\mathcal{C}'|$. □

Proof of Lemma 7. Without loss of generality, let $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We recall that, since G is connected, the combinatorial Laplacian $\mathbf{L}$ is symmetric, its smallest eigenvalue is zero and the remaining eigenvalues are positive. By the spectral theorem, we can write $\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^\top$, where Λ is a $(n-1)\times(n-1)$ diagonal matrix containing the positive eigenvalues of $\mathbf{L}$ in increasing order and the columns of the $n \times (n-1)$ matrix $\mathbf{U}$ are the associated eigenvectors. Then, since each vector $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{1}^\top\mathbf{x} = 0$ can be written as $\mathbf{U}\mathbf{z}$ for a unique vector $\mathbf{z} \in \mathbb{R}^{n-1}$, we have

$$\begin{aligned}
\mathcal{X} &= \{\mathbf{x} \in \mathbb{R}^n \colon \mathbf{x}^\top\mathbf{L}\mathbf{x} \leq \rho, \mathbf{x}^\top\mathbf{x} = 1, \mathbf{1}^\top\mathbf{x} \leq 0\} \\
&= \{\mathbf{U}\mathbf{z} \colon \mathbf{z} \in \mathbb{R}^{n-1}, \\
&\qquad \mathbf{z}^\top\mathbf{U}^\top\mathbf{L}\mathbf{U}\mathbf{z} \leq \rho, \mathbf{z}^\top\mathbf{U}^\top\mathbf{U}\mathbf{z} \leq 1\} \\
&= \{\mathbf{U}\mathbf{z} \colon \mathbf{z} \in \mathbb{R}^{n-1}, \tfrac{1}{\rho}\mathbf{z}^\top\Lambda\mathbf{z} \leq 1, \mathbf{z}^\top\mathbf{z} \leq 1\},
\end{aligned}$$

where in the third identity we have used the fact that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_{n-1}$. Letting $\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^{n-1} \colon \frac{1}{\rho}\mathbf{z}^\top\Lambda\mathbf{z} \leq 1, \mathbf{z}^\top\mathbf{z} \leq 1\}$, we see that

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top\mathbf{y} = \sup_{\mathbf{z}\in\mathcal{Z}} \mathbf{z}^\top\mathbf{U}^\top\mathbf{y} \overset{d}{=} \sup_{\mathbf{z}\in\mathcal{Z}} \mathbf{z}^\top\boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim N(0, \mathbf{I}_{n-1})$ and $\overset{d}{=}$ denotes equality in distribution.

Next, we show that the set $\mathcal{Z}$, which is the intersection of an ellipsoid with the unit ball in $\mathbb{R}^{n-1}$, is contained in an enlarged ellipsoid. The supremum of the Gaussian process $\mathbf{z}^\top\boldsymbol{\xi}$ over $\mathcal{Z}$ will then be bounded by the supremum of the same process over this larger but simpler set, which we will be able to bound using directly a result from [38] based on chaining. To this end, let $\mathbf{A} = \frac{1}{\rho}\Lambda = \mathrm{diag}\{a_i\}_{i=1}^{n-1}$ and $d = \max\{j : a_j < 1\}$. For for a vector $\mathbf{z} \in \mathbb{R}^{n-1}$ set $\mathbf{z}_1 = \mathbf{z}_{[d]}$, $\mathbf{z}_2 = \mathbf{z}_{[n-1]\setminus[d]}$, and $\mathbf{A}_2 = \mathrm{diag}\{a_i\}_{i>d}$. Then, we observe the following chain of implications, holding for vectors $\mathbf{z} \in \mathbb{R}^{n-1}$:

$$\begin{aligned}
&\|\mathbf{z}\| \leq 1, \mathbf{z}^\top\mathbf{A}\mathbf{z} \leq 1 \Rightarrow \|\mathbf{z}_1\| \leq 1, \sum_{i>d} a_i\mathbf{z}_i^2 \leq 1 \\
&\Rightarrow \mathbf{z}_1^\top\mathbf{z}_1 + \mathbf{z}_2^\top\mathbf{A}_2\mathbf{z}_2 \leq 2 \Rightarrow \sum_i \frac{\max\{1, a_i\}}{2}\mathbf{z}_i^2 \leq 1.
\end{aligned}$$

