A Spectral Framework for Anomalous Subgraph Detection

Benjamin A. Miller, *Senior Member, IEEE*, Michelle S. Beard, Patrick J. Wolfe, *Senior Member, IEEE*, and Nadya T. Bliss, *Senior Member, IEEE*

***Abstract*—A wide variety of application domains is concerned with data consisting of entities and their relationships or connections, formally represented as graphs. Within these diverse application areas, a common problem of interest is the detection of a subset of entities whose connectivity is anomalous with respect to the rest of the data. While the detection of such anomalous subgraphs has received a substantial amount of attention, no application-agnostic framework exists for analysis of signal detectability in graph-based data. In this paper, we describe a framework that enables such analysis using the principal eigenspace of a graph's residuals matrix, commonly called the modularity matrix in community detection. Leveraging this analytical tool, we show that the framework has a natural power metric in the spectral norm of the anomalous subgraph's adjacency matrix (signal power) and of the background graph's residuals matrix (noise power). We propose several algorithms based on spectral properties of the residuals matrix, with more computationally expensive techniques providing greater detection power. Detection and identification performance are presented for a number of signal and noise models, including clusters and bipartite foregrounds embedded into simple random backgrounds, as well as graphs with community structure and realistic degree distributions. The trends observed verify intuition gleaned from other signal processing areas, such as greater detection power when the signal is embedded within a less active portion of the background. We demonstrate the utility of the proposed techniques in detecting small, highly anomalous subgraphs in real graphs derived from Internet traffic and product co-purchases.**

***Index Terms*—Graph theory, signal detection theory, spectral analysis, residuals analysis, principal components analysis.**

I. Introduction

In numerous applications, the data of interest consist of entities and the relationships between them. In social network analysis, for example, the data are connections between individuals, such as who knows whom personally, who is in the same organization, or who is connected on a social networking website. In computer networks, we are often interested in which computers communicate with one another. In the natural sciences, we may want to know which chemicals interact in a reaction. Across these varied domains, data regarding connections, relationships, and interactions between discrete entities enhance situational awareness and diversify analysis by incorporating additional contextual information.

When working with relational data, it is common to formally represent the relationships as a graph. A graph $G = (V, E)$ is a pair of sets: a set of vertices, V, comprising the entities, and a set of edges, E, denoting relationships between them. Graph theory provides an abstract mathematical object that has been applied in all of the above contexts. Indeed, graphs have been used to model protein interactions [1] and to represent communication between computers [2]. Graphs—commonly called networks in practice—are used extensively in social network analysis, with many graph algorithms focused on detection of communities [3], [4] and influential figures [5].

As a data structure, graphs have long been utilized by signal processing practitioners. Analysis of graphs derived from radio frequency or image data is common, as a graph structure can help classify similar measurements (see, e.g., [6]). Recent research has also defined traditional digital signal processing kernels—such as filtering and Fourier transforms—for signals that propagate along edges in a graph [7], [8]. When the graph comprises the data itself, rather than a means of storage, significant complications arise. Graphs are discrete, combinatorial structures, and, thus, they lack the convenient mathematical context of Euclidean vector spaces. The ability to perform linear transformations and the analytical tractability of working with Gaussian noise are not available in general when working with relational data. Deriving an optimal detector for a small signal subgraph buried within a large network, then, becomes potentially intractable, as it may require the solution to an NP-hard problem.

Despite these complications, it is desirable to understand notions of detectability of small subgraphs embedded within a large background. The ability to detect small signals in these contexts would be useful in many domains, from the detection of malicious traffic in a computer network to the discovery of threatening activity in a social network. Recent work in this area has considered subgraph detection from a variety of perspectives. Work has been done on detection of specific target subgraphs in random backgrounds [9], with special attention paid in the computer science and statistics

Manuscript received April 23, 2014; revised September 10, 2014 and March 19, 2015; accepted May 01, 2015. Date of publication May 26, 2015; date of current version July 02, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pengfei Xia. This work is sponsored by the Assistant Secretary of Defense for Research & Engineeringunder Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

B. A. Miller is with Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02420 USA (e-mail: bamiller@ll.mit.edu).

M. S. Beard is with Charles Stark Draper Laboratory, Cambridge, MA 02139 USA (e-mail: mbeard@draper.com).

P. J. Wolfe is with the Department of Statistical Science, University College London, London WC1E 6BT U.K. (e-mail: p.wolfe@ucl.ac.uk).

N. T. Bliss is with Arizona State University, Tempe, AZ 85287 USA (e-mail: nadya.bliss@asu.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.2015.2437841

communities to planted cliques [10], [11] and planted clusters [12], [13]. Other work assumes common substructures over the graph, and detects anomalies based on deviations from the "normative pattern" via methods such as minimum description length [14] or analysis of the graph Laplacian [15]. Techniques such as threat propagation [16], [17] and vertex nomination [18] consider a cue vertex as a knowledge prior, giving an initial indication of which vertices are of interest, the objective then being to find the remainder of the subgraph. Community detection in graphs is a widely studied related problem [19], where the communities in the graph are sometimes cast as deviations from a null hypothesis in which the graph has no community structure [20].

The objective of the present contribution is to develop a broadly applicable detection framework for graph-based data. To apply in these varied domains, this framework should be independent of the specific application. We focus specifically on the uncued anomalous subgraph detection problem, where the goal is to detect the presence of a subgraph that is a statistical outlier without a "tip" vertex provided as a cue. As graphs of interest are often extremely large, the framework should have favorable scaling properties as the number of vertices and edges grows. To gain insight into properties that influence subgraph detectability, the framework will ideally have a natural metric for signal and noise power to enable discussion of quantities like signal-to-noise ratio that are intrinsic to signal processing applications.

In this paper, we present a spectral framework to address the uncued subgraph detection problem. This framework is based on a regression-style analysis of residuals in which an observed random graph is compared to its expected value to find outliers. We analyze the graph in the space of the principal eigenvectors of its residuals matrix, which offers two advantages: it allows us to use results from spectral graph theory to elucidate the notion of subgraph detectability, and it works within a linear algebraic framework with which many signal processing researchers are familiar. Within this framework, the spectral norm provides a good metric for signal and noise power, as we demonstrate analytically and empirically. This framework also enables the development of algorithms that work in a low-dimensional space to detect small anomalies, several of which are discussed in this paper.

