Detecting Anomalous Activity on Networks With the Graph Fourier Scan Statistic

James Sharpnack, *Member, IEEE*, Alessandro Rinaldo, and Aarti Singh

***Abstract*—We consider the problem of deciding, based on a single noisy measurement at each vertex of a given graph, whether the underlying unknown signal is constant over the graph or there exists a cluster of vertices with anomalous activation. This problem is relevant to several applications such as surveillance, disease outbreak detection, biomedical imaging, environmental monitoring, etc. Since the activations in these problems often tend to be localized to small groups of vertices in the graphs, we model such activity by a class of signals that are elevated over a (possibly disconnected) cluster with low cut size relative to its size. We analyze the corresponding generalized likelihood ratio (GLR) statistics and relate it to the problem of finding a sparsest cut in the graph. We develop a convex relaxation of the GLR statistic based on spectral graph theory, which we call the graph Fourier scan statistic (GFSS). In our main theoretical result, we show that the performance of the GFSS depends explicitly on the spectral properties of the graph. To assess the optimality of the GFSS, we prove an information theoretic lower bound for the detection of anomalous activity on graphs. Because the GFSS requires the specification of a tuning parameter, we develop an adaptive version of the GFSS. Using these results, we are able to characterize in a very explicit form the performance of the GFSS on a few notable graph topologies. We demonstrate that the GFSS can efficiently detect a simulated Arsenic contamination in groundwater.**

***Index Terms*—Adaptive signal detection, anomaly detection, graph filters, hypothesis testing.**

I. Introduction

In this article, we will take a statistical approach to detecting signals that are localized over a graph. Signal detection on graphs is relevant in a variety of scientific areas, such as surveillance, disease outbreak detection, biomedical imaging, detection using a sensor network, gene network analysis, environmental monitoring and malware detection over a computer network. Recently, the use of graphs to extend traditional methods of signal processing to irregular domains has been proposed [1]–[4]. While this work has largely focused on extending Fourier and wavelet analysis to graphs, little is known about the statistical efficiency of the recently proposed methodology. We show that the Fourier transform over graphs, defined in [5], can be used to detect anomalous patterns over graphs by constructing the Graph Fourier Scan Statistic (GFSS), a novel statistic based on spectral graph theory. We demonstrate the connection between the GFSS and the recently proposed Spectral Scan Statistic [6], and provide strong theoretical guarantees.

Throughout this work, we will assume that there is a known, fixed, undirected graph with p vertices (denoted by the set $V = \{1, \ldots, p\}$), m edges denoted by pairs $(i, j) \in E \subseteq V \times V$, and $p \times p$ weighted adjacency matrix $\mathbf{W}$ (where the weight $W_{i,j} = W_{j,i} \geq 0$ denotes the 'strength' of the connection between vertices $(i, j) \in E$). Assume that we observe a single high-dimensional measurement $\mathbf{y}$ over the graph, whereby for each vertex of the graph, $i \in V$, we make a single, Gaussian-distributed observation y_i. In the context of sensor networks, the measurements y_i are the values reported by each sensor, and the edge weights reflect beliefs about how similar the measurements of two sensors should be. The measurements y_i are noisy, and we are interested in determining if there is a region within the network where these observations are abnormally high. Specifically, we are concerned with the basic but fundamental task of deciding whether there is a 'cluster' of vertices within the graph, $C \subset V$, such that in expectation the observation, $\mathbb{E}[y_i]$, is larger for $i \in C$ than for $i \notin C$. In Section II, we will define precisely our statistical framework, including the assumptions placed on the cluster C and observations $\mathbf{y}$ in relation to the graph. In order to motivate the problem and introduce the GFSS, let us consider the following real data example.

A. Arsenic Groundwater Concentrations in Idaho

Groundwater contamination remains a serious issue globally, where aging infrastructure, shifting population densities, and climate change are among the contributing factors. A study published in 1999, reports levels of Arsenic (As) contamination measured in 20,043 wells throughout the United States [7]. In order to illustrate the usefulness of the GFSS, we analyze the As concentration with the purpose of determining if there is a region that has elevated incidence of high As levels. We will focus on the tested wells within Idaho, which was selected arbitrarily from the other US states. We construct a graph between the wells, where each vertex is a tested well, by creating an edge between two vertices (wells) if either is the kth nearest neighbor of the other (See Fig. 1). For easy visualization, we subsampled the wells by randomly selecting 219 (roughly 10%) of the 2,191 of the Idaho wells. We preprocessed the data by forming the indicator variable, y_i, which was 1 if the measurement made at the

Manuscript received March 28, 2014; revised September 05, 2014 and February 27, 2015; accepted August 25, 2015. Date of publication September 25, 2015; date of current version December 14, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Akbar Sayeed. This research is supported in part by AFOSR under grant FA9550-10-1-0382, NSF under grants DMS-1223137 and IIS-1116458.

J. Sharpnack is with the Statistics Department, University of California Davis, Davis, CA 95616 USA (e-mail: jsharpna@gmail.com).

A. Rinaldo and A. Singh are with Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: arinaldo@cmu.edu; aarti@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2481866

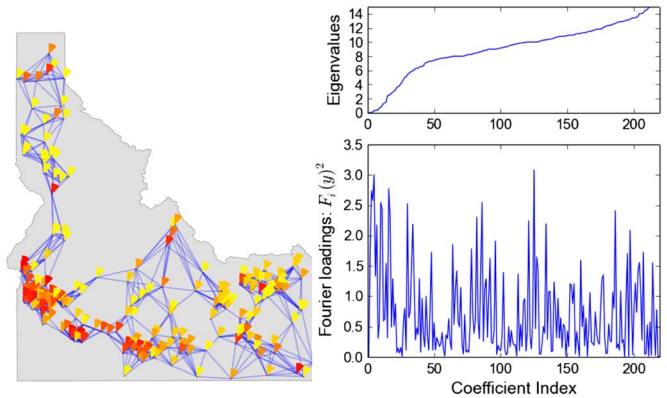

Fig. 1. (**Ground-water Arsenic Concentrations**) The As concentrations within Idaho after (left) where red depicts higher concentrations and yellow depicts lower concentrations. The ordered eigenvalues of the well network are plotted (top right) and the Fourier loadings $\{\mathcal{F}_i(\mathbf{y})^2 = (\mathbf{u}_i^\top \mathbf{y})^2\}_{i=1}^{219}$ are plotted (bottom right). The index in the x-axis for the eigenvalues and Fourier loadings match so that the ith index corresponds to the pair $\lambda_i, \mathbf{u}_i$.

ith well was greater than 10 ppm and 0 otherwise (and we will denote the p dimensional vector, $\mathbf{y}$). The statistical problem that we address in this paper is testing if there is a well-connected set of wells, C, such that the measurements, $\mathbf{y}$, are abnormally high within the active set C.

B. Graph Fourier Scan Statistic

Traditional statistical methods, such as wavelet denoising (i.e. Haar and Daubechies wavelets) that employ the standard multi-resolution analysis (see [8]) are not adapted to irregular domains and sensor distributions that are not grid-like. With this in mind, a natural algorithm for the detection of such anomalous clusters of activity is the generalized likelihood ratio test (GLRT) (also known as the scan statistic or matched filter). Under a signal plus Gaussian noise model, this procedure entails scanning over all permitted clusters and hence is computationally very intensive. In [6], the Spectral Scan Statistic (SSS) was proposed as a relaxation of the combinatorial GLRT. The statistical power of the detector, constructed by thresholding the SSS, was characterized using spectral graph theory. In this paper, we propose another detector which is a low-pass filter based on the graph Fourier transform.

We will show that the resulting Graph Fourier Scan Statistic (GFSS), is in fact a further relaxation of the SSS, but because of its particular form it allows us to very precisely characterize its statistical power and construct an adaptive counterpart. We will begin by introducing a graph Fourier transform, which has been previously proposed in [5] (but other transforms have been proposed, as in [2]). Through the graph Fourier transform, we will define the GFSS, which we introduce next. Define the *combinatorial Laplacian* matrix $\Delta = \mathbf{D} - \mathbf{W}$, where $\mathbf{D} = \text{diag}\{d_i\}_{i=1}^p$ is the diagonal matrix of vertex degrees, $d_i = \sum_{j=1}^p W_{i,j}$. We will denote the eigenvalues and eigenvectors of Δ with $\{\lambda_i, \mathbf{u}_i\}_{i=1}^p$ respectively, where we order the eigenvalues in increasing order. Hence, if $\mathbf{U}$ is the $p \times p$ matrix where the ith column is the eigenvector $\mathbf{u}_i$ and $\Lambda = \text{diag}\{\lambda_i\}_{i=1}^p$ then we have

$$\Delta = \mathbf{U}\Lambda\mathbf{U}^\top.$$

For the measurement vector $\mathbf{y}$ over the vertices, the graph Fourier transform is $\mathcal{F}(\mathbf{y}) = \mathbf{U}^\top\mathbf{y}$. Then the coordinate $\mathcal{F}_i(\mathbf{y}) = \mathbf{u}_i^\top\mathbf{y}$ for i small are the low frequency components of $\mathbf{y}$ and for i large are the high frequency components. In fact, the eigenbasis of the graph Laplacian is commonly used for statistical methods over graphs and point clouds in machine learning. Much of this work has focused on dimension reduction and clustering [9]–[11], there has been some work on using the Laplacian for regression and testing [12], [13]. We demonstrate with the GFSS, and its theoretical analysis, another aspect of the Laplacian eigenbasis in a statistical context.

In order to construct the GFSS, consider a low-pass filter, G, that passes the low-frequency components of $\mathbf{y}$ and attenuates (shrinks) the high-frequency components,

$$G(\mathbf{y}) = \sum_{i=2}^p h(\lambda_i)\left(\mathbf{u}_i^\top\mathbf{y}\right)\mathbf{u}_i, \quad h(\lambda_i) = \min\left\{1, \sqrt{\frac{\rho}{\lambda_i}}\right\},$$

where $\rho > 0$ is a tuning parameter. Because λ_i is increasing in i the attenuation factor, $h(\lambda_i)$, is 1 for i small enough and is non-increasing in i. Then we *define the Graph Fourier Scan Statistic* as the energy of the attenuated signal with an adjustment for the amount of the attenuation (we let $\|.\|$ denote the ℓ_2 norm),

$$\begin{aligned}\hat{t} &= \|G(\mathbf{y})\|^2 - \sum_{i=2}^p h(\lambda_i)^2 \\ &= \sum_{i=1}^p \min\left\{1, \frac{\rho}{\lambda_i}\right\}\left[\left(\mathbf{u}_i^\top\mathbf{y}\right)^2 - 1\right]. \end{aligned} \quad (1)$$

We will explain why the first eigenvector $\mathbf{u}_1$ is ignored in Section IV (notice that the index of the sum begins at 2). We should note here that for any graph Laplacian, $\lambda_1 = 0$ and $u_{1,i} = p^{-1/2}$ for all $i \in V$. If the GFSS is abnormally large then a large amount of the signal $\mathbf{y}$ is in the low-frequency components. We will see in Section III-A that this occurs when there is a well-connected cluster C of vertices that have an abnormally large signal.

In Fig. 1, we have displayed the eigenvalues in increasing order and the squared graph Fourier coefficients (where the index of the eigenvalues matches the index of the coefficients), $\mathcal{F}_i(\mathbf{y})^2$, for the Idaho As concentrations. Because the linear filter $G(\mathbf{y})$ focuses the sensing energy on the low frequency components, the GFSS will be high if the Fourier loadings $\mathcal{F}_i(\mathbf{y})^2$ are large for smaller i.

By forming a k-nearest neighbor (kNN) graph over all 2,191 wells in Idaho with $k = 8$ and applying the GFSS with $\rho = \lambda_{109}$ (the 109th smallest eigenvalue, which was selected simply because $109 = \lfloor 0.05(2191) \rfloor$). The GFSS statistic evaluates to 697.1 and we can obtain a P-value $< 10^{-5}$ by a permutation test (explained in Section VI-A). This indicates that we can be confident that the probability of obtaining a high As measurement is non-constant throughout the graph.

Recall that we also subsampled the well measurements, to form a kNN graph ($k = 8$) over 219 wells (as shown in Fig. 1). By selecting $\rho = \lambda_{10}$ which is selected by the same rule as before ($10 = \lfloor 0.05(219) \rfloor$), the GFSS also obtains a P-value $< 10^{-5}$. So, despite the fact that we used 10% of the samples in this dataset, we can still conclude with confidence that the

signal is not identically distributed over the graph. With this knowledge, targeted ground-water treatment could be recommended and further statistical analysis for locating the contamination would be warranted. After we make a thorough case for the GFSS from a theoretical perspective, we will return to the As detection example in Section VI-A.