Hence, we have the bound

$$\mathbb{E}\sqrt{\hat{s}} \leq \mathbb{E}\sup_{\mathbf{z}\in\mathbb{R}^{n-1}} \mathbf{z}^\top\boldsymbol{\xi} \text{ s.t. } \sum_i 2\max\{1, a_i\}\mathbf{x}_i^2 \leq 1.$$

Recalling that $a_i = \frac{\lambda_{i+1}}{\rho}$, for $i = 1, \ldots, n-1$, where λ_{i+1} is the $(i+1)$th eigenvalue of $\mathbf{L}$, by Proposition 2.2.1 in [38] the right hand side of the previous expression is bounded by $\sqrt{2\sum_{i>1}\min\{1, \rho\lambda_i^{-1}\}}$. □

Supplement to the proof of Theorem 6. The following property of Gaussian processes effectively reduces the study of their supremum to the study of its expectation. It was established by [7] and [10] and can be found in [22].

Lemma 14. *Consider a Gaussian process $\{Z_t\}_{t\in\mathcal{U}}$ where $\mathcal{U}$ is compact with respect to metric*

$$d(s,t) = (\mathbb{E}(Z_s - Z_t)^2)^{1/2}, \quad s, t, \in \mathcal{U},$$

and let $\sigma^2 \geq \sup_{t\in\mathcal{U}} \mathbb{E}Z_t^2$. We have that with probability at least $1-\delta$

$$\left|\sup_{t\in\mathcal{U}} Z_t - \mathbb{E}\sup_{t\in\mathcal{U}} Z_t\right| < \sqrt{2\sigma^2\log\frac{2}{\delta}}.$$

Notice that the natural distance is given by $d(\mathbf{x}_0, \mathbf{x}_1) = (\mathbb{E}((\mathbf{x}_0 - \mathbf{x}_1)^\top\mathbf{y})^2)^{1/2} = \sigma\|\mathbf{x}_0 - \mathbf{x}_1\|$ for $\mathbf{x}_0, \mathbf{x}_1 \in \mathcal{X}$. □

A.3 Proof in Section 4

Proof of Corollary 11 (a). The study of the spectra of trees really began in earnest with the work of [12]. Notably, it became apparent that tree have eigenvalues with high multiplicities, particularly the eigenvalue 1.

[30] gave a tight bound on the algebraic connectivity of balanced binary trees (BBT). They found that for a BBT of depth ℓ, the reciprocal of the smallest eigenvalue ($\lambda_2^{(\ell)}$) is

$$\begin{aligned}\frac{1}{\lambda_2^{(\ell)}} \leq 2^\ell - 2\ell + 2 - \frac{2^\ell - \sqrt{2}(2\ell - 1 - 2^{\ell-1})}{2^\ell - 1 - \sqrt{2}(2^{\ell-1} - 1)} \\ +(3 - 2\sqrt{2}\cos(\frac{\pi}{2\ell - 1}))^{-1} \\ \leq 2^\ell + 105I\{\ell < 4\}\end{aligned} \tag{12}$$

[32] gave a more exact characterization of the spectrum of a balanced binary tree, providing a decomposition of the Laplacian's characteristic polynomial. Specifically, the characteristic polynomial of $\mathbf{L}$ is given by

$$\begin{aligned}\det(\lambda\mathbf{I} - \mathbf{L}) = p_1^{2^{\ell-2}}(\lambda)p_2^{2^{\ell-3}}(\lambda)\dots \\ p_{\ell-3}^{2^2}(\lambda)p_{\ell-2}^2(\lambda)p_{\ell-1}(\lambda)s_\ell(\lambda)\end{aligned} \tag{13}$$

where $s_\ell(\lambda)$ is a polynomial of degree ℓ and $p_i(\lambda)$ are polynomials of degree i with the smallest root satisfying the bound in (12) with ℓ replaced with i. In [33], they extended this work to more general balanced trees.