The remainder of this paper is organized as follows. In Section II, we formally define the subgraph detection problem. Section III provides a brief summary of related work on subgraph detection and graph residuals analysis. Section IV details our proposed subgraph detection framework. In Section V, we outline several algorithms for anomaly detection within the framework. Section VI presents detection results for several simulated datasets, and in Section VII we demonstrate these techniques on real datasets. Finally, in Section VIII, we summarize and discuss open problems and ongoing work.

II. Problem Model

A. Definitions and Notation

In the subgraph detection problem, the observation is a graph $G = (V, E)$. We will denote the sizes of the vertex and edge sets as $N = |V|$ and $M = |E|$, respectively. A subgraph $G_S = (V_S, E_S)$ of G is a graph in which $V_S \subset V$ and $E_S \subset E \cap (V_S \times V_S)$, where the Cartesian product $V \times V$ is the set of all possible edges in a graph with vertex set V. In this paper, we consider graphs whose edges are unweighted and undirected. We will allow the possibility of self-loops, meaning an edge may connect a vertex to itself. Since edges have no weight, two graphs will be combined via their union. The union of two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, is defined as $G_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$.

Working in a spectral framework, we will make use of matrix representations for graphs. The adjacency matrix $A = \{a_{ij}\}$ of G is a binary $N \times N$ matrix. Each row and column is associated with a vertex in V. This implies an arbitrary ordering of the vertices with integers from 1 to N, and we will denote the ith vertex v_i. Then a_{ij} is 1 if there is an edge connecting v_i and v_j, and is 0 otherwise. Similarly, let $A_S = \{s_{ij}\}$ be the adjacency matrix for the signal subgraph. Since we consider undirected graphs, A and A_S are symmetric. Matrix norms will also be used in the discussion of signal and noise power. Unless otherwise noted, the matrix norm will be the spectral norm, i.e., the induced L_2 norm,

$$\|A\| = \max_{\|x\|_2 = 1} \|Ax\|_2, \tag{1}$$

which is equivalent to the absolute value of the largest-magnitude eigenvalue of the matrix.

Our framework is focused on detection of signals within a random background. The analysis presented in this paper is based on the assumption of Bernoulli random graphs, where the probability of an edge between v_i and v_j is a Bernoulli random variable with expected value p_{ij}. Note that the edge probabilities may be different for all pairs of vertices. Since the presence of each edge is a Bernoulli random variable, the expected value of A is given by $P = \{p_{ij}\}$. We refer to P as the probability matrix of the graph.

Another important notion when dealing with graphs is degree. A vertex's degree is the number of edges adjacent to the vertex. The observed degree of vertex v_i will be denoted k_i, and its expected degree is denoted $\mathbb{E}[k_i] = d_i$. Note that $k_i = \sum_{j=1}^{N} a_{ij}$ and $d_i = \sum_{j=1}^{N} p_{ij}$.[1] The vectors of the observed and expected degrees will be denoted k and d, respectively. The volume of the graph, $\mathrm{Vol}(G)$, is the sum of the degrees over all vertices.

B. The Subgraph Detection Problem

In some cases, the observed graph G will consist of only typical background activity. This is the "noise only" scenario. In other cases, most of G exhibits typical behavior, but a small subgraph has an anomalous topology. This is the "signal-plus-noise" scenario. In this case, the noise graph, denoted $G_N = (V_N, E_N)$, and the signal subgraph, $G_S = (V_S, E_S)$, are combined via union.

The objective, given the observation G, is to discriminate between the two scenarios. Formally, we want to resolve the following binary hypothesis test:

$$\begin{cases} \mathcal{H}_0 : & G = G_N \\ \mathcal{H}_1 : & G = G_N \cup G_S. \end{cases} \tag{2}$$

[1] Using this convention, a self-loop only increases a vertex's degree by 1.

Thus, we have the classical signal detection problem: under the null hypothesis $\mathcal{H}_0$, the observation is purely noise, while under the alternative hypothesis $\mathcal{H}_1$, a signal is also present. Here G_N and G_S are both random graphs, with G_N drawn from the noise distribution and G_S drawn from the signal distribution. We will only consider cases in which the vertex set of the signal subgraph is a subset of the vertices in the background, i.e., $V_S \subset V_N = V$.

III. Related Work

While there are many flavors of subgraph detection research, not all of them work under the same assumptions as in this paper. For example, we consider a variety of noise models, which may not have the "normative pattern" required to use techniques based on common subgraphs [14], [15]. Research into anomaly detection in dynamic graphs by Priebe *et al.* [21] uses the history of a node's neighborhood to detect anomalous behavior, but this would not apply in the case of static graphs, which is the focus of this work. As our interest is in uncued techniques, we operate in a different context from the work in [16]–[18]. These methods are complementary to the techniques outlined in this paper, as a set of outlier vertices could be used to seed a cued algorithm and do further exploration.

Previous work has considered optimal detection in the same context we consider in this paper, though in a restricted setting. In [9], the authors consider the detection of a specific foreground embedded (via union) into a large graph in which each possible edge occurs with equal probability (i.e., the random graph model of Erdős and Rényi). In this setting, the likelihood ratio can be written in closed form, as demonstrated by the following theorem.

Theorem 1 (Mifflin et al. [9]): Let G denote the random graph where each possible edge occurs with equal probability p, and let H denote the target graph. The likelihood ratio of an observed graph J is

$$\Lambda_H(J) = \frac{X_H(J)}{\mathbb{E}[X_H(G)]}. \tag{3}$$

Here $X_H(\cdot)$ denotes the number of occurrences of H in the graph. The applicability of this result, therefore, requires a tractable way to count all subgraphs of the observation J that are isomorphic with the target. This is NP-hard in general [22], although there may be feasible methods to accomplish this for certain targets within sparse backgrounds.

While the previous example requires a complicated procedure, detection of random subgraphs embedded into random backgrounds may be an even harder problem. Take, for example, the detection problem where the background and foreground are both Erdős-Rényi, i.e., when the null and alternative hypotheses are given by

$$\begin{cases} \mathcal{H}_0: & \text{each pair of vertices shares an edge with} \\ & \text{probability } p \\ \mathcal{H}_1: & \text{an } N_S\text{-vertex subgraph was embedded whose} \\ & \text{edges were generated with probability } p_S. \end{cases} \tag{4}$$

In this situation, we can derive an optimal detection statistic.