C. Related Work

The problem of statistical hypothesis testing of graph-structured activation has received some attention recently. The GLRT for graphs, also known as the graph scan statistic, is discussed in [14], [15]. Theoretical properties of the GLRT for some specific topologies and specific signal classes have also been derived, e.g. detecting an interval in a line graph or geometric shapes such as rectangles, disks or ellipses in a lattice graph [16], path of activation in a tree or lattice [17], or nonparametric shapes in a lattice graph [18]. In these settings, scanning over the entire signal class or over an epsilon-net for the signal class is often feasible and has been shown to have near-optimal statistical performance. However, for general graphs and signal classes these detectors are infeasible, either because the scan involves too many patterns or due to lack of constructive ways to obtain an epsilon-net. While there has been some work on developing fast graph subset scanning methods [19], these greedy methods sacrifice statistical power. Also, there is work on developing Fourier basis and wavelets for graphs (cf. [1] and references therein), which can potentially serve as an epsilon-net, however the approximation properties of such basis are not well characterized. An exception is [20] where graph wavelets were constructed using a spanning tree and statistical properties of the corresponding wavelet detector have been characterized. In [21], the authors consider the complete graph and study detection under some combinatorial classes such as signals supported over cliques, bi-cliques, and spanning trees. They establish lower bounds on the performance of any detector, and provide upper bounds for some simple but sub-optimal detectors such as averaging all node observations and thresholding.

We build on our previous findings in [6] where the Spectral Scan Statistic was proposed as a convex spectral relaxation of the GLRT and characterize its statistical performance. In another recent work [22], we have also developed a different convex relaxation of the GLRT using Lovasz extension and characterized its properties for detecting graph-structured signals. A comparison of our prior work [6], [20], [22] appears in [23]. Despite the empirical success of the SSS in [6], the statistical guarantees made are in some cases dominated by the guarantees obtained for the energy test statistic (to be introduces in Section III-B) which does not take the graph structure into account. The GFSS attains superior theoretical performance which always outperforms the energy statistic (except in cases in which the graph structure is misleading). Moreover, because the GFSS is formed by attenuating high frequency components via the graph Fourier transform (as in [5]), this paper provides a statistical justification for the use of the combinatorial Laplacian to derive a graph Fourier analysis. Furthermore, the SSS requires perfect knowledge of the tuning parameter ρ, which is not known in general. To this end, we form the adaptive GFSS, which automatically selects ρ. In practice, the adaptive GFSS significantly outperforms the GFSS with a heuristic choice of ρ. The GFSS also may be preferable to more complicated procedures because it is based on a linear filter of the measurements $\mathbf{y}$, which in some computational settings may be advantageous.

D. Contributions

Our contributions are as follows. (1) We examine a new alternative hypothesis, which we call the graph-structured H_1, which generalizes the piece-wise constant graph-structured H_1 proposed in [6]. (2) Following the derivation of the SSS in [6], we show the relationship between the GFSS, SSS, and GLRT. (3) In our main theoretical result, we show that the performance of the GFSS depends explicitly on the spectral properties of the graph. (4) Because the GFSS requires the specification of the tuning parameter, ρ, we develop an adaptive version of the GFSS that automatically selects ρ. We extend our theory to this test. (5) We establish an information theoretic lower bound for the hypothesis testing problem and compare our results to this. (6) Using such results we are able to characterize in a very explicit form the performance of the GFSS on a few notable graph topologies and demonstrate its superiority over detectors that do not take into account the graph structure. (7) We demonstrate the usefulness of the GFSS with the partially simulated Arsenic concentration dataset.

II. Problem Setup

Detection involves the fundamental statistical question: are we observing merely noise or is there some signal amidst this noise? While the As contamination example in Section I-A involves binary measurements, for ease of presentation, we will work with Gaussian measurements with the understanding that many of the results derived may be extended easily to binary observations. We begin by outlining the basic problem of detecting a signal in Gaussian noise, then we will dive into graph-structured signals and the corresponding detection problem. First we begin with the Gaussian sequence space model, in which we make **one** observation at each node of the graph, yielding a vector $\mathbf{y} \in \mathbb{R}^p$ which is modeled as

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^p$ is the unknown signal and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_p)$ is Gaussian noise, with σ^2 known.

We will test whether the signal is constant over all the vertices or if there is some cluster of vertices, $C \subset V$, that have an elevated signal size. This cluster will be unknown, but it will be assumed that it is in a class of clusters that are well-connected within the graph. Throughout this paper we will let $\mathbf{1} = (1, \ldots, 1)$ denote the all 1s vector, let $\mathbf{1}_C$ be the indicator that a vertex is within $C \subset V$ (so that $(\mathbf{1}_C)_i = I\{i \in C\}$, $i \in V$), and $\bar{\mathbf{x}} = n^{-1}(\sum_{i=1}^p x_i)\mathbf{1}$. We now describe our graph-structured alternative class.

A. Graph-Structured H_1

In [6], an alternative class composed of piecewise-constant signals was proposed. Specifically, it was assumed that there is

a true cluster of vertices $C \subset V$ such that the signal $\mathbf{x}$ was constant over C and also constant over its complement $\bar{C} = V \setminus C$. We will make a more general assumption, that there is a true cluster C within which the average signal differs significantly from the average signal within its complement. Define the following signal class,

$$\mathcal{X}(\mu, C) = \left\{ \mathbf{x} \in \mathbb{R}^p : \left| \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|} - \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|} \right| \sqrt{\frac{|C||\bar{C}|}{p}} \geq \mu \right\}$$

for $C \subset V$. It can be shown that for any $\mathbf{x} \in \mathcal{X}(\mu, C)$, $\|\mathbf{x} - \bar{\mathbf{x}}\| \geq \mu$, so μ may be thought of as the minimum amplitude of our signal with its mean subtracted. We will restrict the mean vectors that we would like to detect to the classes $\mathcal{X}(\mu, C)$.

This signal class $\mathcal{X}(\mu, C)$ is defined as if the true cluster C is known, which in general will not be the case. Thus, knowledge of the true cluster C is not assumed, other than that it belongs to a given class $\mathcal{C} \subset 2^V$ that we define next. This class provides a good model for activations that are localized on the graph as we will see. Formally, we define, for some $\rho > 0$ (which is the same ρ in the definition of the GFSS),

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

where $\partial C = \{(i,j) \in E : i \in C, j \in \bar{C}\}$ is the boundary of C and $\mathbf{W}(\partial C) = \sum_{(i,j)\in \partial C} W_{i,j}$. Note that $\mathcal{C}$ is a symmetric class in the sense that $C \in \mathcal{C}$ if and only if $\bar{C} \in \mathcal{C}$. The quantity $\frac{p\mathbf{W}(\partial C)}{|C||\bar{C}|}$ is known in the graph theory literature as the **cut sparsity** [24] and is equivalent, up to factor of 2, to the **cut expansion** $\left(\frac{\mathbf{W}(\partial C)}{\min\{|C|,|\bar{C}|\}}\right)$:

$$\frac{\mathbf{W}(\partial C)}{\min\left\{|C|, |\bar{C}|\right\}} \leq \frac{p\mathbf{W}(\partial C)}{|C||\bar{C}|} \leq 2\frac{\mathbf{W}(\partial C)}{\min\left\{|C|, |\bar{C}|\right\}}$$

The cut expansion of a vertex set C is a measure of the size of the boundary relative to the size of C. Notice that for the same size of activation, a cluster of well connected nodes on the graph has a smaller cut sparsity and cut expansion than a cluster over isolated nodes. We can now define our composite alternative class,

$$\mathcal{X}(\mu, \rho) = \bigcup_{C \in \mathcal{C}(\rho)} \mathcal{X}(\mu, C).$$

We now can formalize our hypothesis testing problem as

$$H_0 : \mathbf{x} = \bar{\mathbf{x}} \text{ v.s. } H_1 : \mathbf{x} \in \mathcal{X}(\mu, \rho). \qquad (4)$$

Hence, we are testing if the signal is constant, or if there is a cluster with a low cut sparsity such that the average signal within the cluster differs from its complement.

Note that this alternative class is much less restrictive than existing work, e.g. [16] considers intervals, rectangles, ellipses, and similar geometrical shapes, or [21] considers cliques, stars, spanning trees. The only other work which considers general non-parametric shapes is [18], however it only considers lattice graphs and it is not clear how to extend the signal class definition used in that work to general graphs.

B. Distinguishability of H_0 and H_1

We will analyze asymptotic conditions under which the hypothesis testing problem described above is statistically feasible, in a sense made precise in the next definition. We will assume that the size of the graph p increases and the relevant parameters of the model, μ, σ, ρ, and eigen-spectrum of Δ, change with p 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 μ/σ, ρ, and the spectrum of the graph.

Definition 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. Recall that H_0 and H_1 index sets of probability measures by which $\mathbf{y}$ may be distributed. We say that the hypotheses H_0 and H_1 are **asymptotically distinguished by the test** T if

$$\sup_{\mathbb{P}_0 \in H_0} \mathbb{P}_0\{T = 1\} \to 0 \quad \text{and} \quad \sup_{\mathbb{P}_1 \in H_1} \mathbb{P}_1\{T = 0\} \to 0 \quad (5)$$

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

In Section III we will derive conditions under which the GFSS asymptotically distinguishes H_0 from H_1, and in Section IV we characterize SNRs under which H_0 and H_1 are asymptotically indistinguishable. We will say that a test is **adaptive** if it can be performed without knowledge of ρ. Naturally, requiring adaptivity may inhibit the quality of our test, as it has for detection within Sobolov-type ellipsoids [25], [26]. We will modify the GFSS to make it adaptive and prove theory regarding its performance.

III. Graph Fourier Scan Statistic

In [6], the Spectral Scan Statistic (SSS) was derived as a convex relaxation of the Generalized Likelihood Ratio (GLR) statistic. The SSS is the result of a convex optimization, while the GFSS is the result of a linear filter that was designed to mimic the SSS. We favor the GFSS because it is simple to implement, performs as well as the SSS in practice, and is the basis of the construction of the adaptive GFSS. These advantages are due to the fact that the GFSS is based on the linear filter, $G(\mathbf{y})$.

A. Derivation of GFSS

The hypothesis testing problem (4) presents two challenges: (1) the model contains an unbounded nuisance parameter $\bar{\mathbf{x}}$ and (2) the alternative hypothesis is comprised of a finite union of composite hypotheses indexed by $\mathcal{C}$. These features set our problem apart from existing work of structured normal means problems (see, e.g. [16]–[18], [21]), which does not consider nuisance parameters and relies on a simplified framework consisting of a simple null hypothesis and a composite hypothesis consisting of unions of simple alternatives.

To derive the GFSS we will first consider the simpler problem of testing the null hypothesis that $\mathbf{x} = \bar{\mathbf{x}}$, i.e. that the signal is constant, versus the alternative composite hypothesis that

$$H_0 : \mathbf{x} = \bar{\mathbf{x}} \quad \text{v.s.} \quad H_1^{\mathbf{x}_0} : \mathbf{x} = \bar{\mathbf{x}} + \mathbf{x}_0 \qquad (6)$$

for one fixed $\mathbf{x}_0 \in \mathbb{R}^p$ such that $\mathbf{x}_0^\top \mathbf{1} = 0$, $\mathbf{x}_0 \neq \mathbf{0}$. A standard approach to solve this testing problem is to compute the likelihood ratio (LR) statistic

$$\log \Lambda_{\mathbf{x}_0}(\mathbf{y}) = \frac{1}{\sigma^2}\left(\mathbf{x}_0^\top \tilde{\mathbf{y}} - \frac{1}{2}\|\mathbf{x}_0\|^2\right) \tag{7}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}} = (\tilde{y}_v, v \in V)$, and to reject the null hypothesis for large values of $\Lambda_{\mathbf{x}_0}(\mathbf{y})$ (the exact threshold for rejection will depend on the choice of the test significance level). The statistic given in (7) is in fact a generalized likelihood ratio statistic, which is due to the existence of the nuisance parameter $\bar{\mathbf{x}}$. In Appendix B, we provide a derivation that shows rigorously how we can eliminate the interference caused by the nuisance parameter by considering test procedures that are independent of $\bar{\mathbf{x}}$. The formal justification for this choice is based on the theory of optimal invariant hypothesis testing (see, e.g., [27]) and of uniformly best constant power tests (see [28]–[33]).