By (13) we know that at most $\ell + (\ell-1) + (\ell-2)2 + ... + (\ell-j)2^{j-1} \leq \ell 2^j$ eigenvalues have reciprocals larger than $2^{\ell-j} + 105I\{j < 4\}$. Let $k = \max\{\lceil \frac{\ell}{c} 2^{\ell(1-\alpha)} \rceil, 2^3\}$, then we have ensured that at most k eigenvalues are smaller than ρ. For n large enough

$$\begin{aligned}\sum_{i>1} \min\{1, \rho\lambda_i^{-1}\} \leq k + \rho \sum_{j>\log k}^{\ell} \ell 2^j 2^{\ell-j} \\ = k + \ell(\ell - \log k)n\rho = O(n^{1-\alpha}(\log n)^2)\end{aligned}$$

□

Proof of Corollary 11 (b). We will construct $\mathcal{C}'$ in Theorem 5 (b) from subtrees of size $4cn^\alpha$. Let C be such a subtree, then for n large enough

$$\begin{aligned}1 - 4cn^{\alpha-1} \geq \frac{1 - cn^{\alpha-1}}{2} \\ \Rightarrow \frac{n|\partial C|}{|C||\bar{C}|} = [4cn^\alpha(1 - 4cn^{\alpha-1})]^{-1} \\ \leq \frac{1}{2}[cn^\alpha(1 - cn^{\alpha-1})]^{-1} = \frac{rho}{2}\end{aligned}$$

Hence the conditions of Theorem 5 (b) hold with $|\mathcal{C}'| = n/(4cn^\alpha) \asymp n^{1-\alpha}$ □

Proof of Corollary 12 (a). By a simple Fourier analysis (see [36]), we know that the Laplacian eigenvalues are $2(2 - \cos(2\pi i_1/p) - \cos(2\pi i_2/p))$ for all $i_1, i_2 \in [p]$. Let us denote the p^2 eigenvalues as $\lambda_{(i_1,i_2)}$ for $i_1, i_2 \in [p]$. Notice that for $i \in [p]$, $|\{(i_1, i_2) : i_1 \vee i_2 = i\}| \leq 2i$. For simplicity let p be even. We know that if $i_1 \vee i_2 \leq p/2$ then $\lambda_{(i_1,i_2)} = 2 - \cos(2\pi i_1/p) - \cos(2\pi i_2/p) \geq 1 - \cos(2\pi(i_1 \vee i_2)/p)$. Thus,

$$\begin{aligned}&\sum_{(i_1,i_2)\neq(1,1)\in[p]^2} 1 \wedge \frac{\rho}{\lambda_{(i_1,i_2)}} \\ &\leq 2\sum_{i\in[p/2]} 2i\left(1 \wedge \frac{\rho}{1 - \cos(2\pi i/p)}\right) \\ &\leq \rho\frac{p^2}{2}\frac{2}{p}\sum_{i\in[p/2]} 2\frac{i/p}{1 - \cos(2\pi i/p)} \\ &\leq \rho\frac{p^2}{2}\int_{1/p}^{1/2} \frac{x dx}{1 - \cos(2\pi x)} \\ &\leq \rho\frac{p^2}{2}\left.\frac{\log(\sin(\pi x)) - \pi x\cot(\pi x)}{2\pi^2}\right|_{1/p}^{1/2} \\ &= \rho\frac{p^2}{2}\frac{(\pi/p)\cot(\pi/p) - \log(\sin(\pi/p))}{2\pi^2}\end{aligned}$$

While we can use the first order expansion of the terms to obtain the behavior,

$$\begin{aligned}(\pi/p)\cot(\pi/p) = 1 + o(\pi/p) \\ -\log(\sin(\pi/p)) = -\log(\pi/p) - \log(1 + o(1))\end{aligned}$$

so we arrive at the following,

$$\begin{aligned}&\sum_{(i_1,i_2)\neq(1,1)\in[p]^2} 1 \wedge \frac{\rho}{\lambda_{(i_1,i_2)}} \\ &\leq \rho\frac{p^2}{4\pi^2}(1 + \log(p/\pi) + o(1)) \\ &= \frac{C}{4\pi^2}p^{1+\beta}(1 + \log(p/\pi) + o(1)) \\ &= O(n^{(1+\beta)/2}\log(p))\end{aligned}$$