Theorem 2: For an observed graph $G = (V, E)$, let X be a subset of V of size N_S, and $E_X \subset E$ be the set of all edges existing between the vertices in X. The likelihood ratio for resolving the hypothesis test in (4) is given by

$$\binom{N}{N_S}^{-1} \left(\frac{1-\hat{p}}{1-p}\right)^{\binom{N_S}{2}} \sum_{\substack{X \subset V \\ |X| = N_S}} \left[\frac{\hat{p}(1-p)}{p(1-\hat{p})}\right]^{|E_X|}, \tag{5}$$

where $\hat{p} = p + p_S - pp_S$.

A proof of Theorem 2 is provided in Appendix A. Even in this relatively simple scenario, computing the likelihood ratio in (5) requires, at least, knowing how many N_S-vertex induced subgraphs contain each possible number of edges. In [12], it is shown that some computable tests asymptotically achieve the information-theoretic bound for dense backgrounds, but there are no known polynomial-time algorithms that achieve the bound in a sparse graph [13]. For more complicated models, calculating the optimal detection statistic is likely to be even more difficult.

The subgraph detection framework presented in this paper is based on graph residuals analysis. The residuals of a random graph are the difference between the observed graph and its expected value.[2] For a random graph G, we analyze its residuals matrix

$$B := A - \mathbb{E}[A]. \tag{6}$$

In the area of community detection, a widely used quantity to evaluate the quality of separation of a graph into communities is modularity, defined in [20]. The modularity of a partition $C = \{C_1, \ldots, C_n\}$ is defined as

$$Q = \sum_{i=1}^{n} \left(e_{ii} - a_i^2\right), \tag{7}$$

where C_i are disjoint subsets of V covering the entire set, e_{ii} is the proportion of edges entirely within C_i, and a_i is the proportion of edge connections in C_i, i.e.,

$$a_i = \sum_{j=1}^{n} e_{ij}, \tag{8}$$

with e_{ij} denoting half the number of edges between C_i and C_j for $i \neq j$ (half to prevent from counting the edge in both e_{ij} and e_{ji}). Note that a_i^2 is the expected proportion of edges within C_i if the edges were randomly rewired (i.e., the degree of each vertex is preserved, but edges are cut and reconnected at random). Indeed, if the edge proportions are the only thing maintained in the rewiring, the fraction of edges from any community that connect to a vertex in C_i will be a_i. Thus, the proportion of the total edges from C_i to C_j will be $a_i a_j$. Taken as an analysis of deviations from an expected topology, modularity is a residuals-based quantity.

In the community detection literature, numerous algorithms exist to maximize Q for a given number of communities. In [3], an algorithm is proposed by casting modularity maximization as optimization of a vector with respect to a matrix. The modularity

[2] This is distinct, it should be noted, from the notion of residual networks when computing network flow [22].

matrix B is given as the observed minus the expected adjacency matrices, i.e., a matrix of the form in (6). To divide the graph into two partitions in which modularity is maximized, we can solve

$$\hat{s} = \underset{s \in \{-1,1\}^N}{\arg\max} \; s^T \left(A - \frac{1}{\mathrm{Vol}(G)} k k^T \right) s, \tag{9}$$

and declare the vertices corresponding to the positive entries of $\hat{s}$ to be in one community, with the negative entries indicating the other. This technique will optimize Q for a partition into two communities. Since this is a hard problem, it is suggested that the principal eigenvector of

$$B = A - \frac{1}{\mathrm{Vol}(G)} k k^T \tag{10}$$

is computed—thereby relaxing the problem into the real numbers—with the same strategy of discriminating based on the sign of eigenvector components used to divide the graph into communities.

This is an example of a community detection algorithm based on spectral properties of a graph, which have inspired a significant amount of work in the detection of communities [3], [23]–[25] and global anomalies [2], [26], [27]. In this paper, we leverage these same properties within a novel framework for detection of small subgraphs whose behavior is distinct from background activity.

IV. Detection Framework

A. Framework Overview

The subgraph detection framework we propose is based on the analysis of graph residuals, as expressed by (6). We may be given $\mathbb{E}[A]$, or it may be estimated from the observed data. This is similar to analysis of variance in linear regression: We compare the observed data to its expectation, and if the deviations from the expected value are not consistent with variations due to noise, then this may indicate the presence of a signal (in this case an anomalous subgraph).

To reduce the dimensionality of the problem, this framework deals with a graph's spectral properties. Using the principal components of the residuals matrix, we can consider a graph in the linear subspace in which its residuals are largest. For some established models, there is also theory regarding the eigenvalues and eigenvectors of these matrices [28]. This technique is used in community detection, and is similar to models in which each vertex has a position in a latent Euclidean space (see, e.g., [29]). The presence of certain anomalous subgraphs will alter the projection of a graph into this Euclidean residuals space. Working within this space, we can compute test statistics and, from these, resolve the hypothesis test (2). While these will not be optimal detection statistics as in Theorems 1 and 2, this framework can be applied to a wide variety of random graph models, is computationally tractable, and, as we demonstrate in subsequent sections, is quite useful for resolving the subgraph detection problem in a variety of scenarios.

We use the modularity matrix from (9) as our baseline residuals model. This has several advantages. First, the "given expected degree" model has been well-studied, and we know properties of its eigenvalues and eigenvectors [30]. Second, the model's expected value term is low-rank, which allows easy computation of the eigenvectors of B without computing a dense $N \times N$ matrix (as noted in [3] and described in [31]). This makes the model computationally tractable for large graphs where algorithms more expensive than $O(M)$ can be prohibitive. This model also has a simple fitting procedure. The observed degree is, in fact, the maximum likelihood estimate for the expected degree in the version of this model where each possible edge is a Poisson random variable [32]. For small edge probabilities, this is a good approximation for Bernoulli random variables. Finally, this model has demonstrated utility for intercommunity behavior; i.e., the probability of connections between vertices in different communities seems to follow such a model (the reason that observed degree was added as a covariate in [33]).

B. Power Metrics

As mentioned previously, one important aspect of a signal processing framework is a metric for signal and noise power. This provides a quantity that enables an intuitive assessment of the detectability of a signal in a given background. Again, vector signals with Gaussian noise provide an intuitive metric based on vector norms, while such quantities are less clear in the context of random graphs.

There are several intuitive quantities that could be used for signal or noise power in the context of random graphs. One natural choice would be number of edges, or perhaps average degree. It seems intuitive that a signal graph with a large number of edges would be easier to detect, and that greater variance in the number of edges in the background would make this more difficult. A related linear algebraic quantity would be the Frobenius norm of the residuals matrix, i.e., the sum of the squared residuals over all ordered pairs of vertices. This would consider each edge probability separately, emphasizing the presence of less-likely edges.