We will outline the derivation of the GLR statistic. The key insight is that every $\mathbf{x} \in \mathcal{X}(\mu, C)$ (for any $C \in \mathcal{C}$) can be decomposed into $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{x}_0$ where $\mathbf{x}_0^\top \mathbf{1} = 0$ and

$$\left|\frac{\mathbf{1}_C^\top \mathbf{x}}{|C|} - \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|}\right| = \left|\frac{\mathbf{1}_C^\top \mathbf{x}_0}{|C|} - \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}_0}{|\bar{C}|}\right|.$$

Hence, $\mathbf{x}_0 \in \mathcal{X}(\mu, C)$ if and only if $\mathbf{x} \in \mathcal{X}(\mu, C)$ and we can reduce the problem to testing (6) for all $\mathbf{x}_0$ within the set

$$\mathcal{X}^\perp(\mu, C) = \left\{\mathbf{x}_0 \in \mathcal{X}(\mu, C) : \mathbf{x}_0^\top \mathbf{1} = 0\right\}.$$

Hence, the testing problem (4) is equivalent to testing (6) for all $\mathbf{x}_0 \in \cup_{C \in \mathcal{C}} \mathcal{X}^\perp(\mu, C)$. When testing against more complex composite alternative, it is customary to consider instead the generalized likelihood ratio (GLR) statistic, which in our case reduces to

$$\hat{L} = \max_{C \in \mathcal{C}} \max_{\mathbf{x} \in \mathcal{X}^\perp(\mu, C)} 2\sigma^2 \log \Lambda_{\mathbf{x}}(\mathbf{y}).$$

Define the following vector,

$$\mathbf{z}_C = \sqrt{\frac{|C||\bar{C}|}{p}}\left(\frac{\mathbf{1}_C}{|C|} - \frac{\mathbf{1}_{\bar{C}}}{|\bar{C}|}\right).$$

Then we can rewrite the log-likelihood ratio statistic specific to a cluster $C \subset V$,

$$\max_{\mathbf{x} \in \mathcal{X}^\perp(\mu, C)} 2\sigma^2 \log \Lambda_{\mathbf{x}}(\mathbf{y}) = \begin{cases} \|\tilde{\mathbf{y}}\|^2, & \text{if } |\mathbf{z}_C^\top \tilde{\mathbf{y}}| \geq \mu \\ \|\tilde{\mathbf{y}}\|^2 - \left(|\mathbf{z}_C^\top \tilde{\mathbf{y}}| - \mu\right)^2, & \text{if } |\mathbf{z}_C^\top \tilde{\mathbf{y}}| < \mu, \end{cases} \tag{8}$$

where recall $\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}}$. We prove (8) in Appendix B. Define the statistic,

$$\hat{g} = \max_{C \in \mathcal{C}(\rho)} \left(\mathbf{z}_C^\top (\mathbf{y} - \bar{\mathbf{y}})\right)^2$$

then it follows that

$$\hat{L} = \begin{cases} \|\mathbf{y} - \bar{\mathbf{y}}\|^2, & \text{if } \hat{g} \geq \mu^2 \\ \|\mathbf{y} - \bar{\mathbf{y}}\|^2 - (\sqrt{\hat{g}} - \mu)^2, & \text{if } \hat{g} < \mu^2. \end{cases}$$

Hence, the GLR for the problem H_0 v.s. H_1 is a function of only $\|\mathbf{y} - \bar{\mathbf{y}}\|$ and $\hat{g}$. Because the signal size μ is unknown, we will focus on the use of $\hat{g}$ as a test statistic with the understanding that an omnibus test can be constructed from both $\|\mathbf{y} - \bar{\mathbf{y}}\|$ and $\hat{g}$. The statistic $\|\mathbf{y} - \bar{\mathbf{y}}\|$ is well studied as a Chi-squared statistic [34], so it will not be addressed in this study.

With simple algebraic manipulations, we find that the statistic, $\hat{g}$, has a very convenient form which is tied to the spectral properties of the graph via its Laplacian.

Lemma 2: Let $\mathbf{K} = \mathbf{I} - \frac{1}{p}\mathbf{1}\mathbf{1}^\top$ and notice that $\tilde{\mathbf{y}} = \mathbf{K}\mathbf{y}$. Then

$$\hat{g} = \max_{\mathbf{x} \in \{0,1\}^p} \frac{\mathbf{x}^\top \tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{K}\mathbf{x}} \text{ s.t. } \frac{\mathbf{x}^\top \Delta \mathbf{x}}{\mathbf{x}^\top \mathbf{K}\mathbf{x}} \leq \rho, \tag{9}$$

where Δ is the combinatorial Laplacian of the graph G.

The statistic $\hat{g}$ is the GLR for the piecewise constant alternative hypothesis proposed in [6], and the proof of Lemma 2 can be found there. We will refer to $\hat{g}$ as the GLR, even though the true GLR, $\hat{L}$, is a function of $\hat{g}$. A brute force approach to solving (9) is out of the question for graphs of any reasonable size, since even for the moderate p, 2^p is astronomically large. An interesting feature of $\hat{g}$ is that the program (9) is directly related to the sparsest cut problem in combinatorial optimization. In particular, the sparsest cut problem is the following combinatorial optimization,

$$\min_{\mathbf{x} \in \{0,1\}^p} \frac{\mathbf{x}^\top \Delta \mathbf{x}}{\mathbf{x}^\top \mathbf{K}\mathbf{x}}.$$

The sparsest cut program is known to be in general NP-hard, with poly-time algorithms known for trees and planar graphs [35]. In fact, just determining if there is an $\mathbf{x}$ in the feasibility set of (9) is NP-hard, indicating that the optimization (9) is in an NP-hard class. Having established the computational difficulty of the GLRT, we consider relaxations of the combinatorial optimization.

In order to obtain a tractable relaxation of the GLR statistic (9), [6] introduced the Spectral Scan Statistic (SSS), defined as

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

Indeed, [6] proved that the SSS is an upper bound to the GLRT statistic:

Proposition 3: The GLR statistic is bounded by the SSS: $\hat{g} \leq \hat{s}$, almost everywhere.

This is due to the fact that the objective and constraint in (9) takes the same form as a generalized eigenvalue problem, and the SSS is a relaxation of the hypercube $[0,1]^p$ to a hypersphere. Notice that because the domain $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \Delta \mathbf{x} \leq \rho, \|\mathbf{x}\| \leq 1, \mathbf{x}^\top \mathbf{1} = 0\}$ is symmetric around the origin, this is precisely the square of the solution to

$$\sqrt{\hat{s}} = \sup_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^\top \mathbf{y} \text{ s.t. } \mathbf{x}^\top \Delta \mathbf{x} \leq \rho,\ \|\mathbf{x}\| \leq 1, \mathbf{x}^\top \mathbf{1} = 0, \tag{10}$$

where we have used the fact that $\mathbf{x}^\top \tilde{\mathbf{y}} = \left(\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{x}\right)^\top \mathbf{y} = \mathbf{x}^\top \mathbf{y}$ because $\mathbf{x}^\top \mathbf{1} = 0$ within $\mathcal{X}$. Hence, the domain of the SSS is the intersection of the ellipsoid $\{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^\top \Delta \mathbf{x} \leq \rho\}$ and the hypersphere $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| \leq 1\}$. While the SSS can be computed with standard first order methods, its properties are not well understood because the domain is this complicated intersection. With that in mind we proposed the GFSS, which is a further relaxation of the SSS.

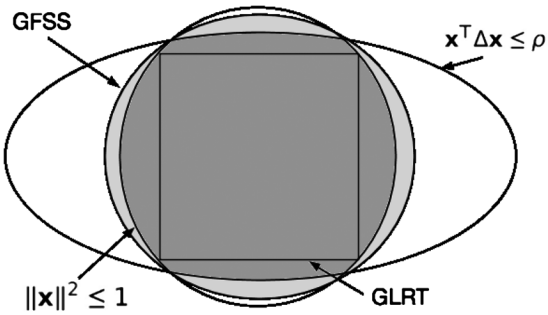

Fig. 2. (**Relaxation Diagram**) A diagram depicting the various statistics based on the domains over which they are optimizators. The domain of the GLR statistic $\hat{g}$ is the hypercube intersected with the ellipsoid $\mathbf{x}^\top \Delta \mathbf{x} \leq \rho$, while the domain of the SSS is the darkly shaded region which is the ellipsoid intersected with the hypersphere ($\|\mathbf{x}\| \leq 1$). The GFSS is the relaxation to an ellipsoid that approximates the domain of the SSS and is the lighter, shaded region unioned with the darker region.

Proposition 4: Recall the definition of the GFSS, $\hat{t}$ in (1). The SSS as a function of ρ can be bounded above and below in the following:

$$\hat{t} + \sum_{i=2}^{p} \min\left\{1, \frac{\rho}{\lambda_i}\right\} \leq \hat{s} \leq 2\left(\hat{t} + \sum_{i=2}^{p} \min\left\{1, \frac{\rho}{\lambda_i}\right\}\right).$$

The proof is provided in Appendix A. The GFSS was based on a linear filter ($G(\mathbf{y})$) that approximates the SSS, $\hat{s}$. The fact that it is a linear filter will allow us to develop the adaptive version of the GFSS, because of our more precise understanding of the distribution of linear filters. The implication of Propositions 3 and 4 is that the GFSS, $\hat{t}$, is a relaxation of the GLRT, $\hat{g}$. The GFSS is a computationally tractable alternative to the GLRT, and as we will see, it is often a vast improvement over the naive test statistics.

The GFSS is in fact the relaxation of the SSS to an ellipsoid that approximates the domain of the SSS. This results in the GFSS being expressed using the linear filter, $G(\mathbf{y})$. In Fig. 2, we have depicted the various regions for which the statistics are relaxations. In the figure, the GFSS domain is nearly identical to the hypersphere, but in high-dimensions the difference between these can be very dramatic and the GFSS domain is very close to the domain of the SSS.

B. Theoretical Analysis of GFSS

A thorough theoretical analysis of the GFSS has several uses. In Corollary 6, we characterize the critical signal-to-noise ratio, enabling us to determine the strength of the GFSS as a detector on theoretical grounds. Theorem 5 will be used to form an adaptive version of the GFSS, which will in turn alleviate the need for specifying ρ.

The following main result bounds the test statistic under H_0 and the graph structured H_1. It is based on the concentration of weighted sums of independent χ^2 random variables found in [36].

Theorem 5: Under the null hypothesis H_0, with probability at least $1-\alpha$ where $\alpha \in (0,1)$,

$$\hat{t} \leq 2\left(\sqrt{\sum_{i=2}^{p} \min\left\{1, \frac{\rho^2}{\lambda_i^2}\right\} \log\left(\frac{1}{\alpha}\right)} + \log\left(\frac{1}{\alpha}\right)\right). \quad (11)$$

Under the alternative hypothesis, H_1, with probability at least $1-\gamma$ where $\gamma \in (0,1)$,

$$\hat{t} \geq \frac{\mu^2}{2\sigma^2} - \frac{2\mu}{\sigma}\sqrt{\log\frac{2}{\gamma}} - 2\sqrt{\sum_{i=2}^{p} \min\left\{1, \frac{\rho^2}{\lambda_i^2}\right\} \log\frac{2}{\gamma}}, \quad (12)$$

for μ/σ large enough.

The proof is provided in Appendix A. Theorem 5 shows that by setting a threshold to be the right hand side of (11), we have a level α test. If we then set the right hand side of (12) to be this threshold, and solve for μ/σ, then we get the lowest SNR such that the test has power $1-\gamma$ under the alternative. The result below allows us to compare the GFSS to other tests on asymptotic theoretical grounds.

Corollary 6: The GFSS $\hat{t}$ can asymptotically distinguish H_0 from H_1 if the SNR is stronger than

$$\frac{\mu}{\sigma} = \omega\left(\sum_{i=2}^{p} \min\left\{1, \frac{\rho^2}{\lambda_i^2}\right\}\right)^{\frac{1}{4}}.$$

Most notably the critical SNR is lower than $p^{1/4}$ which is the critical SNR enjoyed by the energy test statistic. Comparing this to the analogous results for the SSS in [6] (Cor. 8), we see that this is a significant improvement. The most unreasonable assumption that we have made thus far is that the cut sparsity, ρ, is known. The following section develops a test that adapts to ρ.

Before we delve into that, a remark on the computational complexity of the proposed test is in order. The worst-case runtime of computing all the eigenvalues of a $p \times p$ matrix is cubic in p. Thus, we have clearly demonstrated a test that is computationally feasible for any graph topology i.e. it runs in polynomial (cubic) time in the size of the graph p, compared to the GLRT whose computation can be exponential in p in the worst case. However, cubic computational complexity might be prohibitive for very large graphs. It turns out that the GFSS can be calculated with just the top j eigenvectors and a Laplacian solver. Specifically, if $j = \max\{i : \lambda_i < \rho\}$ and let $\mathbf{P}_j$ be the projection onto the span of $\{\mathbf{u}_i\}_{i=2}^{j}$ then the GFSS can be written as

$$\mathbf{y}^\top \mathbf{P}_j \mathbf{y} + \rho \mathbf{y}^\top (\mathbf{I} - \mathbf{P}_j)\Delta^\dagger (\mathbf{I} - \mathbf{P}_j)\mathbf{y} - \sum_{i=2}^{p} \min\left\{1, \frac{\rho}{\lambda_i}\right\}.$$

The first term requires the computation of the top j eigenspace, while the second term requires a Laplacian solver. One can observe that the final term is the expected value of the first two terms applied to a vector drawn from the p-dimensional standard normal, indicating that it can be approximated by Monte carlo sampling. The computation of the first two terms take time $O(kp^2 + m\text{polylog}(m))$, where the constants may depend on the spectral decay, by using the fast Laplacian solvers of [37] (recall that m is the number of edges). Furthermore, when the GFSS will be used on multiple measurement vectors (such as multiple measurements in time), then $\mathbf{P}_j$ needs to only be computed once.

IV. The Adaptive GFSS

Notice that the SSS and GFSS require that we prespecify the cut sparsity parameter, ρ. While the user may have certain shapes in mind, such as large rectangles in a lattice, it is not reasonable to assume that this can be done for arbitrary graph structure. In Theorem 5, we have seen that the choice of the ρ parameter has a significant effect on the theory of the GFSS. In Section VI-A, we will find that the adaptive procedure outperforms the GFSS when a reasonable guess for ρ is used. In order to adapt to ρ we will consider the test statistic, $\hat{t}(\rho)$, as a function of ρ, as it is allowed to vary.

