Detecting Activations over Graphs using Spanning Tree Wavelet Bases

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

Akshay Krishnamurthy
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
akshaykr@cs.cmu.edu

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

Abstract

We consider the detection of clusters of activation over graphs under Gaussian noise. This problem appears in many real world scenarios, such as detecting contamination or seismic activity by sensor networks, viruses in human and computer networks, and groups with anomalous behavior in social and biological networks. We introduce the spanning tree wavelet basis over a graph, a localized basis that reflects the topology of the graph, and a detector based on this construction. We characterize orthonormality and sparsifying properties of the proposed basis, which can be useful for tasks other than detection, such as de-noising, compression and localization. For the detection problem, we provide a necessary condition for asymptotic distinguishability of the null and alternative hypotheses. Then we prove that our detector can correctly detect signals in a low signal-to-noise regime using spanning tree wavelets, for any spanning tree. We then use electric network theory to show that a spanning tree drawn uniformly at random provides a stronger performance guarantee that in many cases matches the necessary condition. For edge transitive graphs, k-nearest neighbor graphs, and ϵ-graphs we obtain nearly optimal performance with the uniform spanning tree wavelet detector.

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.

1 Introduction

This paper focuses on the problem of detecting activations over a graph when observations are corrupted by noise. The problem of detecting graph-structured activations is relevant to many applications including identifying contamination or seismic activity by sensor networks, viruses in human and computer networks, and groups with anomalous attribute values in social and biological networks. Furthermore, these applications require that the method be scalable to large graphs. Luckily, computer science boasts a plethora of efficient graph based algorithms that we can leverage to develop detectors for graph-structured activations that are both statistically and computationally efficient.

1.1 Contributions

In this paper, we will be testing if there is a non-zero piece-wise constant activation pattern on the graph, given observations that are corrupted by Gaussian white noise. We show that correctly distinguishing the null and alternative hypotheses requires that the signal-to-noise ratio (SNR) grows quickly with the allowable number of discontinuities in the activation pattern. For comparison purposes, we show that naive tests that do not incorporate the graph structure are not capable of achieving asymptotic distinguishability at low SNR (Section 2). Since a test based on the scan statistic which matches the observations with all possible activation patterns by brute force is computationally infeasible, we propose a Haar wavelet basis construction for general graphs, which is formed by hierarchically dividing a spanning tree of the graph (Section 3). We find that the size and power of the test can be bounded in terms of the number of signal discontinuities and maximum degree of the spanning tree, immediately giving us a result for any spanning tree. We then propose choosing a spanning tree uniformly at random (this can be done in a computa-

tionally efficient manner), and show that the bound can be improved by a factor to depend on the average effective resistance of the boundary of the activation instead of the size of the boundary (Section 4). With this machinery in place we are able to show that for edge transitive graphs (for example lattices), k-nearest neighbor graphs, and ϵ random geometric graphs, our test is nearly-optimal in that the upper bounds match the fundamental limits of detection up to logarithm factors (Section 5).

1.2 Problem Setup

Consider an undirected graph G defined by a set of vertices V ($|V| = n$) and undirected edges E ($|E| = m$) which are unordered pairs of vertices. Throughout this paper we will assume that the graph G is known and unweighted. The statistical setting that we consider is the normal means model,

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_{n\times n})$, and σ^2 is known. The structure of activation pattern $\mathbf{x}$ is determined by the graph G. Specifically, we assume that there are parameters ρ, μ (possibly dependent on n[1]) such that the class of graph-structured activation patterns $\mathbf{x}$ is given as follows.

$$\mathcal{X} = \left\{ \mathbf{x} : \mathbf{x} = \frac{\mu}{\sqrt{|C|}} \mathbf{1}_C, C \subset V, |\partial C| \leq \rho \right\}$$

where the boundary is defined as $\partial C = \{(v, w) \in E : v \in C, w \notin C\}$ and the indicator $\mathbf{1}_C(v) = \mathbf{1}\{v \in C\}$. Hence, the possible patterns have few edges across which the values of $\mathbf{x}$ differ. This generalizes the notion of functions with few discontinuities to the graphs setting. In other words, the set of activated vertices C have a small *cut size* in the graph G, where cut size is exactly $|\partial C|$. If the number of activated vertices is $|C|$, then values of $\rho < |C| \cdot d_{\max}$ [2] imply that the activation is localized on the graph.

In graph-structured activation detection we are concerned with statistically testing the null and alternative hypotheses,

$$\begin{aligned} H_0 &: \mathbf{y} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \\ H_1 &: \mathbf{y} \sim N(\mathbf{x}, \sigma^2 \mathbf{I}), \mathbf{x} \in \mathcal{X} \end{aligned} \tag{1}$$

H_0 represents business as usual while H_1 encompasses all of the foreseeable anomalous activity. Let a test be a mapping $T(\mathbf{y}) \in \{0, 1\}$, where 1 indicates that we reject the null.

It is imperative that we control both the probability of false alarm, and the false acceptance of the null. To this end, we define our measure of risk to be

$$R(T) = \mathbb{E}_{\mathbf{0}}[T] + \sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{y}}[1 - T]$$

where $\mathbb{E}_{\mathbf{y}}$ denote the expectation with respect to $\mathbf{y} \sim N(\mathbf{x}, \sigma^2 \mathbf{I})$. The test T may be randomized, in which case the risk is $\mathbb{E}_T R(T)$. Notice that if the distribution of the random test T is independent of $\mathbf{x}$, then $\mathbb{E}_T \sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{y}}[1 - T] = \sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{T,\mathbf{y}}[1 - T]$. This setting should not be confused with the Bayesian testing setup (e.g. as considered in [1, 5]) where the patterns are drawn at random instead of the test being randomized. We will say that H_0 and H_1 are *asymptotically distinguished* by a test, T, if in the setting of large graphs $\lim_{n\to\infty} R(T) = 0$. If such a test exists then H_0 and H_1 are asymptotically distinguished, otherwise they are asymptotically indistinguishable.

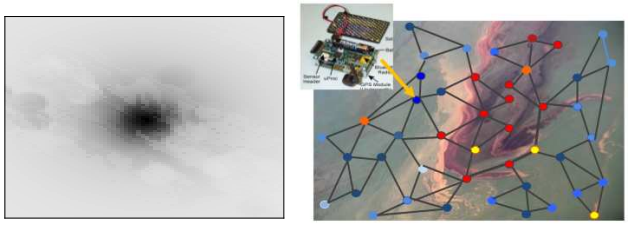

Figure 1: A heatmap of ground motion in a seismic event as recorded by a grid of sensors (left) and a pictorial representation of sensor measurements for water contamination (right).

To aid us in our study we introduce some mathematical terminology. Let the edge-incidence matrix of G be $\nabla \in \mathbb{R}^{m\times n}$ such that for $(v, w) \in E$, $\nabla_{(v,w),v} = 1$, $\nabla_{(v,w),w} = -1$ (the order of (v, w) is chosen arbitrarily) and is 0 elsewhere. For a vector, $\mathbf{w} \in \mathbb{R}^m$, $\text{supp}(\mathbf{w}) = \{e \in E : \mathbf{w}(e) \neq 0\}$ and $\|\mathbf{w}\|_0 = |\text{supp}(\mathbf{w})|$, so $\|\nabla \mathbf{x}\|_0 \leq \rho$ for all $\mathbf{x} \in \mathcal{X}$. We will be constructing spanning trees $\mathcal{T}$ of the graph G, which are connected subsets of E with no cycles that span all vertices in G. Furthermore, we will denote the edge-incidence matrix of $\mathcal{T}$ as $\nabla_{\mathcal{T}}$.