These metrics, however, have a few shortcomings. In both cases, the signal power measurement will be exactly the same for any subgraph with the same number of edges. Consider two different trees: a path, in which each edge can be traversed while visiting each vertex exactly once; and a star, where one vertex is connected to all others. Both will have $N_S - 1$ edges and a Frobenius norm of $2(N_S - 1)$. The star, however, is much more concentrated on one vertex, and this will cause it to stand out more in the eigenspace (it is also much less likely to occur by chance if edges are randomly placed). The power metric we use should provide an indication of a subgraph's likelihood to stand apart from the background in the eigenspace, since this is the space in which we consider the data.

Working within a spectral framework, the spectral norm defined in (1) provides a natural power metric. Using $\|A - \mathbb{E}[A]\|$ as a metric for noise power and $\|A_S\|$ as a metric for signal power, we can determine the detectability of a subgraph in the principal eigenspace. To see this, we first define a new matrix, $\hat{A} = \{\hat{a}_{ij}\}$, which is the adjacency matrix of $\hat{G} = (V, E_S \setminus E)$, i.e., the edges of the anomaly that do not appear in the background. For deterministic foreground graphs, if s_{ij} is 1, then $\hat{a}_{ij}$ is a random variable whose value is 1 with probability $1 - p_{ij}$ and 0 with probability p_{ij}. For a random Bernoulli foreground, if $\mathbb{E}[s_{ij}] = q_{ij}$, then $\hat{a}_{ij}$ is 1 with probability $q_{ij}(1 - p_{ij})$. Thus, when the subgraph is embedded within vertices where the interaction level is low, $\mathbb{E}[\hat{A}] \approx \mathbb{E}[A_S]$. For convenience, we will also denote a partition of the residuals matrix as

$$B = \begin{bmatrix} B_S & B_{SN} \\ B_{SN}^T & B_N \end{bmatrix}, \tag{11}$$

where the rows and columns have been permuted so that the subgraph vertices are those with the smallest indices. The submatrix B_S is the background residuals within the subgraph vertices, B_{SN} is the residuals occurring between the subgraph and the rest of the graph, and B_N includes only the residuals within the complement of the subgraph vertices.

If the spectral norm of the signal subgraph is sufficiently large with respect to the background power, the subgraph will dominate the principal eigenvector of the residuals matrix. This is captured in the following theorem, a proof of which is provided in Appendix B.

Theorem 3: Let B be the residuals matrix of a graph drawn from an arbitrary Bernoulli graph process, and $\hat{A}$ be the adjacency matrix of the subgraph that does not include edges in the background graph. If u is the unit eigenvector associated with the largest positive eigenvalue of $B + \hat{A}$ (the residuals matrix after embedding), then assuming $\|\hat{A}\| > \|B_N\| + \|B_S\|$, the components of u associated with only the signal vertices, denoted u_S, is bounded below as $\|u_S\|_2^2 \geq 1 - \varepsilon$, where

$$\varepsilon = O\left(\frac{(\|\hat{A}\| - \|B_N\|)(\|B_S\| + \|B_{SN}\|) + \|B_{SN}\|^2}{(\|\hat{A}\| - \|B_N\|)^2 + \|B_{SN}\|^2}\right). \tag{12}$$

Consider the implication of Theorem 3 for a fixed background, when embedding on a fixed subset of vertices. The theorem states that as the difference between the signal power and the power of the noise among the non-signal vertices ($\|\hat{A}\| - \|B_N\|$) becomes much larger than the noise power involving subgraph vertices ($\|B_S\| + \|B_{SN}\|$), the principal eigenvector will become concentrated on the foreground vertices. A few aspects of this theorem confirm intuition from other signal processing areas. First, if there is significant noise activity within the subgraph vertices, then $\|\hat{A}\|$ may be significantly smaller than $\|A_S\|$, and $\|B_S\|$ may be relatively large. This means that a signal placed in strong noise will be difficult to detect, which is always the case in detection problems. Also, the bound in the theorem shows that if a relatively strong subgraph is embedded where there is typically very little activity, and where there is relatively little interaction with the remainder of the graph (i.e., small $\|B_S\|$ and $\|B_{SN}\|$), the subgraph will be much easier to detect. Put in traditional signal processing language, the signal will be much easier to detect when it is less correlated with the noise. Working within this framework, we see the same properties of the interaction between signal and noise that affect detectability in domains like radar and communications.

An empirical example is provided in Fig. 1. In this case, a 4096-vertex Erdős-Rényi graph (see Section VI.A1) is generated, with a 15-vertex subgraph with 90% edge probability embedded. The horizontal axis is $1-\delta_{\min}$, where $\delta_{\min}$ is the expression in (35) in Appendix B. The bound holds for all cases considered, and the empirical results often are an order of magnitude below the maximum for both the higher and lower edge probabilities ($p = 4 \times 10^{-4}$ and $p = 1 \times 10^{-6}$, respectively). Only when a case is considered where there is no background connectivity within the subgraph vertices is the bound approached more closely.

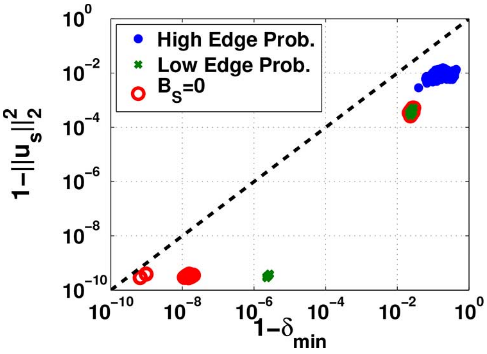

Fig. 1. Empirical comparison to bound in Theorem 3. The bound holds for each case in this scenario with a 4096-vertex random background and a 15-vertex dense signal subgraph, though it is only tight for cases where $B_S = 0$.

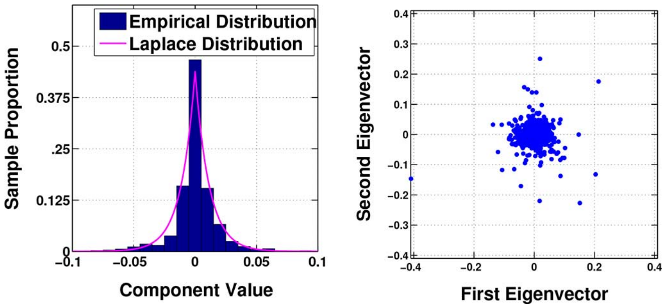

Fig. 2. Distributions of vertex components in principal eigenvectors: a histogram of components in the first eigenvector (left), with a comparison to a Laplace distribution, and a scatterplot (right) in the principal two-dimensional subspace, demonstrating its radial symmetry.