Definition 7: Let $\alpha > 0$ and

$$\tau(\rho)=2\left(\sqrt{\sum_{i=2}^{p}\min\left\{1,\frac{\rho^2}{\lambda_i^2}\right\}\log\left(\frac{p-1}{\alpha}\right)}+\log\left(\frac{p-1}{\alpha}\right)\right).$$

The adaptive GFSS test is the test that rejects H_0 if $\exists\rho>0$ such that

$$\hat{t}(\rho) > \tau(\rho). \tag{13}$$

As we will now show, in order to compute the entire curve $\hat{t}(\rho)$, it is sufficient to evaluate $\hat{t}(\rho)$ only at $p-1$ points. This is because $\hat{t}(\rho)$ is piecewise linear with knots at the eigenvalues. Also, $\tau(\rho)$ is similarly well behaved. Let $j=\max\{i:\lambda_i\le\rho\}$ then

$$\hat{t}(\rho)=\rho\sum_{i=j+1}^{p}\frac{\left(\mathbf{u}_i^\top\mathbf{y}\right)^2-1}{\lambda_i}+\sum_{i=2}^{j}\left(\left(\mathbf{u}_i^\top\mathbf{y}\right)^2-1\right).$$

Hence, $\hat{t}(\rho)$ is piecewise linear with knots at $\{\lambda_i\}_{i=2}^p$. The threshold function can be expressed by

$$\tau(\rho)=\sqrt{4\left(\rho^2\sum_{i=j+1}^{p}\lambda_i^{-2}+j\right)\log\frac{p-1}{\alpha}}+2\log\left(\frac{p-1}{\alpha}\right).$$

Define the following quantities,

$$A=4\log\left(\frac{(p-1)}{\alpha}\right)\sum_{i=j+1}^{p}\lambda_i^{-2},\quad B=4j\log\left(\frac{(p-1)}{\alpha}\right)$$

$$D=2\log\left(\frac{(p-1)}{\alpha}\right),\quad E=\sum_{i=j+1}^{p}\frac{\left(\mathbf{u}_i^\top\mathbf{y}\right)^2-1}{\lambda_i},$$

$$F=\sum_{i=2}^{j}\left(\left(\mathbf{u}_i^\top\mathbf{y}\right)^2-1\right).$$

Then we reject iff

$$\tau(\rho)=\sqrt{\rho^2A+B}+D<\rho E+F=\hat{t}(\rho).$$

Notice that $A,B,D>0$, so $\tau(\rho)$ has strictly positive curvature and is convex. Thus, $\tau(\rho)-\hat{t}(\rho)$ is convex within $\lambda_j\le\rho\le\lambda_{j+1}$ and has a unique minimum. We can minimize the unrestricted function,

$$\rho^*=\arg\min_{\rho}\sqrt{\rho^2A+B}+D-\rho E-F$$

and we find that this is attained at

$$\rho^*=\begin{cases}0, & E^2\ge A\\ \sqrt{\frac{E^2B}{A^2-E^2A}}, & \text{otherwise.}\end{cases}$$

We know by convexity that if $\rho^*<\lambda_j$ then the constrained maximum is attained at λ_j, and if $\rho^*>\lambda_{j+1}$ then it is attained at λ_{j+1}. For each j, we can construct A,B,D,E,F and define

$$\rho_j=\begin{cases}\lambda_j, & E^2\ge A \text{ or } \sqrt{\frac{E^2B}{A^2-E^2A}}\le\lambda_j\\ \lambda_{j+1}, & \sqrt{\frac{E^2B}{A^2-E^2A}}\ge\lambda_{j+1}\\ \sqrt{\frac{E^2B}{A^2-E^2A}}, & \text{otherwise.}\end{cases}$$

Then the following proposition holds,

Proposition 8: The adaptive GFSS test rejects H_0 if and only if

$$\exists j\in\{2,\dots,p\},\quad \tau(\rho_j)<\hat{t}(\rho_j).$$

This proposition states that we only need to compute the GFSS at p points, which implies that it takes $O(p^3)$ time to compute the adaptive GFSS (since computing the eigendecomposition is the most computationally expensive step). The specific choice of threshold function $\tau(\rho)$, (13), gives us a control on the false alarm (type 1 error).

Theorem 9: The probability of false rejection (type 1 error) is bounded by

$$\sup_{\mathbb{P}_0\in H_0}\mathbb{P}_0\left\{\exists\rho,\hat{t}(\rho)>\tau(\rho)\right\}\le\alpha.$$

Consider models from the alternative hypothesis, H_1 as functions of ρ. Let ρ^* be the smallest such ρ^* such that $\mathbf{x}+\boldsymbol{\epsilon}$ is contained in the alternative hypotheses. Then the probability of type 2 error is bounded by $\gamma>0$ if

$$\tau(\rho^*)<\frac{\mu^2}{2\sigma^2}-2\frac{\mu}{\sigma}\sqrt{2\log\left(\frac{2}{\gamma}\right)}-2\sqrt{\sum_{i=2}^{p}\min\left\{1,\frac{\rho^{*2}}{\lambda_i^2}\right\}\log\left(\frac{2}{\gamma}\right)}.$$

The interpretation is that by providing the thresholding function $\tau(\rho)$ we are in effect thresholding at p distinct points which can be controlled theoretically by union bounding techniques. The following corollary describes the SNR rates necessary for asymptotic distiguishability.

Corollary 10: The adaptive GFSS asymptotically distinguishes H_0 from H_1 if

$$\frac{\mu}{\sigma}=\omega\left(\sqrt{\sum_{i=2}^{p}\min\left\{1,\frac{\rho^{*2}}{\lambda_i^2}\right\}\log p+\log p}\right)^{\frac{1}{2}}.$$

So we are able to make all the same theoretical guarantees with the adaptive GFSS as the GFSS with an additional multiplicative term $(\log p)^{1/4}$ and an additive term of $(\log p)^{1/2}$.

We see that in Section VI-A in our simulations that the adaptive GFSS can significantly outperform the GFSS when ρ is unknown. We show in Section VI-B how this theory is applicable by developing corollaries for different specific graph topologies.

V. Information Theoretic Lower Bounds

In order to understand the fundamental limitations of the activity detection problem, we review a lower bound on the performance of any testing procedure. The following lower bound on the critical SNR was derived in [6]. The first part is a simple bound on the performance of the oracle (who has knowledge of the active cluster, C) based on the Neyman-Pearson lemma. The second part is more sophisticated and requires that the graph has symmetries that we can exploit, but it will not be satisfied by many graphs. We will later show that these conditions are satisfied by the specific graph structures that we will analyze in Section VI-B.

Theorem 11. [6]:

(a) H_0 and H_1 are asymptotically indistinguishable if $\mu/\sigma = o(1)$.

(b) 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

$$\forall C \in \mathcal{C}', \quad \frac{p\mathbf{W}(\partial C)}{|C||\bar{C}|} \leq \frac{\rho}{2}$$

i.e., elements of $\mathcal{C}'$ belong to the alternative hypothesis with $\rho/2$ cut sparsity. Furthermore assume that $\frac{c|\mathcal{C}'|}{p} \to 1$. H_0 and H_1 are asymptotically indistinguishable if

$$\frac{\mu}{\sigma} = o\left(|\mathcal{C}'|^{\frac{1}{4}}\right)$$

In [18], the authors also derive a lower bound which scales as $\sqrt{\log(p/|C|)}$ for detection of patterns on the lattice graph which include squares of size $|C|$. However, their results only hold for clusters consisting of a single connected component, whereas H_1 allows for multiple connected components. Thus, our results indicate that detecting clusters with multiple connected components is harder than detecting a cluster with a single connected component, unless the connected component is really large i.e. the activation size $|C|$ is of the same order as the graph size p. In the latter case, both the bound of [18] and our result imply an SNR of $o(1)$ is insufficient for detection on the torus graph.

VI. Specific Graph Models and Experiments

In this section, we demonstrate the power and flexibility of Theorem 5 by analyzing in detail the performance of the GFSS over a simulated As detection example and three important graph topologies: balanced binary trees, the torus graph and Kronecker graphs (see [38], [39]). The explicit goals of this section are as follows:

1) Demonstrate the effectiveness of the GFSS on partially simulated dataset from the Arsenic detection graph.
2) Determine the implications of Theorem 5 in these specific graph examples for some example signal classes;
3) Demonstrate the competitiveness of the GFSS and the adaptive GFSS against the aggregate and max statistics;
4) Provide an example of the general graph structure;

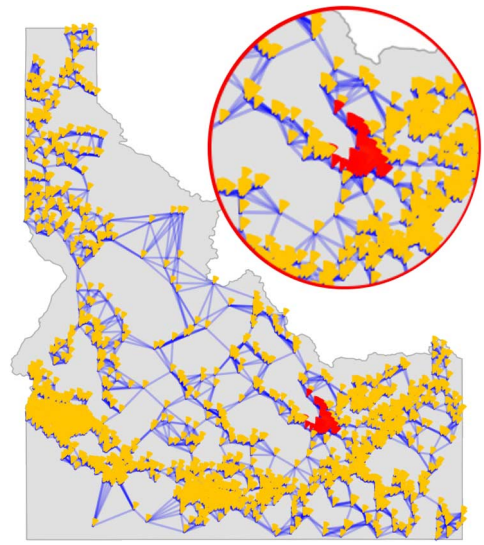

Fig. 3. **(kNN Well Graph)** The test well locations over Idaho are displayed above. The k-nearest neighbor graph with $k = 8$ is used with the activated wells from the second simulation in red (1 large active cluster). The active region has been highlighted in the upper right corner.

A. Arsenic Detection Simulation

In order to compare the GFSS, adaptive GFSS, and some naive estimators, we construct realistic signals over the Arsenic graph (so that we have a ground truth) and generate Gaussian noise over these signals. This will also provide us with an opportunity to make some practical recommendations on how to use the GFSS. We form the k-nearest neighbor graph (with $k = 8$) from the 2,191 well locations (see Fig. 3). In order to construct realistic signals, we will use the locations of the principle aquifers in Idaho [40]. We will assign each test well to its closest aquifer and select randomly a small number of aquifers that we will consider to be contaminated. The signal, x_i, that we will construct is elevated over the wells belong to a contaminated aquifer and zero over the other wells. Specifically, we set the level of elevation in each simulation such that $||\mathbf{x} - \bar{\mathbf{x}}||^2 = 5$ and we generate additive Gaussian noise with $\sigma = 1$.

In the first experiment (Fig. 4 left), we chose 3 aquifers at random which resulted in 112 contaminated wells. By our choice of $||\mathbf{x} - \bar{\mathbf{x}}||^2 = 5$, the signal size at each contaminated well was 0.47 which is substantially less than the noise level σ. In the second experiment (Fig. 4 middle), we chose 1 of the larger aquifers (with greater than 100 wells) at random which resulted in 109 contaminated wells. In the third experiment (Fig. 4 right), we selected 1 of the somewhat smaller aquifer (with greater than 50 wells) at random which resulted in 62 contaminated wells. We simulate the probability of correct detection (rejecting H_0 when the truth is H_1) versus the probability of false alarm (falsely rejecting H_0) by making 1000 draws from the noise distribution (with and without the signal $\mathbf{x}$ for H_1 and H_0 respectively).

For comparison purposes and to demonstrate the effectiveness of utilizing graph structure, we will use an Aggregate and Maximum test statistic, defined as $\sum_{j \in V} y_j$ and $\max_{j \in V} |y_j|$. In all of our simulations we use a signal that is elevated and constant over a cluster and zero elsewhere. In this setting, if the

cluster is very large then the aggregate statistic should be competitive and for a small one the maximum should be more competitive. The purpose of comparing the GFSS to these naive statistics is to demonstrate the gains obtained from exploiting the graph structure.

As can be seen the adaptive GFSS test strictly outperforms all of the test statistics, which demonstrates the importance of adapting to the ρ parameter. This is especially apparent in the second and third simulations, in which the power is significantly better than the GFSS. Because we did not assume that we knew the ρ parameter, the optimal choice of ρ in the non-adaptive GFSS was surely not used. We can see that the effect of suboptimal ρ parameter can be very detrimental to the performance of our procedure. This justifies the development of the adaptive GFSS and we generally recommend its use over the GFSS with a guessed ρ parameter.

Despite not adapting to ρ, the GFSS with the somewhat arbitrary choice of $\rho = \lambda_{109}$ begins to outperform the Aggregate statistic for the smaller contamination as our theory predicts. This is because when the active cluster is smaller the required SNR for the Aggregate statistic increases significantly. The max statistic is nearly powerless in all of the examples because the SNR required for it to be competetive is very large relative to those required by the GFSS. The adaptive GFSS is a substantially better alternative to an arbitrary choice of ρ and statistics that do not take the kNN graph structure into account.

While Theorem 5 can be inverted to obtain a P-value that is valid for finite p, this may be too conservative for practical purposes. We recommend one of two approaches: forming a Z-score that is asymptotically normal under H_0, and using a permutation test. Under H_0, the GFSS, $\hat{t}$, has zero mean and because it is the sum of weighted χ^2_1 random variables it has a variance of $2(\sum_{i=2}^{p} h(\lambda_i)^4)$. Thus, a Z-score can be calculated by $\hat{Z} = \hat{t}/\sqrt{2\sum_{i=2}^{p} h(\lambda_i)^4}$, which can be shown to have an asymptotic standard normal distribution under some regularity conditions. Thus, we can form an asymptotically valid P-value by applying the standard normal inverse CDF to $\hat{Z}$. While this is valid when the noise is Gaussian, in many instances the measurements are not Gaussian and we interpret H_0 to mean that $x_i = \mathbb{E}y_i$ is constant over the graph which is a weaker assumption (recall we had binary observations in Section I-A, but we used the GFSS none-the-less). In this case, we can apply a permutation test, by which we randomly permute the coordinates of $\mathbf{y}$ and maintain the graph structure. We interpret the resulting statistic $\hat{t}$ as a simulation of the GFSS under H_0. Then an estimated P-value would be the fraction of permutations that have a larger $\hat{t}$ then the actual GFSS. This was used to construct the reported P-values in Section I-B.