1.3 Motivating Examples

Localized patterns on graphs are prevalent across a variety of scientific and sociological disciplines. Before presenting our results, we briefly discuss two specific examples that motivate our work.

In an effort to better understand tectonic movements and seismic events, the U.S. Geological Survey Earthquake Hazards Program has deployed sensor grids that provides real-time measurements of ground motion around seismic events [30]. The example snapshot of these measurements (Figure 1) reveals that the majority of the ground movement is localized around the

[1] We suppress dependence on the number of edges m as we focus on graph models where m depends on n.

[2] $d_{\max}$ is the maximum degree of any $v \in G$.

epicenter of the earthquake, as one would expect. In this example, the signal strength is quite high and it seems easy to detect the seismic event, but for smaller events the signal-to-noise ratio is much lower, making it difficult to identify these earthquakes. Seismologists are interested in detecting such small seismic events, as they provide information that can assist in predicting larger, more dangerous earthquakes, motivating the development of detectors for weak, localized patterns over graphs.

Water contamination remains a major concern for human and environmental health. The introduction of pathogens into water sources is a common source of human diseases, e.g. polio virus, enterotoxigenic E. coli etc. As the world witnessed during the Deepwater Horizon oil spill in 2010, the spread of contaminant (in this case oil) had far reaching consequences for the Gulf of Mexico ecosystem. It stands to reason that accurate and early detection of contaminants through sensors distributed in water bodies is of particular importance. In this setting, there is a network of sensors, each of which is providing noisy measurements of contaminant levels. One can thus determine edges between sensors by their relative proximities, and exploit this graph structure to boost the power of the statistical test.

The localized activation patterns in these examples are characterized by discontinuities in the signal and hence well-modeled by piecewise smooth functions on graphs. The guarantees on our testing procedure currently only hold for piecewise constant patterns (since the Haar wavelets provide a good approximation for such patterns). However, one can envision extending our Haar construction to smoother wavelets and this constitutes an important direction of future work.

1.4 Related Work

The statistical problem that we are addressing can be broadly classified as a high-dimensional Gaussian goodness-of-fit test. This is a well studied problem when the structure of H_1 derives from a smooth function space such as an ellipsoid, Besov space or Sobolev space [18, 19]. The function space $\mathcal{X}$ that we are proposing is combinatorial in nature. This and related statistical problems have only recently been studied theoretically [1, 4, 5, 6, 3, 27, 21], although to the best of our knowledge none have addressed the problem under arbitrary graph structure. More broadly, this work falls under the purview of multiple hypothesis testing, which has a rich history [7]. Unfortunately, aside from a few special cases [15], the multiple tests are assumed to be unstructured, making any such work not applicable to our setting.

In this paper, we evaluate our method by its ability to distinguish H_0 from H_1, however the procedure is based on constructing a wavelet basis over graphs which is relevant for other problems such as denoising and compression. Wavelets are multi-resolution bases that can represent inhomogeneous signals efficiently using a few non-zero wavelet coefficients which makes them attractive for denoising, compression and detection. As a result, they have been used extensively in mathematics, signal processing, statistics and physics [24]. They have also been used with great success in statistics, with extensive theoretical guarantees (for example, see [10, 17, 31]). Recently there has been some attention paid to developing wavelets for graphs. Unfortunately, most of these have either focused on graphs with a known hierarchical structure [14, 26, 28], or do not come with approximation or sparsifying properties that can be used for our class of graph functions $\mathcal{X}$ [9, 16].

2 Universal Lower Bound and Unstructured Tests

In order to more completely understand the problem of detecting weak activations on graphs, we prove that there is a universal minimum signal strength under which H_0 and H_1 are asymptotically indistinguishable. The proof is based on a result developed in [5], with a new construction of prior distribution over worst case patterns.

Theorem 1. *Hypotheses H_0 and H_1 defined in Eq. (1) are asymptotically indistinguishable if*

$$\frac{\mu}{\sigma} = o\left(\sqrt{\min\left\{\frac{\rho}{d_{\max}}\log\left(\frac{nd_{\max}^2}{\rho^2}\right), \sqrt{n}\right\}}\right)$$

where $d_{\max}$ is the maximum degree of graph G.

Proof. We begin by constructing a prior distribution over $\mathcal{X}$. Consider $k = \min\{\rho/d_{\max}, \sqrt{n}\}$ and form a subset $\mathcal{S} \subseteq 2^n$ that consists of all subsets of k vertices, i.e. $|S| = k$ for all $S \in \mathcal{S}$. For every $S \in \mathcal{S}$ construct a pattern $\mathbf{x} = \frac{\mu}{\sqrt{|S|}}\mathbf{1}_S$ and denote the collection of such patterns as $\mathcal{X}'$. Note that for all $\mathbf{x} \in \mathcal{X}'$, $\|\mathbf{x}\|_2 = \mu$ and by construction $|\{(v,w) \in E : x_v \neq x_w\}| \leq kd_{\max} \leq \rho$. Hence, $\mathcal{X}' \subseteq \mathcal{X}$. The prior we construct assigns uniform probability to all patterns in $\mathcal{X}'$ and zero probability to patterns in $\mathcal{X} \setminus \mathcal{X}'$. This gives us a prior distribution π over $\mathcal{X}$.

The Bayes risk associated with a prior π is defined as $R^* = \inf_T R^*(T) = \inf_T\{\mathbb{E}_{\mathbf{0}}[T] + \mathbb{E}_\pi \mathbb{E}_{\mathbf{y}}[1 - T]\}$. Notice that $R(T) \geq R^*(T) \geq R^*$ for all prior distributions π. Our prior satisfies the conditions of Proposition 3.4 in

[1], and hence we get that $R^* \geq \delta$ for all μ such that

$$\frac{\mu}{\sigma\sqrt{k}} \leq \sqrt{\log\left(1 + \frac{n\log(1+4(1-\delta)^2)}{k^2}\right)}$$

from which the theorem follows by considering the two cases $\rho/d_{\max} \leq \sqrt{n}$ and $\rho/d_{\max} > \sqrt{n}$. □

We illustrate that there is a gap between this lower bound and tests that do not incorporate the graph structure. We analyze two test statistics: vertex-wise thresholding and the vertex averaging detector. Vertex-wise thresholding is typically used when it's believed that the parameter $\mathbf{x}$ is sparse, in the sense that it has few non-zero coordinates, or the signal strength at each vertex is high. The following proposition characterizes its performance.

Proposition 2. *Consider the vertex-wise thresholding test statistic* $\max_{v\in V}|y_v|$. *This asymptotically distinguishes* H_0 *and* H_1 *if*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\max_{C:|\partial C|\leq\rho}|C|\log n}\right)$$

and notice that $\max_{C:|\partial C|\leq\rho}|C| \geq n/2$.