V. Detection Algorithms

For relatively large subgraph anomalies, a simple "energy detector" based on the spectral norm of the residuals matrix will provide good detection performance. It is desirable, however, to be able to detect much smaller subgraphs, which may not stand out in the principal eigenvector. A few techniques have been developed within this framework to detect subtler anomalies [34]–[37], which we outline in this section.

A. Chi-Squared Statistic in Principal Components

The first algorithm is based on the symmetry of the projection of B into its two principal components. This will enable the detection of subgraphs that do not stand out in the first eigenvector. We have empirically observed for several random graph models that, when projecting the residuals into their principal two components, the result is rather radially symmetric. For sparse graphs, the entries in the principal eigenvectors resemble a Laplace distribution, as shown on the left in Fig. 2, which is consistent with behavior observed in sparse Erdős-Rényi graphs. The right-hand plot in Fig. 2 demonstrates the symmetry of the residuals in the top two eigenvectors.

When an anomaly is embedded within the graph, as previously discussed, the subgraph vertices will stand apart from the background. Therefore, we compute a statistic that is based on symmetry in this space to detect the presence of an anomaly. The detection statistic is a chi-squared statistic based on a 2 × 2 contingency table, where the table contains the number of vertices projected into each quadrant of the two-dimensional space. (That is, the number of rows of $[u_1, u_2]$, where u_1 and u_2 are (column) eigenvectors of B, fall into each quadrant.) This yields a 2×2 matrix $O = \{o_{ij}\}$ of the observed numbers of points in each section. From the observation, we compute the expected number of points under the assumption of independence, $\bar{O} = \{\bar{o}_{ij}\}$, where

$$\bar{o}_{ij} = \frac{(o_{i1} + o_{i2})(o_{1j} + o_{2j})}{N}. \tag{13}$$

The chi-squared statistic is then calculated as

$$\chi^2([u_1 \;\; u_2]) = \sum_i \sum_j \frac{(o_{ij} - \bar{o}_{ij})^2}{\bar{o}_{ij}}, \tag{14}$$

and, to favor radial symmetry, we maximize the statistic over rotation in the plane, computing

$$\chi^2_{\max} = \max_\theta \chi^2 \left([u_1 \; u_2] \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}^T \right). \tag{15}$$

The statistic $\chi^2_{\max}$ is used to detect an anomalous subgraph.

When the spectral norm is a reliable detection statistic, thresholding along the principal eigenvalue is often an effective method to identify the vertices that are exhibiting anomalous behavior. Working in multiple dimensions, while it enables the detection of smaller subgraphs, makes the process of identification more complicated. In this setting, we use a method based on k-means clustering to identify the subgraph vertices. Within the two-dimensional space, we compute a small number of clusters and declare the smallest cluster with at least a minimum number of vertices to be the signal subgraph.

B. *Eigenvector L_1 Norms*

It is also desirable to detect signal subgraphs that do not stand out in the principal two components of the residuals matrix, and extending the algorithm of Section V.A to an arbitrary number of dimensions may not be feasible. One method to detect such anomalies relies on the subgraphs being separable in the space of a single eigenvector. As mentioned previously, the entries in the eigenvectors of the background alone resemble numbers drawn from a Laplace distribution. Thus, if a subgraph were to stand out in a single eigenvector, that eigenvector will have a substantially smaller L_1 norm than for the background alone. The L_1 norm of a vector x, $\|x\|_1 = \sum_i |x_i|$, is much smaller when it is concentrated on a small subset of entries, provided that it is unit-normalized in an L_2 sense. For this reason, the L_1 norm serves as a proxy for sparsity in applications such as compressed sensing [38].

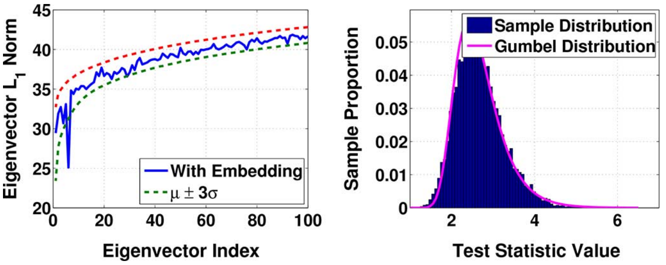

Fig. 3. An example of using eigenvector L_1 norms for subgraph detection. When a small, dense subgraph is embedded into a background with a skewed degree distribution, the L_1 norm of one of the eigenvectors of the residuals matrix becomes much smaller than usual, as shown on the left. Under the null hypothesis, the largest negative deviation from the mean will resemble a Gumbel distribution, plotted on the right.

The following algorithm enables detection when an eigenvector is concentrated on the vertices of the subgraph. This will occur when, for example, a dense subgraph is embedded on relatively low degree vertices, as discussed in Appendix C. We compute the eigenvectors corresponding to the m largest eigenvalues. By measuring cases with no embedding present, we obtain the mean μ_i and standard deviation σ_i for the L_1 norm of the ith eigenvector. For each of the eigenvectors u_i, $1 \le i \le m$, we subtract the mean and normalize by the standard deviation. The smallest (i.e., most negative) value is then used as a test statistic, since we are interested in cases where the norm is small. The test statistic is given by

$$-\min_{1 \le i \le m} \frac{\|u_i\|_1 - \mu_i}{\sigma_i}. \tag{16}$$

An example demonstrating this method is provided in Fig. 3. The example uses a 4096-vertex graph with a skewed degree distribution (using the CL model described in Section VI.A2), with a 15-vertex subgraph with average degree 10.5 randomly embedded into the background. The analysis is run on the 100 eigenvectors associated with the largest positive eigenvalues. While the L_1 norms of most eigenvectors in the resulting matrix fall within three standard deviations of the mean for their index, the L_1 norm of the 6th eigenvector is over 10 standard deviations below the mean, which is extremely unlikely to occur under the null hypothesis. Under the null hypothesis, the test statistic (16) will resemble a Gumbel distribution (commonly used to model extreme values), as shown in the plot on the right. When an embedding occurs that creates a deviation as large as that in the left-hand plot, it will take on a value much larger than the maximum under normal circumstances.