B. Balanced Binary Trees

Balanced trees are graph structures of particular interest because they provide a simple hierarchical structure. Furthermore, the behavior of the graph spectra for the balanced binary tree provides a natural multiscale basis [13], [41]. We begin this analysis of the GFSS by applying it to the balanced binary tree (BBT) of depth ℓ. We consider the class of signals defined by $\rho = [cp^{\alpha}(1 - cp^{\alpha-1})]^{-1}$ where $0 < c \leq 1/2$, $0 < \alpha \leq 1$. This class is interesting as it includes, among others, clusters of constant signal which are subtrees of size at least cp^{α} (subtrees can be isolated from a tree by cutting a single edge and hence have cut size 1).

Corollary 12: Let G be a balanced binary tree with p vertices, and let $\rho = p[cp^{\alpha}(p - cp^{\alpha})]^{-1}$.

(a) The GFSS can asymptotically distinguish H_0 from signals within H_1 if the SNR is stronger than

$$\frac{\mu}{\sigma} = \omega\left(p^{\frac{1-\alpha}{4}}(\log p)^{\frac{1}{4}}\right).$$

(b) The adaptive GFSS distinguishes the hypotheses of (a) if

$$\frac{\mu}{\sigma} = \omega\left(p^{\frac{1-\alpha}{4}}(\log p)^{\frac{1}{2}}\right).$$

(c) H_0 and H_1 are asymptotically indistinguishable if

$$\frac{\mu}{\sigma} = o\left(p^{\frac{1-\alpha}{4}}\right).$$

The conclusion is that for the BBT the GFSS and the adaptive GFSS is near optimal with respect to critical SNR. The proof (Appendix A) is based on the special form of the spectrum of the BBT. So in this case, the GFSS consistently dominates the naive statistics and the theoretical results are very close to the lower bounds for any α.

We simulate the probability of correct detection versus the probability of false alarm. These are given for the four statistics in Fig. 5 as the test threshold, and hence the probability of false alarm, is varied. The GFSS is computed with the correct ρ, which is in general unknown. Different statistics dominate under different choices of cluster size parameter, α. When $\alpha = 1$, corresponding to large clusters, where the size is on the same order as p, the aggregate statistic is competitive with the adaptive statistic. When $\alpha = 0.5$, corresponding to clusters of size $\asymp p^{1/2}$, the aggregate becomes less competitive and the max more competitive than the $\alpha = 1$ case, and the GFSS remains the dominating test. In each case, we set $c = 1/2$, which ensures that the $\alpha = 1$ case does not select the entire tree.

C. Torus Graph

A torus graph is a lattice or two-dimensional grid that is wrapped around so that rightmost vertices are same as leftmost vertices, and topmost vertices are same as bottom vertices. Formally, the $\ell \times \ell$ torus graph ($p = \ell^2$) is defined as follows. Let the vertex set $V = (\mathbb{Z} \bmod \ell)^2$ where points $(i_1, i_2), (j_1, j_2)$ are connected by an edge if and only if $|(i_1 - j_1) \bmod \ell| + |(i_2 - j_2) \bmod \ell| = 1$ (here $|i \bmod \ell|$ means the smallest absolute value of representatives). The class of clusters in the torus under consideration $\mathcal{C}(\rho)$ are those that have sparsity $p\mathbf{W}(\partial C)/(|C||\bar{C}|) \leq \rho$. For example, rectangles of size $k \times k$ within the torus have cut sparsity $4kp/(k^2(p - k^2)) \asymp 4/k$. This means that if we would like to include rectangles of size roughly $k \times k$, it is sufficient that $\rho \asymp 4/k$.

The torus is an important example as it models a mesh of sensors in two dimensions. We will analyze the performance guarantees of the GFSS over our running example, the 2-dimensional torus graph. To include squares of size $p^{1-\beta}$, we set $\rho \asymp p^{-(1-\beta)/2}$. The following result is due to a detailed analysis of the spectrum of the torus.

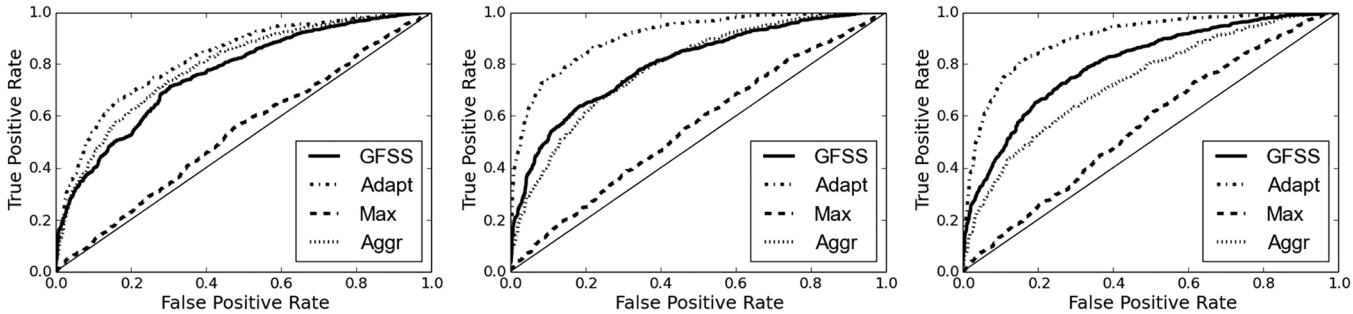

Fig. 4. (**Arsenic Contamination Simulations**) Simulations of the size (false positive rate) and the power under H_1 for the As simulations of the GFSS, adaptive GFSS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for 3 contaminated aquifers (left), 1 large contaminated aquifer (middle), and 1 smaller contaminated aquifer (right).

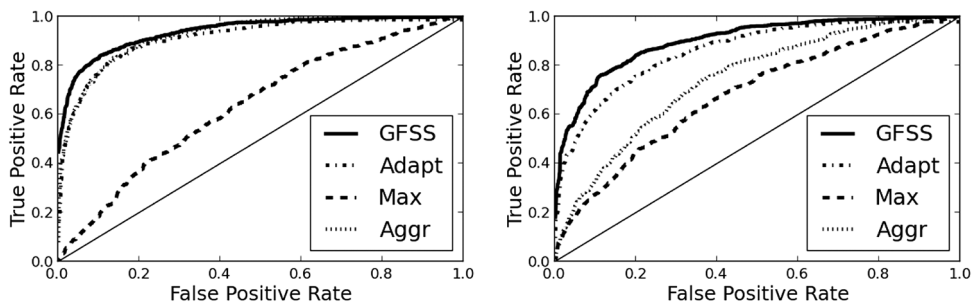

Fig. 5. (**BBT Comparisons**) Simulations of the size (false positive rate) and the power under H_1 for the balanced binary tree of the GFSS, adaptive GFSS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for the tree of depth $\ell = 6$, $p = 2^{\ell+1} - 1 = 127$, with choice of $\alpha = 1$ (left) and $\alpha = 0.5$ (right).

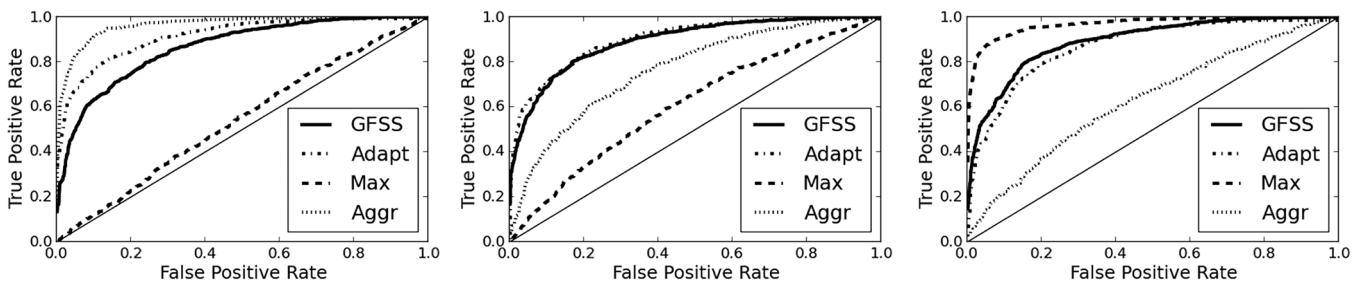

Fig. 6. (**Torus Comparisons**) Simulations of the size (false positive rate) and the power under H_1 for the Torus of the GFSS, adaptive GFSS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for side length of $\ell = 30$, $p = \ell^2 = 900$, with choice of $\beta = 0$ (top left), $\beta = .5$ (top right) and $\beta = .75$ (bottom).

Corollary 13: Let G be the $\ell \times \ell$ square torus ($p = \ell^2$), and let $\rho = cp^{-(1-\beta)/2}$ for $\beta \in [0, 1)$.

(a) (a) The GFSS can asymptotically distinguish H_0 from H_1 if the SNR satisfies

$$\frac{\mu}{\sigma} = \omega\left(p^{\frac{3}{20}+\frac{1}{10}\beta}\right).$$

(b) (b) The adaptive GFSS can asymptotically distinguish the hypotheses of (a) if

$$\frac{\mu}{\sigma} = \omega\left(p^{\frac{3}{20}+\frac{1}{10}\beta}(\log p)^{\frac{1}{4}}\right).$$

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

$$\frac{\mu}{\sigma} = o\left(p^{\frac{\beta}{4}}\right).$$

The implication of Cor. 13 is that when $\beta > 0$ (the clusters are not too large), the GFSS is consistent under an SNR lower than $p^{1/4}$. Regardless of the β parameter the GFSS never achieves the lower bound for the torus graph, which suggests an approach that exploits the specific structure of the torus may yet outperform the GFSS. We simulate the performance of the test statistics over a 30×30 torus, with $\beta = 0, .5, .75$ with $c = 1/2$ When β is small (large clusters), we suffer an additional factor of $p^{3(1-\beta)/20}$ in the upper bound. Despite the theoretical shortcomings of in this case, the simulations (Fig. 6) suggest that the GFSS is significantly superior to the naive tests for medium sized clusters.

D. Kronecker Graphs

Much of the research in complex networks has focused on observing statistical phenomena that is common across many data sources. The most notable of these are that the degree distribution obeys a power law ([42]) and networks are often found to have small diameter ([43]). A class of graphs that satisfy

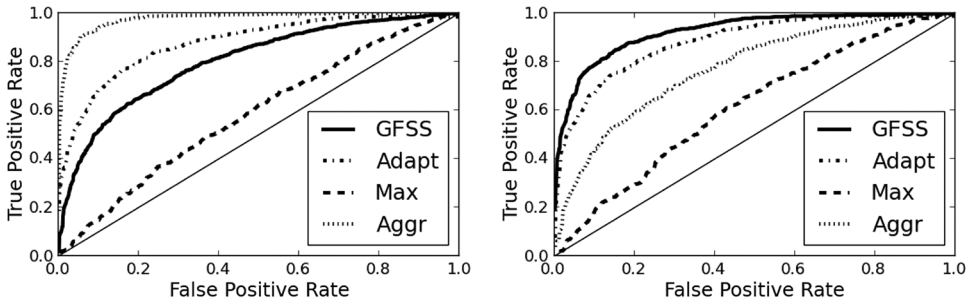

Fig. 7. (**Kronecker Comparison**) Simulations of the size (false positive rate) and the power under H_1 for the Kronecker graph of the GFSS, adaptive GFSS (Adapt), Max statistic (Max), and Aggregate statistic (Aggr). The figures are for a base graph of size $p_0 = 6$ and Kronecker power of $\ell = 3$, so $p = p_0^\ell = 216$. The cuts were chosen at the coarsest scale, $k = 1$, (left) and at the second coarsest, $k = 2$ (right).

these, while providing a simple modeling platform are the Kronecker graphs (see [38], [39]). Let H_1 and H_2 be graphs on p_0 vertices with Laplacians Δ_1, Δ_2 and edge sets E_1, E_2 respectively. The Kronecker product, $H_1 \otimes H_2$, is the graph over vertices $[p_0] \times [p_0]$ 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_0 vertices. Then the graph G for $\ell > 0$ levels is defined as

$$\frac{1}{p_0^{\ell-1}} H \otimes \frac{1}{p_0^{\ell-2}} H \otimes \ldots \otimes \frac{1}{p_0} H \otimes H.$$

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 14: Let G be the Kronecker product of the base graph H described above with $p = p_0^\ell$ vertices, and let $\rho \asymp p_0^{2k-\ell-1}$ (which includes cuts within the k coarsest scale).

(a) The GFSS can asymptotically distinguish H_0 from signals from H_1 if the SNR is stronger than

$$\frac{\mu}{\sigma} = \omega\left(p^{\frac{k}{2\ell}} \left(\operatorname{diam}(H)\right)^{\frac{1}{4}}\right)$$

where $\operatorname{diam}(H)$ is the diameter of the base graph H.