The intuition is that the thresholding test statistic performs poorly on large clusters which can be of size nearly n. Hence, then there can be a significant gap between the lower bound of Theorem 1 and the upper bound of Proposition 2. In other words, vertex thresholding does not take advantage of the pattern structure when it is localized on the graph. On the other hand, if the signal is unstructured or very sparse (number of activations do not increase with size of the graph), then the max statistic is nearly optimal (up to log factors).

Another natural test statistic is vertex averaging, $|\frac{1}{n}\sum_{v\in V} y_v|$, whose performance is characterized below.

Proposition 3. *Consider the test statistic* $|\frac{1}{n}\sum_{v\in V} y_v|$. *A necessary and sufficient condition for this test to asymptotically distinguish* H_0 *and* H_1 *is:*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\max_{C:|\partial C|\leq\rho}\frac{n}{|C|}}\right)$$

Because $|C|$ can be as small as a singleton if $\rho > d_{\min}$ then there can be a significant gap between the lower bound of Theorem 1 and the upper bound of Proposition 3. Intuitively, if the number of activated vertices is small, then globally averaging the observations at all vertices is suboptimal.

We will show that testing with spanning tree wavelets is near optimal (up to log factors) and hence bridges the gap in performance exhibited by unstructured tests (both when the structure is localized and when the number of activated vertices is small).

3 Spanning Tree Wavelets

In this section, we present an algorithm for constructing a wavelet basis given a spanning tree $\mathcal{T}$ and we characterize the performance of a test statistic, which thresholds the largest basis coefficient obtained by projecting the observations $\mathbf{y}$ onto the basis elements, for the detection problem specified in (1).

Informally, we would like to construct a basis $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$ (here $\mathbf{b}_i$ is a column of $\mathbf{B}$) such that each edge $e \in \mathcal{T}$ supports very few basis elements, where we say that an edge e supports a basis element $\mathbf{b} \in \mathbf{B}$ if $e \in \text{supp}(\nabla_{\mathcal{T}}\mathbf{b})$. As we will show, upper bounding the number of basis elements supported by any edge will be essential in analyzing the performance of our test statistic $||\mathbf{B}^T\mathbf{y}||_\infty$.

We construct our wavelet basis $\mathbf{B}$ recursively, by first finding a seed vertex in the spanning tree such that the subtrees adjacent to the seed have at most $\lceil n/2\rceil$ vertices and then by including basis elements localized on these subtrees. We recurse on each subtree, adding higher-resolution elements to our basis, and consequently constructing a complete wavelet basis. The first phase of the algorithm ensures that the depth of the recursion is at most $\lceil\log n\rceil$ and the second recursion phase ensures that each edge supports at most $\lceil\log d_{\mathcal{T}}\rceil$ basis elements per recursive call, where $d_{\mathcal{T}}$ is the maximum degree of the spanning tree. Combining these two shows that each edge supports at most $\lceil\log d_{\mathcal{T}}\rceil\lceil\log n\rceil$ basis elements.

Finding a balancing vertex in the tree parallels the technique in [25], which finds a balancing edge. The algorithm starts from any vertex $v \in \mathcal{T}$ and moves along $\mathcal{T}$ to a neighboring vertex w that lies in the largest connected component of $\mathcal{T} \setminus v$. The algorithm repeats this process (moving from v to w) until the largest connected component of $\mathcal{T}\setminus w$ is larger than the largest connected component of $\mathcal{T} \setminus v$ at which point it returns v. We call this the *FindBalance* algorithm (Algorithm 1).

Once we have a balancing vertex v, we form wavelets that are constant over the connected components of $\mathcal{T}\setminus v$ such that any vertex is supported by at most $\log d$ wavelets. Let d_v be the degree of the balancing vertex v and let $c_1, \ldots c_{d_v}$ be the connected components of $\mathcal{T}\setminus v$ (with v added to the smallest component). Our algorithm acts as if $c_1, \ldots c_{d_v}$ form a chain structure and constructs the Haar wavelet basis over them. We call this algorithm *FormWavelets*:

Algorithm 1 FindBalance

Require: $\mathcal{T}$ is a spanning tree of G and initialize $v \in V$ arbitrarily
loop
Let T' be the component of $\mathcal{T}\backslash\{v\}$ of largest size
Let w be the unique neighbor of v in T'.
Let T'' be the component of $\mathcal{T}\backslash\{w\}$ of largest size.
Stop and return v if $|T''| \geq |T'|$.
$v \leftarrow w$.
end loop

Algorithm 2 FormWavelets

Require: $C = \{c_i\}_{i=1}^{d_v}$
(1) Let $C_1 = \cup_{i \leq |C|/2} c_i$ and $C_2 = \cup_{i > |C|/2} c_i$.
(2) Form the following basis element and add it to $\mathbf{B}$:

$$\mathbf{b} = \frac{\sqrt{|C_1||C_2|}}{\sqrt{|C_1|+|C_2|}} \left[\frac{1}{|C_1|} \mathbf{1}_{C_1} - \frac{1}{|C_2|} \mathbf{1}_{C_2} \right]$$

(3) Recurse at (1) with $C \leftarrow \{c_i\}_{i \leq |C|/2}$ and $C \leftarrow \{c_i\}_{i > |C|/2}$ separately.

Algorithm 3 recursively constructs basis elements using the *FindBalance* and *FormWavelets* routines on subtrees of $\mathcal{T}$. We initialize $\mathcal{T}$ to be a spanning tree of the graph and start with no elements in our basis. In the appendix, we verify that this algorithm outputs an orthonormal basis (see Lemma 14). We also prove that the construction takes $O(n\lceil \log n \rceil \lceil \log d_{\max} \rceil)$ time. (This does not include the time it takes to construct the spanning tree, which may be considerably longer depending on the algorithm used.)

Algorithm 3 Spanning Tree Wavelet Construction

(0) Initialize $\mathbf{B}$ with the basis element $\frac{1}{\sqrt{n}}\mathbf{1}$.
(1) Let v be the output of *FindBalance* applied to $\mathcal{T}$.
(2) Let $\mathcal{T}_1, .., \mathcal{T}_{d_v}$ be the connected components of $\mathcal{T}\backslash v$ and add v to the smallest component.
(3) Add the basis elements constructed in *FormWavelets* when applied to $\mathcal{T}_1, ..., \mathcal{T}_{d_v}$.
(4) For each $i \in \{1, \ldots, d_v\}$, recursively apply (1) - (4) on $\mathcal{T}_i$ as long as $|\mathcal{T}_i| \geq 2$.

As we will see, controlling the wavelet domain sparsity $||\mathbf{B}^T\mathbf{x}||_0$ for the patterns $\mathbf{x} \in \mathcal{X}$ is essential in analyzing the performance of the statistic $||\mathbf{B}^T\mathbf{y}||_\infty$. The main theoretical guarantee of our basis construction algorithm is that patterns with small cut sizes in G are sparse when projected onto $\mathbf{B}$. Specifically, we prove this sparsifying property in the appendix:

Lemma 4. *Let ∇ be the incidence matrix of G and $\nabla_{\mathcal{T}}$ be the incidence matrix of $\mathcal{T}$ (where $\mathcal{T}$ has degree at most d). Thus, $||\nabla \mathbf{x}||_0$ is the cut size of pattern $\mathbf{x}$. Then for any $\mathbf{x} \in \mathcal{X}$,*

$$\begin{aligned} ||\mathbf{B}^T\mathbf{x}||_0 &\leq ||\nabla_{\mathcal{T}}\mathbf{x}||_0 \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil \\ &\leq ||\nabla \mathbf{x}||_0 \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil \end{aligned}$$

Equipped with Lemma 4 we can now characterize the performance of $||\mathbf{B}^T\mathbf{y}||_\infty$ on any pattern $\mathbf{x}$. Our bound depends on the choice of the spanning tree $\mathcal{T}$, specifically via the quantity $||\nabla_{\mathcal{T}}\mathbf{x}||_0$, the cut size of $\mathbf{x}$ in $\mathcal{T}$. The proof of the following can be found in the appendix.