The occurrence of tightly connected subgraphs highly aligned with eigenvectors was documented independently in [39], and a similar anomaly detection method using eigenvector kurtosis in [40]. Here, we use this phenomenon to find subgraphs whose internal connectivity is much larger than the expectation, given the background model. When an anomaly is detected according to (16), the corresponding eigenvector is thresholded to determine the subgraph vertices.

C. *Sparse Principal Component Analysis*

While analysis of eigenvector L_1 norms enables the detection of some subgraphs that do not separate in the principal components of the residuals space, this technique has some shortcomings. In particular, as consecutive eigenvalues get closer together, the direction of the eigenvectors becomes unstable. Therefore, we cannot rely on the test statistic being sufficiently changed because an eigenvector points in the direction of the subgraph.

There is, however, a similar technique that enables the detection of small subgraphs with large residuals. Rather than first computing the eigenvectors of the residuals matrix and then finding an eigenvector with a small L_1 norm, we can find a vector that is *nearly* an eigenvector whose L_1 norm is constrained. This is a technique known as sparse principal component analysis (sparse PCA) [41]. This method has been used in the statistics literature to find high variance in the space of a limited number of variables. We utilize it here for a similar goal: to find large residuals in the space of a small number of vertices.

The problem is formulated as follows. The goal is to find a vector that is projected substantially onto itself by the residuals matrix, but with few nonzero components. Put formally, the objective is to solve

$$\begin{aligned}\hat{x} &= \arg\max_{\|x\|_2=1} x^T B x \\ &\text{subject to } \|x\|_0 \leq N_S,\end{aligned} \tag{17}$$

where $\|\cdot\|_0$ denotes the L_0 quasi-norm (the number of nonzero components in a vector). This, however, is an integer programming problem and is NP-hard. We therefore use a relaxation with an L_1 constraint, recast as a penalized optimization:

$$\hat{x} = \arg\max_{\|x\|_2=1} x^T B x - \lambda \|x\|_1. \tag{18}$$

This problem is still not in an easily solvable form, due to the quadratic equality constraint. We use an additional relaxation, following the method of [41], to achieve a semidefinite program that can be solved using well-documented techniques:

$$\begin{aligned}\hat{X} &= \arg\max_{X \in S_n} \operatorname{tr}(BX) - \lambda \mathbf{1}^T |X| \mathbf{1} \\ &\text{subject to } \operatorname{tr}(X) = 1,\end{aligned} \tag{19}$$

where $\operatorname{tr}(\cdot)$ denotes the matrix trace, $|\cdot|$ replaces each entry in a matrix with its absolute value, and S_n is the set of positive semidefinite matrices in $\mathbb{R}^{n \times n}$. The principal eigenvector of $\hat{X}$, denoted $\hat{x}$, is then returned (and should be sparse, given the constraints). The subgraph detection statistic is $\|\hat{x}\|_1$. If no small subgraph has sufficiently large residuals, the vector should be relatively diffuse and have a relatively large L_1 norm. For vertex identification, the sparse principal component is thresholded, and the vertices corresponding to the components of the vector greater than the threshold are declared to be part of the anomalous subgraph.

One drawback of this technique is its computational complexity. As mentioned in the introduction, one goal of this work is to develop techniques that scale to very large graphs. The algorithms described in Sections V.A and V.B rely on a partial eigendecomposition. Using the Lanczos method for computing m eigenvectors and eigenvalues of a matrix, and leveraging sparseness of the graphs, this requires a running time of $O((Mm + Nm^2 + m^3)h)$, where h is the number of restarts in the algorithm [42]. Thus, if the number of eigenvectors to compute is fixed, this algorithm scales linearly in the number of edges in its per-restart running time. Sparse PCA, as described in [41], has a running time that is $O(N^4\sqrt{\log N}/\epsilon)$, where ϵ controls accuracy of the solution. This implies that sparse PCA will not scale to extremely large datasets without additional optimization, which is a problem for future work. We present results using this technique to demonstrate the feasibility of detecting exceptionally small anomalies using the framework outlined in this paper.

VI. Simulation Results

A. Noise Models

There are many models for random graphs, with varying degrees of complexity. In this section, we outline three random models that will be used for background noise in our experiments.

1) Erdős-Rényi (ER) Random Graphs: The simplest random graph model was proposed by Erdős and Rényi in [43]. In this model, given a vertex set V and a number $p \in (0, 1)$, an edge occurs between any two vertices in V with probability p. In matrix form, $p_{ij} = p$ for all i and j. This model is subsumed by the model for a random graph with a given expected degree sequence assumed by (9), where, in this case, all vertices have the same expected degree.

2) Chung-Lu (CL) Random Graphs: The "given expected degree" model has been studied extensively by Chung and Lu [30]. Similarly to the dynamic preferential attachment model of [44], in this model, the probability of two nodes sharing a connection increases with their popularity. Formally, each vertex v_i is given an expected degree d_i, and the probability of vertices v_i and v_j sharing an edge is given by $p_{ij} = (d_i d_j)/\sum_{\ell=1}^{|V|} d_\ell$, yielding a rank-1 probability matrix

$$P = \frac{1}{\sum_{i=1}^{|V|} d_i} dd^T. \tag{20}$$

Using the observed degree as the expected degree—shown to be an approximately asymptotically unbiased estimator in [45]—the standard formulation of the modularity matrix (10) perfectly fits this model for background behavior.

3) R-MAT Stochastic Kronecker Graphs: To include a slightly more complicated model, we also consider the Recursive Matrix (R-MAT) stochastic Kronecker graph [46]. In this model, a base probability matrix

$$P_b = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{21}$$

is given, where $a, b, c,$ and d are nonnegative values that sum to 1, and edge probabilities are defined by the n-fold Kronecker product of P_b, denoted $\hat{P} = \{\hat{p}_{ij}\} = \bigotimes_{i=1}^{n} P_b$. This results in matrices with 2^n vertices. The graph is generated by an iterative method where one edge is added at each iteration with probabilities defined by $\hat{P}$. If the total number of iterations is t, the edge probabilities are given by

$$p_{ij} = 1 - (1 - \hat{p}_{ij})^t. \tag{22}$$

If the base probability matrix has rank 1, this generator will produce graphs with a similar structure to the CL model. When this is not the case, however, this model creates graphs with mild community structure, as shown in [46], thereby presenting a more challenging noisy background for our subgraph detection framework.