(b) The adaptive GFSS can distinguish the hypotheses of (a) if

$$\frac{\mu}{\sigma} = \omega\left(p^{\frac{k}{2\ell}} \left(\operatorname{diam}(H) \log p\right)^{\frac{1}{4}}\right).$$

(c) H_0 and H_1 are asymptotically indistinguishable if

$$\frac{\mu}{\sigma} = o\left(p^{\frac{k}{4\ell}}\right).$$

The proof and an explanation of ρ is in the appendix. The implication of Cor. 14 is that only for k small is the GFSS nearly optimal. Generally, one will suffer a multiplicative term of $p^{k/4\ell}$. As we can see from the simulations the $k = 1$ case is exactly when the aggregate statistic dominates (see Fig. 7). When $1 < k < \ell$, the GFSS improves on the aggregate and the max statistics. Throughout these simulations we set $\rho = p_0^{2k-\ell-1}$.

One may rightly ask if the gap between the upper bounds (Corollaries 12 (b), 13 (b), 14 (b)) and the lower bounds (Corollaries 12 (c), 13 (c), 14 (c)) is just due to a lack of theoretical know-how and the test is actually optimal. We attempt to assess this concern by plotting the performance of the GFSS with the SNR increasing according to the scaling dictated by the upper bounds (Fig. 8). For the BBT because the curve does not change significantly with p (as the tree depth l increases), the upper bound is supposed to be tight. In the torus graph, for large rectangles ($\beta = 0$) the upper bound appears to be correct, while for moderately sized rectangles ($\beta = .5$) there may be a gap between our theoretical bound, 13 (b), and the actual performance of GFSS. For the Kronecker graph there appears to be a gap for both scalings ($k = 1$ and $k = 2$) of cluster size, indicating that the performance of the GFSS may be better than predicted by our theory.

VII. Conclusion

We studied the problem of how to tractably detect anomalous activities in networks under Gaussian noise. We motivated this problem by using it to detect Arsenic in Idaho groundwater. We analyzed the GLRT, outlined the derivation of the spectral scan statistic, and proposed the graph Fourier scan statistic. We completely characterized the performance of the GFSS for any graph in terms of the spectrum of the combinatorial Laplacian. The theoretical analysis of the GFSS naturally led to the development of the adaptive GFSS.

Our experiments have demonstrated the ability of the adaptive GFSS to outperform the GFSS and the naive estimators. We applied the main result to three graph models: balanced binary trees, the lattice and Kronecker graph. Our theoretical results are superior to those proven for the spectral scan statistic and we have shown that the GFSS is nearly optimal for the balanced binary tree. We demonstrated that though the theoretical performance of the GFSS for the Torus graph and Kronecker graph may be sub-optimal, there is experimental evidence to indicate that this is partly an artifact of the theoretical analysis technique. We see that not only is it statistically sub-optimal to ignore graph structure, but in many of these cases the GFSS gives a near optimal performance.

Appendix A
Proofs

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

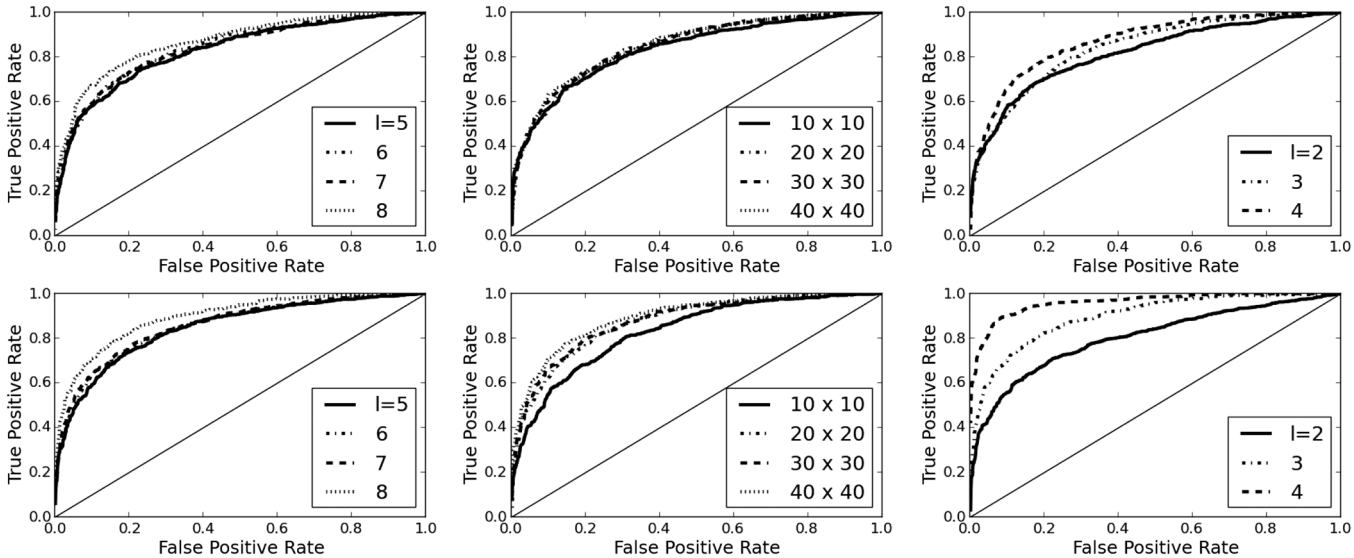

Fig. 8. (**Rescaling by Theoretical Bounds**) The size (false positive rate) and power (true positive rate) of the GFSS as p increases for the following graph and signal models: BBT with $\alpha = 1$ (top left) and $\alpha = .5$ (bottom left); Torus with $\beta = 0$ (top middle) and $\beta = .5$ (bottom middle); Kronecker graph with base graph size $p_0 = 6$ and $k = 1$ (top right) and $k = 2$ (bottom right). The SNR was allowed to scale according to Cor. 12 (b) (left), Cor. 13 (b) (middle), Cor. 14 (b) (right).

Proof of Proposition 4: To prove the claim we will first rewrite the SSS in an equivalent but more convenient form which we will then bound from above and below using the GFSS. To this end we recall the arguments from Lemma 7 of [6]. Since G is connected, the combinatorial Laplacian Δ is symmetric, its smallest eigenvalue is zero and the remaining eigenvalues are positive. By the spectral theorem, we can write $\Delta = \mathbf{U}\Lambda\mathbf{U}^\top$, where Λ is a $(p-1)\times(p-1)$ diagonal matrix containing the positive eigenvalues of Δ, $\lambda_2, \ldots, \lambda_p$, in increasing order. The columns of the $p \times (p-1)$ matrix $\mathbf{U}$ are the associated eigenvectors. Then, since each vector $\mathbf{x} \in \mathbb{R}^p$ with $\mathbf{1}^\top \mathbf{x} = 0$ can be written as $\mathbf{U}\mathbf{z}$ for a unique vector $\mathbf{z} \in \mathbb{R}^{p-1}$, we have

$$\begin{aligned}\mathcal{X} &= \{\mathbf{x} \in \mathbb{R}^p : \mathbf{x}^\top \Delta \mathbf{x} \le \rho, \mathbf{x}^\top \mathbf{x} = 1, \mathbf{1}^\top \mathbf{x} \le 0\} \\ &= \{\mathbf{U}\mathbf{z} : \mathbf{z} \in \mathbb{R}^{p-1}, \mathbf{z}^\top \mathbf{U}^\top \Delta \mathbf{U}\mathbf{z} \le \rho, \mathbf{z}^\top \mathbf{U}^\top \mathbf{U}\mathbf{z} \le 1\} \\ &= \left\{\mathbf{U}\mathbf{z} : \mathbf{z} \in \mathbb{R}^{p-1}, \frac{1}{\rho}\mathbf{z}^\top \Lambda \mathbf{z} \le 1, \mathbf{z}^\top \mathbf{z} \le 1\right\}\end{aligned}$$

where in the third identity we have used the fact that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_{p-1}$. Letting $\mathcal{Z} = \left\{\mathbf{z} \in \mathbb{R}^{p-1} : \frac{1}{\rho}\mathbf{z}^\top \Lambda \mathbf{z} \le 1, \mathbf{z}^\top \mathbf{z} \le 1\right\}$, we see that the SSS can be equivalently expressed as

$$\sqrt{\widehat{s}} = \sup_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \mathbf{y} = \sup_{\mathbf{z} \in \mathcal{Z}} \mathbf{z}^\top \mathbf{U}^\top \mathbf{y}. \tag{14}$$

Next, let $\mathbf{A} = \frac{1}{\rho}\Lambda = \operatorname{diag}\{a_i\}_{i=1}^{p-1}$, where $a_i = \lambda_{i+1}/\rho$, for $i = 1, \ldots, p-1$. If $\mathbf{z} \in \mathbb{R}^{p-1}$ satisfies $\|\mathbf{z}\| \le 1$ and $\mathbf{z}^\top \mathbf{A}\mathbf{z} \le 1$, then

$$\sum_{i=1}^{p} \max\{1, a_i\} z_i^2 \le \|\mathbf{z}\|^2 + \mathbf{z}^\top \mathbf{A}\mathbf{z} \le 2.$$

Similarly, if $\sum_{i=1}^{p} \max\{1, a_i\} z_i^2 \le 1$, then we must have $\max\{\|\mathbf{z}\|, \mathbf{z}^\top \mathbf{A}\mathbf{z}\} \le 1$ as well. Now let $\mathbf{A}'$ be the $(p-1)$-dimensional diagonal matrix with entries $\max\{1, a_i\}$, $i = 1, \ldots, p-1$ and set $\mathcal{Z}_1 = \{\mathbf{z} \in \mathbb{R}^{p-1} : \mathbf{z}^\top \mathbf{A}'\mathbf{z} \le 1\}$ and $\mathcal{Z}_2 = \{\mathbf{z} \in \mathbb{R}^{p-1} : \mathbf{z}^\top \mathbf{A}'\mathbf{z} \le 2\}$. Thus we have shown that

$$\mathcal{Z}_1 \subset \mathcal{Z} \subset \mathcal{Z}_2.$$

Using (14), the previous inclusions imply the following bounds on the square root of the SSS:

$$\sup_{\mathbf{z} \in \mathcal{Z}_1} \mathbf{z}^\top \mathbf{U}^\top \mathbf{y} \le \sqrt{\widehat{s}} \le \sup_{\mathbf{z} \in \mathcal{Z}_2} \mathbf{z}^\top \mathbf{U}^\top \mathbf{y}$$

which in turn are equivalent to the bounds

$$\sup_{\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{U}\mathbf{A}'\mathbf{U}^\top \mathbf{z} \le 1\}} \mathbf{y}^\top \mathbf{z} \le \sqrt{\widehat{s}} \le \sup_{\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{U}\mathbf{A}'\mathbf{U}^\top \mathbf{z} \le 2\}} \mathbf{y}^\top \mathbf{z}$$

since every $\mathbf{z} \in \mathbb{R}^{p-1}$ can be written as $\mathbf{U}^\top \mathbf{z}$ for some $\mathbf{z} \in \mathbb{R}^p$.[1]

All that remains is to show that

$$\hat{t} = \sup_{\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{U}\mathbf{A}'\mathbf{U}^\top \mathbf{z} \le 1\}} \mathbf{y}^\top \mathbf{z}.$$

This can be seen by strong duality for convex programs,

$$\begin{aligned}&\sup_{\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{U}\mathbf{A}'\mathbf{U}^\top \mathbf{z} \le 1\}} \mathbf{y}^\top \mathbf{z} \\ &= \sup_{\{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{A}'\mathbf{z} \le 1\}} (\mathbf{U}^\top \mathbf{y})^\top \mathbf{z} \\ &= \sup_{\{\mathbf{z} \in \mathbb{R}^{p-1}\}} \inf_{\eta \ge 0} (\mathbf{U}^\top \mathbf{y})^\top \mathbf{z} - \eta(\mathbf{z}^\top \mathbf{A}'\mathbf{z} - 1) \\ &= \inf_{\eta \ge 0} \sup_{\{\mathbf{z} \in \mathbb{R}^{p-1}\}} (\mathbf{U}^\top \mathbf{y})^\top \mathbf{z} - \eta(\mathbf{z}^\top \mathbf{A}'\mathbf{z} - 1).\end{aligned}$$

The solution to the maximization problem is $\mathbf{z} = (2\eta\mathbf{A}')^{-1}(\mathbf{U}^\top \mathbf{y})$, and plugging this in it becomes

$$\inf_{\eta \ge 0} (\mathbf{U}\mathbf{y})^\top (4\eta\mathbf{A}')^{-1} (\mathbf{U}^\top \mathbf{y}) + \eta$$

[1]In fact, $\mathbf{z} = \mathbf{U}^\top \mathbf{z}_1 = \mathbf{U}^\top \mathbf{z}_2$ if and only if the difference $\mathbf{z}_1 - \mathbf{z}_2$ belongs to the linear subspace of $\mathbb{R}^p$ spanned by the constant vectors.

which is minimized at

$$\eta = \sqrt{(\mathbf{U}\mathbf{y})^\top (4\mathbf{A}')^{-1}(\mathbf{U}\mathbf{y})}.$$

Plugging this in completes our proof. □

Proof of Theorem 5: We will use the following lemma regarding the concentration of χ^2 random variables.