Theorem 5. *Perform the test in which we reject the null if $||\mathbf{B}\mathbf{y}||_\infty > \tau$. Set $\tau = \sigma\sqrt{2\log(n/\delta)}$. If*

$$\frac{\mu}{\sigma} \geq \min\left\{ \sqrt{2||\nabla_{\mathcal{T}}\mathbf{x}||_0 \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil}, \sqrt{\frac{n}{|C|}} \right\} g(n,\delta),$$

where $g(n\delta) = \sqrt{\log(2/\delta)} + \sqrt{\log(2n/\delta)}$ is poly-logarithmic in $n, 1/\delta$, then under H_0, $\mathbb{P}_{H_0}\{\text{Reject } H_0\} \leq \delta$, and under H_1, $\mathbb{P}_{H_1}\{\text{Reject } H_0\} \geq 1-\delta$.

Remark 6. *For any tree we have $||\nabla_{\mathcal{T}}\mathbf{x}||_0 \leq ||\nabla \mathbf{x}||_0$ for all patterns $\mathbf{x}$, hence we achieve asymptotic distinguishability of H_0 and H_1 if*

$$\frac{\mu}{\sigma} = \omega(\sqrt{2\rho \log d_{\mathcal{T}}} \log(n)) \tag{2}$$

Comparing this result to the lower bound in Theorem 1, we see that modulo logarithmic factors, this is larger by a factor of $1/\sqrt{d_{\max}}$. For sequences of graphs with bounded degrees, the detector is asymptotically optimal, ignoring logarithmic factors, and the choice of spanning tree does not affect the asymptotic performance. When maximum degrees are increasing in n, there is a growing gap between our current upper bound given by (2) and the lower bound. We will see that choosing a spanning tree uniformly at random allows us to achieve near optimality when the maximum degrees are increasing in n.

4 Uniform Spanning Tree Basis

The uniform spanning tree (UST) is a random spanning tree generation technique that we will use to construct wavelet bases. Because the UST is randomly generated, the test statistic $\|\mathbf{B}^T\mathbf{y}\|_\infty$ when conditioned on $\mathbf{y}$ will also be random. We will first examine the deep connection between electrical networks, USTs and random walks. We will then leverage results from cut sparsification to relate the performance of the UST detector to effective resistances.

4.1 Cuts and Effective Resistance

Effective resistances have been extensively studied in electrical network theory. We define the combinatorial Laplacian of G to be $\Delta = \nabla^\top \nabla$. A *potential difference* is any $\mathbf{z} \in \mathbb{R}^{|E|}$ such that it satisfies *Kirchoff's potential law*: the total potential difference around any cycle is 0. Algebraically, this means that $\exists \mathbf{x} \in \mathbb{R}^{|V|}$ such that $\nabla \mathbf{x} = \mathbf{z}$. The *Dirichlet Principle* states that any solution to the following program gives an absolute potential $\mathbf{x}$ that satisfies Kirchoff's potential law:

$$\min_{\mathbf{x}} \mathbf{x}^\top \Delta \mathbf{x} \text{ s.t. } \mathbf{x}_S = \mathbf{v}_S$$

for source/sinks $S \subset V$ and some voltage constraints $\mathbf{v}_S \in \mathbb{R}^{|S|}$. By Lagrangian calculus, the solution to the above program is given by $\mathbf{x} = \Delta^\dagger \mathbf{v}$ where $\mathbf{v}$ is 0 over S^C and $\mathbf{v}_S$ over S, and $\dagger$ indicates the Moore-Penrose pseudoinverse. The effective resistance between a source $v \in V$ and a sink $w \in V$ is the potential difference required to create a unit flow between them. Hence, the effective resistance between v and w is $r_{v,w} = (\delta_v - \delta_w)^\top \Delta^\dagger (\delta_v - \delta_w)$, where δ_v is the Dirac delta function.

An extremely useful characterization of effective resistance is the random walk interpretation. Let X_t be the location of a random walker on G at time t. The hitting time $H(v,w)$ is then $H(v,w) = \mathbb{E}[\min\{t > 0 : X_t = w\} | X_0 = v]$. We find that the effective resistance is related to the hitting time by,

$$r_{v,w} = \frac{H(v,w) + H(w,v)}{2m}$$

where recall $m = |E|$. The numerator is also known as the commute time. As we will see, this characterization of effective resistance is useful when working with specific graph models.

4.2 UST Wavelet Detector

We will now examine the performance of the wavelet detector, given a spanning tree that is drawn according to a UST. First, we will explore the construction of the UST and examine key properties. The UST is a random spanning tree, chosen uniformly at random from the set of all distinct spanning trees. The foundational Matrix-Tree theorem [20] describes the probability of an edge being included in the UST. The following lemma can be found in [22] and [23].

Lemma 7. *Let G be a graph, $\mathcal{T}$ a draw from* $\mathrm{UST}(G)$*, and let e be any edge in E. Then,*

$$\mathbb{P}\{e \in \mathcal{T}\} = r_e$$

Hence, we can expect that for a given cut in the graph, the cut size in the tree will look like the sum of effective resistances of edges in the cut. While it is infeasible to enumerate all spanning trees of a graph, the Aldous-Broder algorithm is an efficient method for generating a draw from $\mathrm{UST}(G)$ [2]. The algorithm simulates a random walk $\{X_t\}$ on G, stops when all of the vertices have been visited, and defines the spanning tree $\mathcal{T}$ by the edges $\{(X_{H(X_0,v)-1}, v) : v \in V\}$. The computational complexity of the Aldous-Broder algorithm is the expected cover time of G, which is $O(n \log n)$ for expander graphs, d-regular graphs, and many other graph models, but is $O(n^3)$ in the worst case [8]. As we can construct the spanning tree wavelet basis is $\tilde{O}(n)$ time, drawing from $\mathrm{UST}(G)$ dominates the computational complexity of our detector.

In order to control $\|\nabla_{\mathcal{T}} \mathbf{x}\|_0$, we need to control the overlap between a cut in the graph and the UST. Clearly the UST does not independently sample edges, but it does have the well documented property of negative association, that the inclusion of an edge decreases the probability that another edge is included. The following lemma states a concentration result for the UST, based on negative association, and can be found in [12]. The proof is a simple extension of the concentration results in [13].

Lemma 8. *Let $B \subset E$ be a fixed subset of edges, and $|\mathcal{T} \cap B|$ denote the number of edges in $\mathcal{T}$ also in B.*

$$\mathbb{P}\{|\mathcal{T} \cap B| \geq (1+\delta) \sum_{e \in B} r_e\} \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^{\sum_{e \in B} r_e}$$

We use this result to give conditions under which the UST wavelet detector asymptotically distinguishes H_0 from H_1.