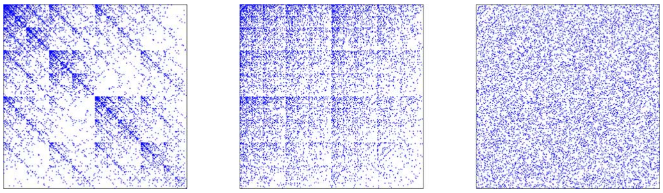

Fig. 4. Sparsity patterns for background graphs: an R-MAT graph (left), a Chung-Lu graph (center), and an Erdős-Rényi graph (right).

These three models represent varying degrees of complexity for the detection framework. The ER model is overspecified by the given expected degree model used in the modularity matrix, the CL model matches the formula exactly, and the R-MAT model is mismatched due to its mild community structure. In the simulations in Section VI.C, the R-MAT graphs are generated using a base probability matrix with $a = 0.5$, $b = c = 0.125$, and $d = 0.25$, and the algorithm is run for $12N$ iterations, resulting in an average degree of approximately 12. The graph is unweighted and its directionality is removed via the "clip-and-flip" procedure as in [46], i.e., the edges below the main diagonal in the adjacency matrix are removed, and those above the main diagonal are made undirected. For CL backgrounds, the expected degree sequence is defined by the edge probabilities of the R-MAT background, i.e., $d_i = \sum_{j=1}^{|V|} p_{ij}$, where p_{ij} is defined in (22). The ER backgrounds use an edge probability that yields an average degree the same as the more complicated models.

Example sparsity patterns of the adjacency matrices, each with 1024 vertices, are shown in Fig. 4. Note the moderate community structure in the R-MAT graph. While the CL graph has vertices of varying degree, it does not have the same structure of the R-MAT. One particularly visible difference is the lack of connections between low-degree vertices and high-degree vertices in the R-MAT graph, seen in the upper-right and lower-left corners of the matrix. Both of these graphs contain more variation than the ER graph, where the uniform randomness can be seen in its sparsity pattern.

B. Signal Subgraph

Two random graph models are used for the anomalous signal subgraph. In one case, an ER graph with probability parameter p_S is generated and combined with randomly selected vertices from the background. Here, the expected adjacency matrix is an $N_S \times N_S$ matrix where every entry is p_S, and thus has spectral norm $p_S N_S$. The second subgraph we consider is a random bipartite graph, where the vertex set is split into two subsets and no edge can occur between vertices in the same subset. Letting N_1 and N_2 be the numbers of vertices in each subset, there are $N_1 N_2$ possible edges between the two vertex subsets, and, as in the ER subgraph case, each of these possible edges is generated with equal probability p_S. For the bipartite subgraph, the expected adjacency matrix has the form

$$\mathbb{E}[A_S] = \begin{bmatrix} \mathbf{0}_{N_1 \times N_1} & p_S \mathbf{1}_{N_1 \times N_2} \\ p_S \mathbf{1}_{N_2 \times N_1} & \mathbf{0}_{N_2 \times N_2} \end{bmatrix}, \tag{23}$$

which has spectral norm $p_S\sqrt{N_1 N_2}$. This subgraph provides us with a signal where the average degree does not equal the spectral norm (unless $N_1 = N_2$), demonstrating that the spectral norm is a more appropriate power metric.

C. Monte Carlo Simulations

The results in this section detail the outcomes of several 10 000-trial Monte Carlo simulations. In each simulation, a background graph is generated and may or may not have a signal subgraph embedded on a subset of its vertices. The subgraph may be a 15-vertex cluster or a bipartite graph with $N_1 = 12$ and $N_2 = 25$. Test statistics outlined in Section V are computed on the resulting graph, creating several empirical distributions that can be used to discriminate between $\mathcal{H}_0$ and $\mathcal{H}_1$. Residuals matrices are formed using either the exact expected value,[3] or a rank-1 approximation based on the observed degrees, as in (9). The expected degree sequence from the R-MAT model is used for CL backgrounds, and ER backgrounds use the same average degree. For R-MAT and CL backgrounds, we consider cases where the foreground vertices are selected uniformly at random from all background vertices, and cases where they are randomly selected from the set of vertices with expected degree at most five.

For 4096-vertex graphs, ER graphs always achieved near-perfect detection performance. Identification and detection performance for CL and R-MAT backgrounds are summarized in Fig. 5. A few phenomena in the results confirm our intuition. First, note that CL backgrounds have extremely similar performance, whether the expected value term is given or estimated. This is because the observed degree is a good estimate for expected degree, and the small embedding has a minimal effect on the expected value term, as shown in Appendix D. (The small but noticeable difference when using a bipartite foreground emphasizes the impact of the number of subgraph vertices.) The R-MAT backgrounds have much more substantial performance differences, due to the model mismatch. In fact, when the true expected value is given, performance is better than with the CL background. This is likely due to the lower variance in the noise, caused by smaller connection probabilities among low-degree vertices. Detection performance improves going from the spectral norm statistic to the chi-squared statistic, and improves further when analyzing the eigenvector L_1 norms. Also, when the subgraph is embedded only on vertices with expected degree at most five, performance significantly increases for L_1 norm analysis, while it degrades for the other statistics (since it is likely to be more orthogonal to the principal eigenvectors). Note also that, for the spectral norm and chi-squared statistics, the bipartite embedding is more detectable than the cluster with the same average degree, since the bipartite foreground has a higher spectral norm. This does not hold for the L_1 norm statistic, since the cluster embedding, while less powerful, is concentrated on a smaller subset of vertices, making it more detectable using this statistic.

One interesting aspect of the L_1 norm technique is its non-monotonic behavior when using the estimated rank-1 expected value. In both detection and identification, performance improves as the subgraphs increase in density up to a certain point, after which performance degrades and then improves again. This is due to clustering of eigenvalues caused by the model mismatch, as shown in Fig. 6. The figure presents a

[3] Due to time and memory constraints, a rank-100 approximation for the R-MAT expected value was used instead of the true probability matrix.