which in conjunction with (9) completes our proof. □

Proof of Corollary 13 (a). The Kronecker product of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ is defined as $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{(n\times n)\times(n\times n)}$ such that $(\mathbf{A} \otimes \mathbf{B})_{(i_1,i_2),(j_1,j_2)} = A_{i_1,j_1}B_{i_2,j_2}$. Some matrix algebra shows that if H_1 and H_2 are graphs on p vertices with Laplacians $\mathbf{L}_1, \mathbf{L}_2$ then the Laplacian of their Kronecker product, $H_1 \otimes H_2$, is given by $\mathbf{L} = \mathbf{L}_1 \otimes \mathbf{I}_p + \mathbf{I}_p \otimes \mathbf{L}_2$ ([28]). Hence, if $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^p$ are eigenvectors, viz. $\mathbf{L}_1\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $\mathbf{L}_2\mathbf{v}_2 = \lambda_2\mathbf{v}_2$, then $\mathbf{L}(\mathbf{v}_1 \otimes \mathbf{v}_2) = (\lambda_1 + \lambda_2)\mathbf{v}_1 \otimes \mathbf{v}_2$, where $\mathbf{v}_1 \otimes \mathbf{v}_2$ is the usual tensor product. This completely characterizes the spectrum of Kronecker products of graphs.

We should argue the choice of $\rho \propto p^{2k-\ell-1}$, by showing that it is the results of cuts at level k. We say that an edge $e = ((i_1, ..., i_\ell), (j_1, ..., j_\ell))$ has scale k if $i_k \neq j_k$.

Furthermore, a cut has scale k if each of its constituent edges has scale at least k. Each edge at scale k has weight $p^{k-\ell}$ and there are $p^{\ell-1}$ such edges, so cuts at scale k have total edge weight bounded by

$$p^{\ell-1}\sum_{i=1}^{k} p^{i-\ell} = p^{k-1}\frac{p-\frac{1}{p^{k-1}}}{p-1} \leq \frac{p^k}{p-1}$$

Cuts at scale k leave components of size $p^{\ell-k}$ intact, meaning that $\rho \propto p^{2k-\ell-1}$ for large enough p.

We now control the spectrum of the Kronecker graph. Let the eigenvalues of the base graph H be $\{\nu_j\}_{j=1}^p$ in increasing order. The eigenvalues of G are precisely the sums

$$\lambda_i = \frac{1}{p^{\ell-1}}\nu_{i_1} + \frac{1}{p^{\ell-2}}\nu_{i_2} + ... + \frac{1}{p}\nu_{i_{\ell-1}} + \nu_{i_\ell}$$

for $i = (i_j)_{j=1}^{\ell} \subseteq [p]$. The eigenvalue distribution $\{\lambda_i\}$ stochastically bounds

$$\lambda_i \geq \sum_{j=1}^{\ell} \frac{1}{p^{\ell-j}}\nu_2 I\{\nu_{i_j} \neq 0\} \geq \frac{\nu_2}{p^{Z(i)}}$$

where $Z(i) = \min\{j : \nu_{i_{\ell-j}} \neq 0\}$. Notice that if i is chosen uniformly at random then $Z(i)$ has a geometric distribution with probability of success $(p-1)/p$. Also $\rho/(\frac{\nu_2}{p^{Z(i)}}) = p^{Z(i)+2k-\ell-1}/\nu_2 \geq 1$ if $Z(i) \geq \ell+1-2k+\log_p \nu_2$, so

$$\begin{aligned}
&\frac{1}{p^\ell}\sum_{i\in[p]^\ell} \min\{1, \frac{\rho}{\lambda_i}\} \leq \frac{p^{2k-\ell-1}}{\nu_2} \\
&+ \sum_{Z=1}^{\lfloor \ell+1-2k+\log_p \nu_2 \rfloor} \frac{p^{Z+2k-\ell-1}}{\nu_2}\frac{1}{p^Z}\frac{p-1}{p} \\
&\leq \frac{(\ell+2)p^{2k-\ell-1}}{\nu_2}
\end{aligned}$$

This followed from the geometric probability mass function. We also know that the algebraic connectivity, ν_2, is bounded from below by $4p^{-2}$, so the following result holds.

□

Proof of Corollary 13 (b). Similarly to the proof of Corollary 11 (b), we form $\mathcal{C}'$ as the connected components of the graph with all the edges at coarseness less than $k-2$. So we have more than quadrupled the size of the clusters without increasing their cut size. Hence, $|\mathcal{C}'| \asymp p^{k-2} \asymp n^{k/\ell}/p^2$.

□