Theorem 9. *Let $r_{\max} = \max_{\mathbf{x} \in \mathcal{X}} \sum_{e \in supp(\nabla \mathbf{x})} r_e$ (the maximum effective resistance of the cut of a pattern in $\mathcal{X}$). If*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{r_{\max} \log d_{\max}} \log n \right)$$

then H_0 and H_1 are asymptotically distinguished by the test statistic $\|\mathbf{B}^T \mathbf{y}\|_\infty$ where $\mathbf{B}$ is the UST wavelet basis.

Proof. Let $r_B = \sum_{e \in B} r_e$ for $B \subset E$. By some basic calculus, and the fact that $\log(1+x) \geq x/(1+x/2)$, we see that

$$\left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^{r_B} \leq \exp\left(-\frac{\delta^2 r_B}{2+\delta} \right)$$

Rewriting the Lemma 8, we obtain with probability $> 1 - \gamma$

$$\begin{aligned} |\mathcal{T} \cap B| &\leq r_B + \sqrt{2 r_B \log \frac{1}{\gamma} + \frac{1}{4} (\log \frac{1}{\gamma})^2} + \frac{1}{2} \log \frac{1}{\gamma} \\ &\leq \left(r_B + \sqrt{2 r_B \log \frac{1}{\gamma}} + \log \frac{1}{\gamma} \right) \end{aligned}$$

Now, because $\|\nabla_{\mathcal{T}}\mathbf{x}\|_0 = |\mathcal{T} \cap B|$ for $B = \text{supp}(\nabla_{\mathcal{T}}\mathbf{x})$, we know by Theorem 5 if

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\left(r_B + \sqrt{2r_B \log\frac{1}{\gamma}} + \log\frac{1}{\gamma}\right)\log d_{\mathcal{T}} \log n}\right)$$

then H_0 and H_1 are asymptotically distinguished. The result follows because we guarantee this for all B. □

5 Specific Graph Models

In this section we study our detection problem for several different families of graphs. Specifically, we control the effective resistance r_e for each graph family, which when combined with Theorem 9 gives a lower bound on the SNR for which $||\mathbf{B}^T\mathbf{y}||_\infty$ asymptotically distinguishes H_0 and H_1.

In Theorem 9, we showed that the distinguishability regime depends on the effective resistances of the cuts induced by the class of activation patterns $\mathcal{X}$. On its own, it is not immediately clear that this result is an improvement over the bound in Remark 6 that we would obtain from any spanning tree. However, Foster's theorem highlights why we expect the effective resistance to be less than the cut size.

Theorem 10 (Foster's Theorem [11, 29])**.**

$$\sum_{e\in E} r_e = n - 1$$

Hence, if we select an edge uniformly at random from the graph, we expect its effective resistance to be $(n-1)/m \approx \bar{d}^{-1}$ (the reciprocal of the average degree $\bar{d}$). Hence, we expect the effective resistance of a cut to be $\approx \rho/\bar{d}$, which would be an improvement over the results of Remark 6 and yield a closer match to the lower bounds established in Theorem 1. In several example graphs we will formalize this intuition. We complement these results with two types of simulations verifying different aspects of our theory. The first verifies the upper bound in Lemma 4 for a variety of graph models by plotting $||\mathbf{B}^T\mathbf{x}||_0$ versus $\rho\log(d)\log(n)$ for several randomly generated signals. These plots (see Figure 2) demonstrate the validity of our bound since in all cases $||\mathbf{B}^T\mathbf{x}||_0 \le \rho\log(d)\log(n)$, but, more importantly, the readily-observable linear relationship between these two quantities suggests that one should not expect an improvement on this bound by more than a constant factor.

The second simulation verifies the performance of our spanning tree wavelets detector on various graph models. In Figure 3, we plot the power of our test statistic (with Type I error fixed at 5%) as a function of signal strength μ for several values of n, where we allow ρ to scale with n to ensure a non-empty $\mathcal{X}$. These simulations demonstrate that as expected for sufficiently large signal strength, our statistic can separate H_0 from H_1. More importantly, the threshold signal strength for which detection is possible increases with n and ρ, as predicted by our theory.

5.1 Edge Transitive Graphs

An edge transitive graph, G, is one such that for any edges e_0, e_1, there is a graph automorphism that maps e_0 to e_1. Examples of edge transitive graphs include the l-dimensional torus and the complete graph K_n. For such a graph, every edge has the same effective resistance, and Foster's Theorem then shows that $r_e = (n-1)/m$ where m is the number of edges. Moreover since edge transitive graphs must be d-regular for some degree d, we see that $m = \Theta(nd)$ so the $r_e = \Theta(1/d)$. This leads us to the following corollary, which we note matches the lower bound in Theorem 1 modulo logarithmic terms if $\rho/d \le \sqrt{n}$:

Corollary 11. *Let G be edge transitive with common degree d. Then for each edge $e \in E(G)$, $r_e = (n-1)/m$. Consider the hypothesis testing problem* (1) *where the set $\mathcal{X}$ is parameterized by ρ. If:*

$$\frac{\mu}{\sigma} = \omega\left(\sqrt{\frac{\rho}{d}\log d \log n}\right)$$

Then the UST wavelet detector, $||\mathbf{B}^T\mathbf{y}||_\infty$, asymptotically distinguishes H_0 and H_1.

5.2 kNN Graphs

Oftentimes in applications, the graph topology is derived from data. In this case, the randomness of the data means that the graph itself is inherently random. Commonly, these graphs are modeled as random geometric graphs, and in this section we will devote our attention to the *symmetric k-nearest neighbor graphs.* Specifically, suppose that $\mathbf{z}_1, ..., \mathbf{z}_n$ are drawn i.i.d. from a density p supported over $\mathbb{R}^D$. Then we form the graph G over V by connecting vertices i, j if $\mathbf{z}_i$ is amongst the k-nearest neighbors of $\mathbf{z}_j$ or vice versa. Some regularity conditions of p are needed for our results to hold; they can be found in [32].

To bound the effective resistance r_e, Corollary 9 in [32] shows that $H_{ij}/2m \to 1/d_j$ and by the definition of r_e we see that $r_{ij} \to \frac{1}{d_i} + \frac{1}{d_j} \le \frac{2}{k}$, since $d_i \ge k$ for each i. A formal analysis leads to the following corollary, which we prove in Appendix C with more precise concentration arguments:

Corollary 12. *Let G be a k-NN graph with $k/n \to 0$ and $k(k/n)^{2/D} \to \infty$ and where the density p satisfies the regularity conditions in [32]. Consider the hypothesis testing problem* (1) *where the set $\mathcal{X}$ is parameterized by ρ. If:*

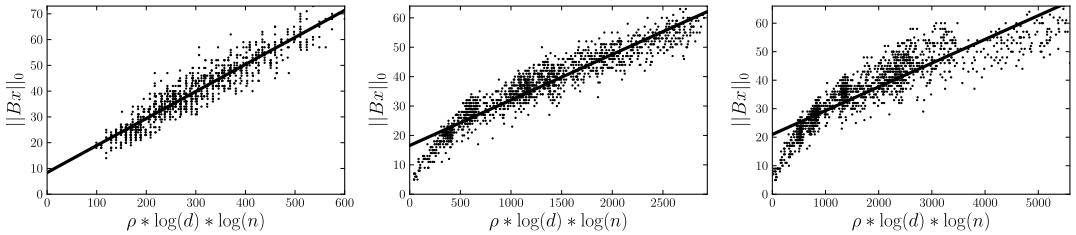

Figure 2: Spanning tree wavelet basis sparsity as a function of $\rho \log d \log n$ for, from left to right, 2-dimensional torus, k-NN, and ϵ graphs.