Lemma 15 ([36]): Let for $i \in \{2,\dots,p\}$, $a_i \geq 0$ and $\{X_i\}_{i=1}^p$ be independent χ_1^2 random variables. Define $Z = \sum_{i=1}^p a_i(X_i - 1)$

$$\mathbb{P}\left\{Z \geq 2\|\mathbf{a}\|_2\sqrt{x} + 2\|\mathbf{a}\|_\infty x\right\} \leq e^{-x}$$
$$\mathbb{P}\left\{Z \leq -2\|\mathbf{a}\|_2\sqrt{x}\right\} \leq e^{-x}$$

Recall the notation of the proof of Prop. 4. The probability of error under the null, (11), follows from Lemma 15. Consider any of the alternatives, then $\hat{t}$ can be written,

$$\begin{aligned}\hat{t} &= \mathbf{y}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{y} - \operatorname{tr}(\mathbf{A}')^{-1} \\ &= \mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \boldsymbol{\epsilon} \\ &\quad + \boldsymbol{\epsilon}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \boldsymbol{\epsilon} - \operatorname{tr}(\mathbf{A}')^{-1} \\ &\stackrel{d}{=} \mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top (\mathbf{A}')^{-1}\boldsymbol{\epsilon} \\ &\quad - \operatorname{tr}(\mathbf{A}')^{-1}\end{aligned}$$

where $\stackrel{d}{=}$ denotes equality in distribution (which follows from rotational invariance of the isonormal Gaussian). By Gaussian concentration, with probability at least $1-\alpha$,

$$\mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\boldsymbol{\epsilon} \geq -\sqrt{2\mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-2}\mathbf{U}^\top \mathbf{x}\log\left(\frac{1}{\alpha}\right)}.$$

Because $\mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top$ is positive definite with eigenvalues bounded by 1, we have that $\mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-2}\mathbf{U}^\top \mathbf{x} \leq \mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{x}$. We will now show that $\mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{x} \geq \mu^2/2$ under H_1. Recall that by the dual norm (as derived in the proof of Prop. 4),

$$\mathbf{x}^\top \mathbf{U}\mathbf{A}'\mathbf{U}^\top \mathbf{x} = \sup_{\mathbf{z}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{z} \leq 1} (\mathbf{z}^\top \mathbf{x})^2. \tag{15}$$

Let $\mathbf{x} \in \mathcal{X}(\mu, C)$. Let $\mathbf{K}_C$ be the projection onto the span of $\mathbf{1}_C, \mathbf{1}_{\bar{C}}$ and orthogonal to $\mathbf{1}$. So,

$$\mathbf{K}_C\mathbf{x} = \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|}\mathbf{1}_C + \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|}\mathbf{1}_{\bar{C}} - \bar{\mathbf{x}}.$$

Let $\mathbf{z} = \mathbf{K}_C\mathbf{x}/\|\mathbf{K}_C\mathbf{x}\|$ that $\mathbf{z}^\top \mathbf{x} = \|\mathbf{K}_C\mathbf{x}\|$. Let

$$\bar{x}_C = \frac{\mathbf{1}_C^\top \mathbf{x}}{|C|} \quad \text{and} \quad \bar{x}_{\bar{C}} = \frac{\mathbf{1}_{\bar{C}}^\top \mathbf{x}}{|\bar{C}|}.$$
$$\begin{aligned}\bar{x}_C - \bar{x} &= \left(\frac{1}{|C|} - \frac{1}{p}\right)\mathbf{1}_C^\top \mathbf{x} - \frac{1}{p}\mathbf{1}_{\bar{C}}^\top \mathbf{x} \\ &= \frac{|\bar{C}|}{p}(\bar{x}_C - \bar{x}_{\bar{C}}).\end{aligned}$$

Then similarly, $\bar{x}_{\bar{C}} - \bar{x} = \frac{|C|}{p}(\bar{x}_{\bar{C}} - \bar{x}_C)$. And so,

$$\begin{aligned}(\mathbf{z}^\top \mathbf{x})^2 &= \|\mathbf{K}_C\mathbf{x}\|^2 \\ &= |C|\frac{|\bar{C}|^2}{p^2}(\bar{x}_C - \bar{x}_{\bar{C}})^2 + |\bar{C}|\frac{|C|^2}{p^2}(\bar{x}_{\bar{C}} - \bar{x}_C)^2 \\ &= \frac{|C||\bar{C}|}{p}(\bar{x}_{\bar{C}} - \bar{x}_C)^2 \geq \mu^2.\end{aligned}$$

Furthermore,

$$\mathbf{z}^\top \Delta \mathbf{z} = \frac{p\mathbf{W}(\partial C)}{|C||\bar{C}|} \leq \rho$$

while $\|\mathbf{z}\| = 1$. Thus, $\mathbf{z}\mathbf{U}\mathbf{A}'\mathbf{U}^\top \mathbf{z} \leq 2$ and so, $(\mathbf{z}/\sqrt{2})^\top \mathbf{U}\mathbf{A}'\mathbf{U}^\top(\mathbf{z}/\sqrt{2}) \leq 1$. By substituting $\mathbf{z} \leftarrow \mathbf{z}/\sqrt{2}$ in (15) we arrive at

$$\Rightarrow \mathbf{x}^\top \mathbf{U}(\mathbf{A}')^{-1}\mathbf{U}^\top \mathbf{x} \geq \frac{\mu^2}{2}.$$

The error bound, (12) follows from these facts and the Lemma 15 applied to $\boldsymbol{\epsilon}^\top(\mathbf{A}')^{-1}\boldsymbol{\epsilon} - \operatorname{tr}(\mathbf{A}')^{-1}$. □

Proof of Corollary 12: The study of the spectra of trees really began in earnest with the work of [44]. Notably, it became apparent that trees have eigenvalues with high multiplicities, particularly the eigenvalue 1. [45] 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

$$\begin{aligned}\frac{1}{\lambda_2^{(\ell)}} &\leq 2^\ell - 2\ell + 2 - \frac{2^\ell - \sqrt{2}(2\ell - 1 - 2^{\ell-1})}{2^\ell - 1 - \sqrt{2}(2^{\ell-1} - 1)} \\ &\quad + \left(3 - 2\sqrt{2}\cos\left(\frac{\pi}{2\ell - 1}\right)\right)^{-1} \\ &\leq 2^\ell + 105I\{\ell < 4\}\end{aligned} \tag{16}$$

[46] 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 Δ is given by

$$\begin{aligned}\det(\lambda\mathbf{I} - \Delta) = p_1^{2^{\ell-2}}(\lambda)p_2^{2^{\ell-3}}(\lambda) \\ \dots p_{\ell-3}^{2^2}(\lambda)p_{\ell-2}^2(\lambda)p_{\ell-1}(\lambda)s_\ell(\lambda)\end{aligned} \tag{17}$$

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 (16) with ℓ replaced with i. In [47], they extended this work to more general balanced trees.

By (17) we know that at most $\ell + (\ell-1) + (\ell-2)2 + \dots + (\ell-j)2^{j-1} \leq \ell 2^j$ eigenvalues have reciprocals larger than $2^{\ell-j} + 105I\{j < 4\}$. Let $k = \max\left\{\left\lceil \frac{\ell}{c}2^{\ell(1-\alpha)}\right\rceil, 2^3\right\}$, then we have ensured that at most k eigenvalues are smaller than ρ. For n large enough

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

Proof of Cor. 13:

(a) By a simple Fourier analysis (see [13]), we know that the Laplacian eigenvalues are $2(2-\cos(2\pi i_1/\ell)-\cos(2\pi i_2/\ell))$ for all $i_1,i_2 \in [\ell]$. Let us denote the ℓ^2 eigenvalues as $\lambda_{(i_1,i_2)}$ for $i_1,i_2 \in [\ell]$. Notice that for $i \in [\ell]$, $|\{(i_1,i_2): i_1 \vee i_2 = i\}| \le 2i$. For simplicity let ℓ be even. We know that if $i_1 \vee i_2 \le \ell/2$ then $\lambda_{(i_1,i_2)} = 2-\cos(2\pi i_1/\ell)-\cos(2\pi i_2/\ell) \ge 1-\cos(2\pi(i_1 \vee i_2)/\ell)$. Let $k \ll \ell$ which we will specify later. Thus,

$$\begin{aligned}
&\sum_{(i_1,i_2)\neq(1,1)\in[\ell]^2} 1 \wedge \frac{\rho^2}{\lambda^2_{(i_1,i_2)}} \\
&\le 2\sum_{i\in\left[\frac{\ell}{2}\right]} 2i\left(1\wedge\frac{\rho^2}{\left(1-\cos\left(\frac{2\pi i}{\ell}\right)\right)^2}\right) \\
&\le 4\sum_{i=1}^{k} i + \rho^2\frac{\ell^2}{2}\frac{2}{\ell}\sum_{k<i\le\frac{\ell}{2}} 2\frac{\frac{i}{\ell}}{\left(1-\cos\left(\frac{2\pi i}{\ell}\right)\right)^2} \\
&\le 4k^2 + \rho^2\frac{\ell^2}{2}\int_{\frac{k}{\ell}}^{\frac{1}{2}} \frac{x\,dx}{\left(1-\cos(2\pi x)\right)^2} \\
&= 4k^2 + \rho^2\frac{\ell^2}{2}\left(\frac{1}{4\pi^4}\frac{\ell^3}{k^3}+O\left(\frac{\ell}{k}\right)\right)
\end{aligned}$$

The above followed by the Taylor expansion about 0 of the integral. Let us choose k to such that $k \approx \rho^{2/5}\ell$. The inequalities above hold regardless of the choice of k, as long $k \ll \ell$, so we have the freedom to tune it to our liking. Plugging this in we obtain,

$$\sum_{(i_1,i_2)\neq(1,1)\in[\ell]^2} 1\wedge\frac{\rho^2}{\lambda^2_{(i_1,i_2)}} = O\left(\rho^{\frac{4}{5}}\ell^2\right) = O(p^{3/5+2\beta/5})$$

(b) We will construct $\mathcal{C}'$ in Theorem 11 (b) from disjoint squares of size a constant multiple of $p^{1-\beta}$, making $|\mathcal{C}'| \asymp p^{\beta}$.

□

Proof of Corollary 14: 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 Δ_1,Δ_2 then the Laplacian of their Kronecker product, $H_1\otimes H_2$, is given by $\Delta = \Delta_1\otimes\mathbf{I}_p + \mathbf{I}_p\otimes\Delta_2$ ([48]). Hence, if $\mathbf{v}_1,\mathbf{v}_2\in\mathbb{R}^p$ are eigenvectors, viz. $\Delta_1\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $\Delta_2\mathbf{v}_2 = \lambda_2\mathbf{v}_2$, then $\Delta(\mathbf{v}_1\otimes\mathbf{v}_2) = (\lambda_1+\lambda_2)\mathbf{v}_1\otimes\mathbf{v}_2$, 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 $\rho \asymp p^{2k-\ell-1}$, by showing that it is the results of cuts at level k. We say that an edge $e = ((i_1,\ldots,i_\ell),(j_1,\ldots,j_\ell))$ 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

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

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

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

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

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

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 $(p-1)/p$. Hence,

$$\begin{aligned}
&\frac{1}{p^\ell}\sum_{i\in V^\ell}\min\left\{1,\frac{\rho^2}{\lambda_i^2}\right\} \\
&\le \mathbb{E}_Z\min\left\{1,\frac{\rho^2 p^{2Z}}{\nu_2^2}\right\} \\
&\le \mathbb{P}_Z\{Z\ge 2k-\ell-1+\log_p\nu_2\} \\
&\quad + \frac{1}{\nu_2^2}\sum_{z=1}^{\lfloor \ell+1-2k+\log_p\nu_2\rfloor} p^{2(2k-\ell-1+z)}\mathbb{P}_Z\{Z=z\} \\
&\le p^{2k-\ell-1+\log_p\nu_2} \\
&\quad + \frac{1}{\nu_2^2}\sum_{z=1}^{\lfloor \ell+1-2k+\log_p\nu_2\rfloor} p^{2(z+2k-\ell-1)}\frac{1}{p^z}\frac{p-1}{p} \\
&= O\left(\left(\nu_2+\nu_2^{-1}\right)p^{2k-\ell-1}\right) = O\left(p^{2k-l}\mathrm{diam}(H)\right)
\end{aligned}$$

where $\mathrm{diam}(H)$ is the diameter of the base graph H. Hence,

$$\sum_{i\in V^\ell}\min\left\{1,\frac{\rho^2}{\lambda_i^2}\right\} = O\left(n^{\frac{2k}{\ell}}\mathrm{diam}(H)\right)$$

□

APPENDIX B
THE LR STATISTIC

Below we will provide the details for the derivation of the LR statistic (7) for testing the null hypothesis that $\mathbf{x} = \bar{\mathbf{x}}$ versus the alternative hypothesis

$$\mathbf{x} = \alpha\mathbf{1} + \mathbf{x}_0, : \alpha,\ \delta\in\mathbb{R}$$

for one given $\mathbf{x}_0 \neq 0$, $\mathbf{x}_0^\top\mathbf{1} = 0$. The unknown parameter α is a nuisance parameter.