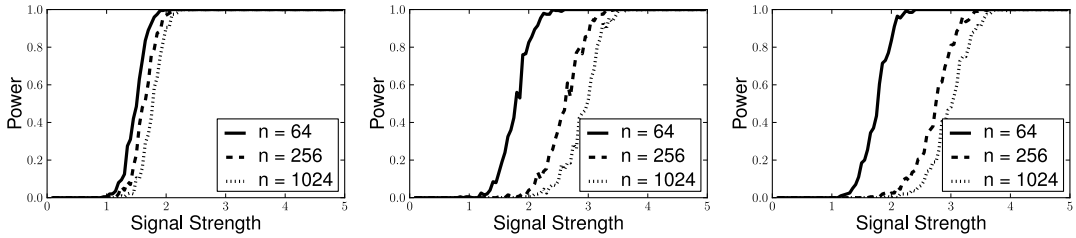

Figure 3: Power as a function of signal strength for different values of n for 2-dimensional torus, k-NN, and ϵ graphs. ρ scales like $\sqrt{n}$, $n^{2/3}$ and $n^{4/5}$ respectively.

$$\frac{\mu}{\sigma} = \omega(\sqrt{\frac{\rho}{k}\log k \log n})$$

Then the UST wavelet detector, $||\mathbf{B}^T\mathbf{y}||_\infty$, asymptotically distinguishes H_0 and H_1.

5.3 ϵ-Graphs

The ϵ-graph is another widely used random geometric graph in machine learning and statistics. As with the k-NN graph, the vertices are embedded into $\mathbb{R}^D$ and edges are added between pairs of vertices that are within distance ϵ of each other. As with the k-NN graph, Corollary 8 from [32] shows that $H_{ij} \to m/d_j$ for each pair of vertices. Thus, we know that $r_{ij} \to 1/(d_i) + 1/(d_j)$. If the density p from which we draw data points is bounded from below by some constant, then we can uniformly lower bound all of the degrees d_i using fairly elementary concentration results, which results in an upper bound on r_e. Formalizing this intuition, we have the following corollary, which we prove in Appendix C:

Corollary 13. *Let G be a ϵ-graph with points $X_1, \ldots X_n$ drawn from a density p satisfying the regularity conditions in [32] and lower bounded by some constant $p_{\min}$, and upper bounded by a constant $p_{\max}$ (independent of n). Let $\epsilon \to 0, n\epsilon^{D+2} \to \infty$ and consider the testing problem* (1) *where the set $\mathcal{X}$ is parameterized by ρ. $||\mathbf{B}^T\mathbf{y}||_\infty$ asymptotically distinguishes H_0 and H_1 if:*

$$\frac{\mu}{\sigma} = \omega(\sqrt{\frac{\rho}{n\epsilon^D}\log(n\epsilon^D)}\log n)$$

6 Discussion

We studied the detection of piece-wise constant activation patterns over graphs, and provided a necessary condition for the asymptotic distinguishability of signals that are assumed to have few discontinuities. We gave a novel spanning tree wavelet construction, that is the extension of the Haar wavelet basis, for arbitrary graphs and proposed a detector relying on the largest wavelet coefficient obtained by projecting the observations onto the basis. The wavelet detector constructed using a uniform spanning tree was shown to have strong theoretical guarantees that in many cases gives us near optimal performance. This means that under adversarial choice of patterns, our randomized algorithm asymptotically distinguishes H_0 from H_1 at near optimal signal-to-noise ratios. In this paper, we primarily focused on detection, however, the spanning tree wavelet construction we propose and resulting sparsifying properties of the basis could be of independent interest for compression and denoising. Also, we focused on structures given by undirected, unweighted, known graphs. There are many natural extensions of this work including weighted and directed edges and how to construct detectors when we are not certain about the graph structure.

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.

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] D. Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, 3(4):450–465, 1990.

[3] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detection of correlations. *The Annals of Statistics*, 40(1):412–435, 2012.

[4] 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.

[5] 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.

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

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[8] A. Broder. Generating random spanning trees. *Foundations of Computer Science*, 1989.

[9] R. Coifman and M. Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006.

[10] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, pages 1200–1224, 1995.

[11] R. Foster. The average impedance of an electrical network. *Contributions to Applied Mechanics (Reissner Anniversary Volume)*, pages 333–340, 1949.

[12] W. Fung and N. Harvey. Graph sparsification by edge-connectivity and random spanning trees. *Arxiv preprint arXiv:1005.0265*, 2010.

[13] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM (JACM)*, 53(3):324–360, 2006.

[14] M. Gavish, B. Nadler, and R. R. Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *ICML*, pages 367–374, 2010.

[15] P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732, 2010.

[16] D. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[17] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*. Springer Berlin, 1998.

[18] Y. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability and its Applications*, 31:333, 1987.

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

[20] G. Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497–508, 1847.

[21] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. *Advances in Neural Information Processing Systems*, 2011.

[22] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[23] R. Lyons and Y. Peres. Probability on trees and networks. Book in preparation., 2000.

[24] S. Mallat. *A wavelet tour of signal processing*. Academic Pr, 1999.

[25] J. Pearl and M. Tarsi. Structuring causal trees. *Journal of Complexity*, 2(1):60–77, 1986.

[26] I. Ram, M. Elad, and I. Cohen. Generalized tree-based wavelet transform. *Signal Processing, IEEE Transactions on*, 59(9):4199–4209, 2011.

[27] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pages 985–1012, 2009.

[28] A. Singh, R. Nowak, and R. Calderbank. Detecting weak but hierarchically-structured patterns in networks. In *Artificial Intelligence and Statistics, AISTATS*, 2010.

[29] P. Tetali. Random walks and the effective resistance of networks. *Journal of Theoretical Probability*, 4(1):101–109, 1991.

[30] United States Geological Survey. Shakemaps, November 2012.

[31] B. Vidakovic. *Statistical modeling by wavelets*. Wiley Online Library, 1999.

[32] U. Von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large graphs are often misleading. *ReCALL*, 2010.

A Proofs for Section 2

Proof of Proposition 2. By Cirelson's theorem, we know that

$$\mathbb{P}\{\max_{v\in C}\epsilon_v \geq \mathbb{E}\max_{v\in C}\epsilon_v + u\} \leq e^{-u^2/2\sigma^2}$$

and we have that $\mathbb{E}\min_{v\in C}\epsilon_v$ is within a constant factor of $-\sigma\sqrt{2\log|C|}$ by the Majorizing-Measure theorem. Hence, under H_1^C we have that $\max_{v\in C} y_v \geq \mu/\sqrt{|C|} + L^{-1}\sigma\sqrt{2\log|C|} - \sigma\sqrt{2\log(1/\delta)}$ with probability at least $1-\delta$ for some universal constant L. By similar reasoning, under H_0 we have that $\max_{v\in V}|y_v| \leq L\sigma\sqrt{2\log n} + \sigma\sqrt{2\log(2/\delta)}$. We have analogous opposite bounds that prove the necessary condition. Hence, a threshold, τ, distinguishes H_0 and H_1 if

$$\begin{aligned}\mu/\sqrt{|C|} + L^{-1}\sigma\sqrt{2\log|C|} - \sigma\sqrt{2\log 1/\delta} &> \tau \\ &> L\sigma\sqrt{2\log n} + \sigma\sqrt{2\log 1/\delta}\end{aligned}$$

which occurs if $\mu/\sigma - \sqrt{2|C|\log n} = \omega(1)$. Choosing C such that $|C| \geq n/2$ we have the sufficient condition. □

Proof of Proposition 3. Under H_0, $\frac{1}{n}\sum_{v\in V} y_v$ is normally distributed with mean 0 and variance σ^2/n. Meanwhile under H_1^C, $\frac{1}{n}\sum_{v\in V} y_v$ is normally distributed with mean $\mu\sqrt{|C|}/n$ and variance σ^2/n. Hence, the test statistic asymptotically distinguished H_0 from H_1 if and only if

$$\Phi\left(-\frac{\mu}{\sigma}\sqrt{\frac{|C|}{n}}\right) = o(1)$$

where Φ is the CDF of the standard normal. □

B Proofs for Section 3

B.1 Properties of the Spanning Tree Wavelet Basis

Lemma 14. *The output* $\mathbf{B}$ *of the Spanning Tree Wavelet Construction has the following properties:*

1. $\mathbf{B}$ *is an orthonormal basis of* $\mathbb{R}^n$*, in particular there are* n *vectors.*
2. $\mathbf{B}$ *can be computed in* $O(n\log n\log d_{\mathcal{T}})$ *time.*

Proof. Before we dive into the Spanning Tree Wavelet Construction, we must analyze the FormWavelets routine. FormWavelets operates on the subtrees c_i as if they form a chain structure, and it constructs the Haar Wavelet basis over this chain structure. A fairly straightforward inductive argument shows that, coupled with the all-ones vector, the output of FormWavelets is an orthonormal basis over the subtrees c_i. In particular, a linear combination of these vectors can be used to obtain a vector that is constant and non-zero on a single subtree and zero elsewhere.

Analyzing the Spanning Tree Wavelet Construction also proceeds by induction. On the two-node graph, running the Spanning Tree Wavelet Construction gives a basis with two elements $(1/\sqrt{2}, 1/\sqrt{2})^T$ and $(1/\sqrt{2}, -1/\sqrt{2})^T$ which is clearly an orthobasis. For any tree $\mathcal{T}$, FormWavelets gives an orthonormal basis over the subtrees, and the inductive hypothesis gives the basis for each of these individual subtrees. By induction, recursing on each subtree $\mathcal{T}_i$ gives a set of orthonormal vectors that are also orthogonal to $\mathbf{1}_{\mathcal{T}_i}$. Consequently $\mathbf{B}$ is an orthobasis.

Computing the basis involves finding a balancing vertex and then recursively forming the basis elements. Finding a balancing vertex can be done in linear time by precomputing all of the subtree sizes using a depth-first tree traversal. Computing the basis just involves another pre-order traversal of the tree, where at each level in the tree we construct $O(\lceil\log d_{\mathcal{T}}\rceil)$ basis elements, with a total of $O(n\lceil\log d_{\mathcal{T}}\rceil)$ non-zero coefficients. Repeating this across the $\lceil\log n\rceil$ levels gives the running time bound. □

B.2 Proof of Lemma 4

Before we proceed with the proof, we state and prove two results on the performance of the algorithm:

Lemma 15. *Let* $\mathcal{T}$ *be a tree.* FindBalance *returns a vertex* v *such that the largest connected component of* $\mathcal{T}\backslash v$ *is of size at most* $\lceil|\mathcal{T}|/2\rceil$ *in* $O(|\mathcal{T}|)$ *time.*

Proof. Let the objective be the size of the largest connected components of $\mathcal{T}\backslash v$. Every move in *FindBalance* reduces the objective by at least 1 and the objective can be at most $|\mathcal{T}|-1$ so it must terminate in less than $|\mathcal{T}|$ moves. Now at any step of *FindBalance*, if the objective is greater than $\lceil|\mathcal{T}|/2\rceil$, the cumulative size of the remaining connected components is less than $\lfloor|\mathcal{T}|/2\rfloor$. Hence, in the next step the connected component formed by these is less than $\lceil|\mathcal{T}|/2\rceil$. Thus, the program cannot terminate at a move directly after the objective is greater than $\lceil|\mathcal{T}|/2\rceil$. □

We will also require the following claim. Indeed, controlling the depth of the recursion in the wavelet construction is the sine qua non for controlling the sparsity, $\|\mathbf{B}^T\mathbf{x}\|_0$.

Claim 16. *The wavelet construction has recursion depth at most* $\lceil\log d_{\mathcal{T}}\rceil\lceil\log n\rceil$.

Proof. Whenever *FormWavelet* is applied it increases the height of the dendrogram by at most $\lceil \log d_{\mathcal{T}} \rceil$. By lemma 15 the size of the remaining components is halved, so the algorithm terminates in at most $\lceil \log n \rceil$ steps. □

Proof of Lemma 4. We will show that any edge $e \in \mathcal{T}$ supports at most $\lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil$ basis elements in $\mathbf{B}$, and this will imply the result. We will say that an edge e supports a basis element $\mathbf{b}$ if $e \subseteq supp(\nabla_{\mathcal{T}} \mathbf{b})$. It follows that for a basis element $\mathbf{b}$, if $\mathbf{b}^T \mathbf{x} \neq 0$ then $\exists e$ that supports b. Let no_basis(e) be the number of basis elements that are supported by e (no_basis(e) $= 0$ if $e \notin supp(\nabla_{\mathcal{T}} \mathbf{x})$). We then have

$$||\mathbf{B}^T \mathbf{x}||_0 \leq \sum_{e \in supp(\nabla_{\mathcal{T}} \mathbf{x})} \text{no\_basis}(e)$$

Consider some edge e. If e supports some subtree $\mathcal{T}_{sub}$ (we use this interchangeably with supporting a basis element formed by partitioning $\mathcal{T}_{sub}$ into two groups), then e supports at most one of $\mathcal{T}_{sub}$'s subtrees. This implies that no_basis(e) is upper bounded by the depth of the recursion. By the claim, we find that,

$$\begin{aligned} ||\mathbf{B}^T \mathbf{x}||_0 &\leq \sum_{e \in supp(\nabla_{\mathcal{T}} \mathbf{x})} \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil \\ &\leq ||\nabla_{\mathcal{T}} \mathbf{x}||_0 \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil \end{aligned}$$

proving the first claim. The second claim is obvious from the fact that $\mathcal{T}$ contains a subset of the edges in $\mathcal{G}$, so every cut has larger cut size in $\mathcal{G}$ than it does in $\mathcal{T}$. □