To eliminate the dependence on α and simplify the problem we will resort to invariant testing theory [27]. In fact, the testing problem remains invariant under the action of the group of translations, i.e. additions of constant vectors, of the mean of $\mathbf{y}$. To take advantage of such invariance we proceed as follows. Let $\mathbf{B}$ be a $(p-1)\times p$ whose rows form an orthonormal basis for $\mathcal{R}^\perp(\mathbf{1})$, the linear subspace of $\mathbb{R}^p$ orthogonal to the subspace of vectors in $\mathbb{R}^p$ with constant entries (the matrix $\mathbf{U}^\top$ as defined in the proof of Prop. 4 would suffice). Then, a maximal invariant

with respect with respect to such a group is the $(p-1)$-dimensional random vector

$$\mathbf{z} := \mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{x}_0 + \mathbf{B}\boldsymbol{\epsilon}.$$

Since $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_{p-1}$ and $\mathbf{B}\mathbf{x}_0 = \mathbf{x}_0$, $\mathbf{z}$ has a $N_{p-1}(\mathbf{x}_0, \sigma^2\mathbf{I}_{p-1})$ distribution, which no longer depends on the nuisance parameter α. The hypothesis testing problem (6) is then equivalent to the problem of testing $H_0 : \mathbb{E}[\mathbf{z}] = 0$ versus the alternative $H_1^{\mathbf{x}_0} : \mathbb{E}[\mathbf{z}] = \mathbf{x}_0$. It is also worth pointing out that, as our calculations below show, the choice of the orthonormal basis of $\mathcal{R}^\perp(\mathbf{1})$ comprising the rows of the matrix $\mathbf{B}$ does not matter in the construction of the optimal test.

The LR statistic is

$$\frac{\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{z}-\mathbf{x}_0\|^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{z}\|^2\right\}}$$

which is equal to

$$\exp\left\{\frac{1}{2\sigma^2}\left(\|\mathbf{x}_0\|^2 - 2\mathbf{x}_0^\top\tilde{\mathbf{y}}\right)\right\} \tag{18}$$

because $\mathbf{z} = \mathbf{B}\mathbf{y} = \mathbf{y} - \bar{\mathbf{y}} = \tilde{\mathbf{y}}$.

We now will verify (8). Let $\mathbf{x} \in \mathcal{X}^\perp(\mu, C)$ and notice that $\mathbf{z}_C^\top\mathbf{1} = 0$. Moreover,

$$\|\mathbf{z}_C\|^2 = \frac{|C||\bar{C}|}{p}\left(\frac{1}{|C|} + \frac{1}{|\bar{C}|}\right) = 1.$$

Define $\alpha_C = \mathbf{z}_C^\top\mathbf{x}$ and $\mathbf{x}_1 = \mathbf{x} - \alpha_C\mathbf{z}_c$, then

$$\begin{aligned} 2\sigma^2\log\Lambda_{\mathbf{x}}(\mathbf{y}) &= 2\mathbf{x}^\top\tilde{\mathbf{y}} - \|\mathbf{x}\|^2 \\ &= 2\alpha_C\mathbf{z}_C^\top\tilde{\mathbf{y}} + 2\mathbf{x}_1^\top\tilde{\mathbf{y}} - \alpha_C^2 - \|\mathbf{x}_1\|^2. \end{aligned}$$

Notice that

$$\mathcal{X}^\perp(\mu, C) = \{\mathbf{x} \in \mathbb{R}^p : |\mathbf{x}^\top\mathbf{z}_C| \geq \mu, \mathbf{x}^\top\mathbf{1} = 0\}$$

and so for the $\mathbf{x}$ defined above, $\mathbf{x} \in \mathcal{X}^\perp(\mu, C)$ iff $|\alpha_C| \geq \mu$. We can see that $2\mathbf{x}_1^\top\tilde{\mathbf{y}} - \|\mathbf{x}_1\|^2$ is maximized at $\mathbf{x}_1 = \tilde{\mathbf{y}} - \mathbf{z}_C^\top\tilde{\mathbf{y}}\mathbf{z}_C$. Furthermore, $2\alpha_C\mathbf{z}_C^\top\tilde{\mathbf{y}} - \alpha_C^2$ is maximized at $\alpha_C = \mathbf{z}_C^\top\tilde{\mathbf{y}}$, but this is restricted if $|\mathbf{z}_C^\top\tilde{\mathbf{y}}| \leq \mu$. Hence, evaluating the objective gives us

$$\begin{aligned} &\max_{\mathbf{x}\in\mathcal{X}^\perp(\mu,C)} 2\sigma^2\log\Lambda_{\mathbf{x}}(\mathbf{y}) \\ &\qquad = \begin{cases} \|\tilde{\mathbf{y}}\|^2, & \text{if } \left|\mathbf{z}_C^\top\tilde{\mathbf{y}}\right| \geq \mu \\ \|\tilde{\mathbf{y}}\|^2 - \left(\left|\mathbf{z}_C^\top\tilde{\mathbf{y}}\right| - \mu\right)^2, & \text{if } \left|\mathbf{z}_C^\top\tilde{\mathbf{y}}\right| < \mu. \end{cases} \end{aligned}$$

REFERENCES

[1] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[2] A. Sandryhaila and J. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, 2013.

[3] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 53–94, 2006.

[4] F. Murtagh, "The Haar wavelet transform of a dendrogram," *J. Classification*, vol. 24, pp. 3–32, 2007.

[5] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 129–150, 2011.

[6] J. Sharpnack, A. Rinaldo, and A. Singh, "Changepoint detection over graphs with the spectral scan statistic," in *Proc. AISTATS*, 2013, pp. 545–553.

[7] M. Focazio, A. Welch, S. Watkins, D. Helsel, and M. Horn, "A retrospective analysis on the occurrence of arsenic in ground-water resources of the united states and limitations in drinking-water-supply characterizations," Washington, DC, USA, U.S. Geological Survey Water-Resources Investigation Rep. 99-4279, 1999.

[8] W. Härdle, G. Kerkyacharian, A. Tsybakov, and D. Picard, *Wavelets, Approximation, Statistical Applications*. New York, NY, USA: Springer, 1998.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. NIPS*, 2001, vol. 14, pp. 585–591.

[10] A. Y. Ng, M. I. Jordan, and Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," in *Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 849–856.

[11] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh, "Noise thresholds for spectral clustering," in *Adv. Neural Inf. Process. Syst.*, 2011, pp. 954–962.

[12] J. Nilsson, F. Sha, and M. I. Jordan, "Regression on manifolds using kernel dimension reduction," in *Proc. 24th ACM Int. Conf. Mach. Learn.*, 2007, pp. 697–704.

[13] J. Sharpnack and A. Singh, "Identifying graph-structured activation patterns in networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2010.

[14] C. E. Priebe, "Scan statistics on graphs," Johns Hopkins Univ., Baltimore, MD, USA, Tech. Rep. 650, 2004.

[15] H. Wang, M. Tang, Y. Park, and C. E. Priebe, "Locality statistics for anomaly detection in time series of graphs," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 703–717, 2014.

[16] E. Arias-Castro, D. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2402–2425, 2005.

[17] E. Arias-Castro, E. Candes, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *Ann. Statist.*, vol. 36, no. 4, pp. 1726–1757, 2008.

[18] E. Arias-Castro, E. Candes, and A. Durand, "Detection of an anomalous cluster in a network," *Ann. Statist.*, vol. 39, no. 1, pp. 278–304, 2011.

[19] S. Speakman and D. B. Neill, "Fast graph scan for scalable detection of arbitrary connected clusters," in *Proc. Int. Soc. Disease Surveill. Annu. Conf.*, 2010.

[20] J. Sharpnack, A. Krishnamurthy, and A. Singh, "Detecting activations over graphs using spanning tree wavelet bases," in *Proc. Artif. Intell. Statist. (AISTATS)*, 2013.

[21] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, "On combinatorial testing problems," *Ann. Statist.*, vol. 38, no. 5, pp. 3063–3092, 2010.

[22] J. L. Sharpnack, A. Krishnamurthy, and A. Singh, "Near-optimal anomaly detection in graphs using Lovász extended scan statistic," in *Adv. Neural Inf. Process. Syst.*, 2013, pp. 1959–1967.

[23] J. Sharpnack and A. Singh, "Near-optimal and computationally efficient detectors for weak and sparse graph-structured patterns," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2013.

[24] V. V. Vazirani, *Approximation Algorithms*. New York, NY, USA: Springer, 2001.

[25] V. Spokoiny, "Adaptive hypothesis testing using wavelets," *Ann. Statist.*, vol. 24, no. 6, pp. 2477–2498, 1996.

[26] P. Ji and M. Nussbaum, "Sharp adaptive nonparametric testing for Sobolev ellipsoids," 2012, arXiv preprint arXiv:1210.8162.

[27] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*. New York, NY, USA: Springer-Verlag, 2005.

[28] M. Fouladirad, L. Freitag, and I. Nikiforov, "Optimal fault detection with nuisance parameters and a general covariance matrix," *Int. J. Adapt. Control Signal Process.*, vol. 22, no. 5, pp. 431–439, 2008.

[29] M. Fouladirad and I. Nikiforov, "Optimal statistical fault detection with nuisance parameters," *Automatica*, vol. 41, no. 7, pp. 1157–1171, 2005.

[30] L. Fillatre, "Asymptotically uniformly minimax detection and isolation in network monitoring," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3357–3371, 2012.

[31] L. L. Scharf and B. Friedlander, "Matched sub-space detectors," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2146–2157, 1994.

[32] B. Baygün and A. O. Hero, "Optimal simultaneous detection and estimation under a false alarm constraint," *IEEE Trans. Signal Process.*, vol. 41, no. 3, pp. 688–703, 1995.

[33] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*, vol. 54, pp. 426–482, 1943.
[34] Y. Ingster and I. Suslina, *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. New York, NY, USA: Springer-Verlag, 2003, vol. 169.
[35] D. Matula and F. Shahrokhi, "Sparsest cuts and bottlenecks in graphs," *Discrete Appl. Math.*, vol. 27, no. 1, pp. 113–123, 1990.
[36] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000.
[37] I. Koutis, G. L. Miller, and R. Peng, "Approaching optimality for solving SDD linear systems," in *Proc. 51st Annu. IEEE Symp. IEEE Found. Comput. Sci. (FOCS)* , 2010, pp. 235–244.
[38] J. Leskovec and C. Faloutsos, "Scalable modeling of real graphs using kronecker multiplication," in *Proc. 24th ACM Int. Conf. Mach. Learn.*, 2007, pp. 497–504.
[39] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, pp. 985–1042, 2010.
[40] Principal Aquifers of the 48 Conterminous United States, Hawaii, Puerto Rico, the U.S. Virgin Islands, 2003 [Online]. Available: http://water.usgs.gov/lookup/getspatial?aquifers_us
[41] A. Singh, R. D. Nowak, and R. Calderbank, "Detecting weak but hierarchically-structured patterns in networks," 2010, arXiv preprint arXiv:1003.0205.
[42] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *ACM SIGCOMM Comput. Commun. Review*, vol. 29, pp. 251–262, 1999.
[43] S. Milgram, "The small world problem," *Psychol. Today*, vol. 2, no. 1, pp. 60–67, 1967.
[44] M. Fiedler, "Eigenvectors of acyclic matrices," *Czech. Math. J.*, vol. 25, no. 4, pp. 607–618, 1975.
[45] J. Molitierno, M. Neumann, and B. Shader, "Tight bounds on the algebraic connectivity of a balanced binary tree," *Electron. J. Linear Algebra*, vol. 6, pp. 62–71, 2000.
[46] O. Rojo, "The spectrum of the laplacian matrix of a balanced binary tree," *Linear Algebra Appl.*, vol. 349, no. 1, pp. 203–219, 2002.
[47] O. Rojo and R. Soto, "The spectra of the adjacency matrix and laplacian matrix for some balanced trees," *Linear Algebra Appl.*, vol. 403, pp. 97–117, 2005.
[48] R. Merris, "Laplacian graph eigenvectors," *Linear Algebra Appl.*, vol. 278, no. 1, pp. 221–236, 1998.

James Sharpnack (M'14) received a B.S. in Math and Physics from Ohio State University and a Master's in Statistics and Ph.D. in Machine Learning and Statistics from Carnegie Mellon University. He was previously a postdoctoral researcher at UC San Diego in the Mathematics department. He is an assistant professor at UC Davis in the Statistics Department. His research focuses on the development of algorithms with provably low statistical risk and low computational complexity to new statistical problems with complex structural and modeling assumptions. Specifically, he has focused on graph-structured normal means testing and regression.

Alessandro Rinaldo received the Bachelor's degree in Economics from the Bocconi University in Milan, and the M.S. and Ph.D. degrees in Statistics from Carnegie Mellon University in 2001 and 2005, respectively. He is currently an Associate Professor in the Department of Statistics at Carnegie Mellon University. His research interests include high-dimensional statistics, machine learning, network models and geometric data analysis. His work has been recognized by an NSF CAREER Award and the Umesh Gavaskar Memorial Thesis Award.

Aarti Singh received her B.E. in Electronics and Communication Engineering from the University of Delhi in 2001, and M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Wisconsin-Madison in 2003 and 2008, respectively. She was a Postdoctoral Research Associate at the Program in Applied and Computational Mathematics at Princeton University from 2008–2009, before joining the School of Computer Science at Carnegie Mellon in 2009 where she is currently an Assistant Professor. Her research interests lie at the intersection of machine learning, statistics and signal processing, and focus on designing statistically and computationally efficient algorithms that can leverage inherent structure of the data in the form of clusters, graphs, subspaces and manifold using direct, compressive and active queries. Her work is recognized by an NSF Career Award, a United States Air Force Young Investigator Award, A. Nico Habermann Faculty Chair Award, Harold A. Peterson Best Dissertation Award, and a best student paper award.