B.3 Proof of Theorem 5

Proof. Under the null $\mathbf{x} = 0$, and we have that

$$||\mathbf{B}^T \mathbf{y}||_\infty = ||\mathbf{B}^T \boldsymbol{\epsilon}||_\infty < \sigma \sqrt{2 \log(n/\delta)}$$

with probability at least $1 - \delta$. So, as long as $\tau = \sigma \sqrt{2 \log(n/\delta)}$ then we control the probability of false alarm (type 1 error). For a element $\mathbf{x}$ of the alternative, let the index, i^*, achieve the maximum of $\mathbf{B}^T \mathbf{x}$ (i.e. $||\mathbf{B}^T \mathbf{x}||_\infty = |\mathbf{B}^T \mathbf{x}|_{i^*}$). Then $|\mathbf{B}^T \mathbf{y}|_{i^*} \geq |\mathbf{B}^T \mathbf{x}|_{i^*} - \sigma \sqrt{2 \log(1/\delta)}$ with probability at least $1 - \delta$ and

$$|\mathbf{B}^T \mathbf{x}|_{i^*}^2 = ||\mathbf{B}^T \mathbf{x}||_\infty^2 \geq \frac{\sum_{i:(\mathbf{B}^T \mathbf{x})_i \neq 0} (\mathbf{B}^T \mathbf{x})_i^2}{||\mathbf{B}^T \mathbf{x}||_0} = \frac{||\mathbf{x}||_2^2}{||\mathbf{B}^T \mathbf{x}||_0}$$

Taking square roots and combining this with Lemma 4,

$$||\mathbf{B}^T \mathbf{x}||_\infty \geq \frac{||\mathbf{x}||_2}{\sqrt{||\nabla_{\mathcal{T}} \mathbf{x}||_0 \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil}}$$

from which we have the result that under H_1,

$$||\mathbf{B}^T \mathbf{y}||_\infty \geq \frac{||\mathbf{x}||_2}{\sqrt{||\nabla_{\mathcal{T}} \mathbf{x}||_0 \lceil \log d_{\mathcal{T}} \rceil \lceil \log n \rceil}} - \sigma \sqrt{2 \log(1/\delta)}$$

Furthermore, because the first wavelet coefficient is just the vertex average, $\mathbf{B}_0^T \mathbf{y}$, we know that under H_1 with probability at least $1 - \delta$,

$$|\mathbf{B}_0^T \mathbf{y}| \geq \sqrt{\frac{|C|}{n}} \mu - \sigma \sqrt{2 \log(1/\delta)}$$

Forcing the maximum of these lower bounds to be greater than τ gives us our result. □

C Proofs For Section 5

C.1 Proof of Corollary 12

First we restate Corollary 9 from [32]:

Corollary 17. *Consider an unweighted symmetric or mutual k-NN graph built from a sequence $X_1, \ldots, X_n$ drawn i.i.d. from a density p. Then there exists constants c_1, c_2, c_3 such that with probability at least $1 - c_1 n \exp(-k c_2)$ we have uniformly for all $i \neq j$ that:*

$$\left| \frac{k}{2m} H_{ij} - \frac{k}{d_j} \right| \leq c_3 \frac{n^{2/d}}{k^{1+2/d}}$$

Proof of Corollary 12. We focus on the symmetric k-NN graph in which we connect v_i to v_j if v_i is in the k-nearest neighbors of v_j or vice versa. In this graph, every node has degree $\geq k$ which will be crucial in our analysis. Our goal is to bound the effective resistance of every edge, so that we can subsequently bound r_{max} and apply Corollary 9. From the definition of r_e we have:

$$\begin{aligned} r_{ij} &= \frac{1}{2} \left(\frac{H_{ij}}{m} + \frac{H_{ji}}{m} \right) \\ &\leq 2 c_3 \frac{n^{2/d}}{k^{2+2/d}} + \frac{1}{d_i} + \frac{1}{d_j} \\ &\leq 2 c_3 \frac{n^{2/d}}{k^{2+2/d}} + \frac{2}{k} \end{aligned}$$

Where the first line is the definition of r_{ij}, the second line follows from Corollary 17 and the last line follows from the fact that $d_i \geq k$ for each vertex. Since $k(k/n)^{2/d} \to \infty$, we see that $r_{ij} = O(\frac{1}{k})$. Moreover, with this scaling of k, that the probability in Corollary 17 is going to 1. We can therefore bound r_{max} as:

$$r_{max} \leq \rho \left(2 c_3 \frac{n^{2/d}}{k^{2+2/d}} + \frac{2}{k} \right) = O\left(\frac{\rho}{k} \right)$$

Since the first term is going to zero with n. Plugging in this bound on r_{max} into Theorem 9 gives the result (and substituting $d = k$). □

C.2 Proof of Corollary 13

As before, we first state Corollary 8 from [32]:

Corollary 18. *Consider an unweighted ϵ-graph built from the sequence $X_1, \ldots, X_n$ drawn i.i.d. from the density p .Then there exists constants $c_1, \ldots c_5 > 0$ such that with probability at least $1 - c_1 n \exp(-c_2 n\epsilon^D) - c_3 \exp(-c_4 n\epsilon^D)/\epsilon^D$, we have uniformly for all $i \neq j$ that:*

$$\left| \frac{n\epsilon^D}{2m} H_{ij} - \frac{n\epsilon^D}{d_j} \right| \leq \frac{c_5}{n\epsilon^{D+1}}$$

Proof of Corollary 13. Some manipulation of the result in Corollary 18 reveals that:

$$H_{ij} \leq \frac{2m}{d_j} + \frac{2c_5 m}{n^2 \epsilon^{2D+2}}$$

Under our scaling, the second term goes to zero and the probability in Corollary 18 goes to one, so $H_{ij} = O(m/d_j)$. We will now give a lower bound on d_j. If X_i is in the ball of radius ϵ centered at X_j, then we connect X_i and X_j. Thus d_j is exactly the number of vertices in the $B(X_j; \epsilon)$. The regularity condition on p in [32] requires that there exists constants α and ϵ_0 such that for all $\epsilon < \epsilon_0$ and for all $x \in \text{supp}(p)$, $\text{vol}(B(x;\epsilon) \cap \text{supp}(p)) \geq \alpha \text{vol}(B(x;\epsilon))$. By this fact, the fact that the density is lower bounded by $p_{\min}$, and by the fact that $\epsilon \to 0$, we know that for sufficiently large n, $p(B(X_j;\epsilon)) \geq p_{\min} \alpha c_D \epsilon^D$ where $c_D \epsilon^D$ is the volume of a D-dimensional ball of radius ϵ. The probability that $X_i \in B(X_j;\epsilon)$ is distributed as a Bernoulli random variable with mean $\geq \alpha p_{min} c_D \epsilon^D$. By Hoeffding's inequality and a union bound we get that:

$$d_j \geq n\alpha p_{min} c_D \epsilon^D + \sqrt{n \log(n)} = \Omega(n\epsilon^D)$$

for all vertices j with probability $\geq 1 - 1/n$. Using the definition of $r_{i,j}$ along with the bound on H_{ij} and d_j we have that uniformly for all pairs i, j:

$$r_{i,j} = O(\frac{1}{n\epsilon^D})$$

Plugging in this bound into Theorem 9 gives us the result. □