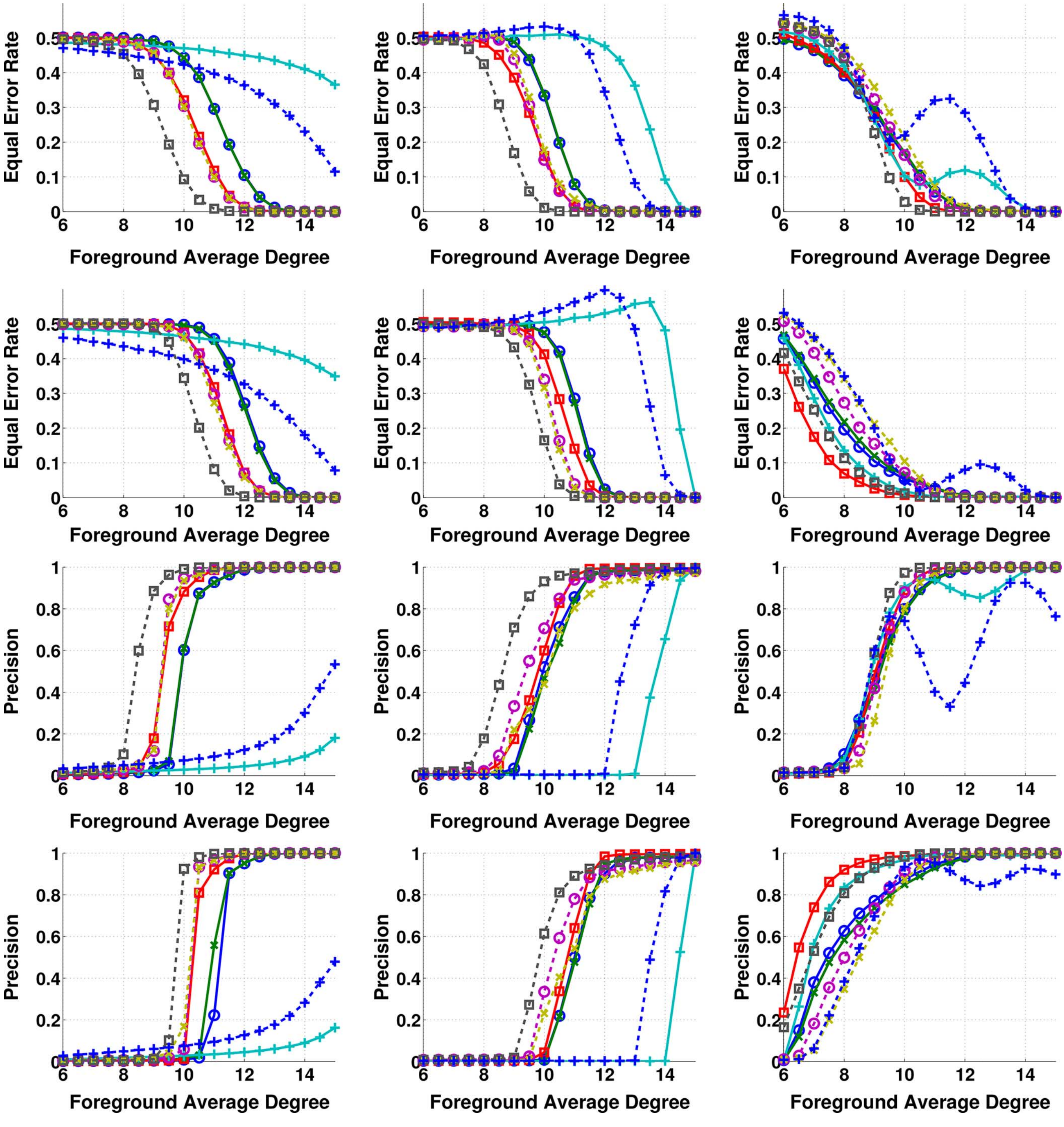

Fig. 5. A summary of detection and identification performance. The equal error rate (EER) for each background and foreground is shown as the average foreground degree increases from 6 to 15. Results are shown for cluster subgraphs (solid line) and bipartite subgraphs (dashed line), for R-MAT graphs with the true expected value (□), R-MAT graphs with an estimated rank-1 expected value $(+)$, CL graphs with given expected degrees (○), and CL graphs using observed degrees $(\times)$. Performance improves as the test statistic goes from the spectral norm (left column), to the chi-squared statistic (center column), to the largest deviation in L_1 norm (right column). Detection performance with the L_1 norm-based statistic improves when the subgraph is embedded on low-degree vertices (second row), rather than choosing the vertices uniformly at random (first row). The same performance trends typically hold for the vertex identification algorithms (uniform random embedding in third row, degree-biased embedding in fourth row), shown here in terms of precision at a 35% recall rate. The non-monotone behavior using L_1 norms is caused by a cluster of larger eigenvalues in the R-MAT background, which, as discussed in Appendix C, makes detection more difficult with this method.

histogram of eigenvalues for the R-MAT graph minus the estimated rank-1 expected value matrix, $\|k\|_1^{-1}kk^T$. (The vertical axis is the average number of eigenvalues that fell into a given bin over the 10 000 Monte Carlo trials.) Most of the eigenvalues are below 12, while there is always 1 that is greater than 16 and 11 in the cluster that spans approximately 12 to 15. Since, as discussed in Appendix C, having eigenvalues that are close together hinders performance with this method, performance improves when the subgraph can be localized in an eigenvector as its eigenvalue approaches the gap at 12, but it will be more difficult once it falls in the cluster of larger values. Using the true expected value instead of the rank-1 approximation does

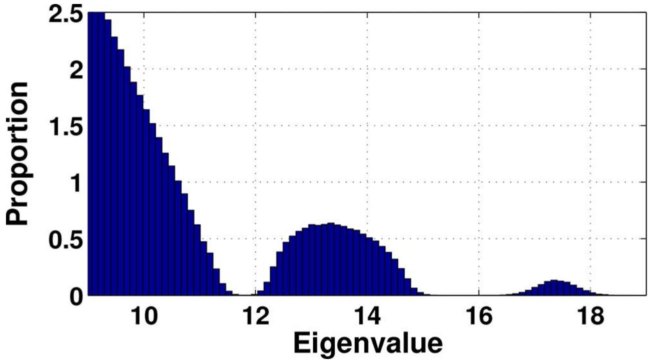

Fig. 6. Histogram of eigenvalues from an R-MAT matrix using an estimated rank-1 expected value. The two clusters of larger eigenvalues are responsible for the non-monotonic behavior in the L_1-norm statistic shown in Fig. 5.

not yield this behavior, since there is no model mismatch. The mismatch between R-MAT and the rank-1 expected value also causes the slight degradation in performance using the chi-squared statistic before it rapidly improves. This may be because the embedded subgraph actually improves the symmetry of the projection by balancing out the mismatch, before finally overpowering it.

The identification results on the bottom half of Fig. 5 follow similar trends, with one notable exception. Performance is shown in terms of precision at a 35% recall rate (precision is emphasized since the foreground vertex set is much smaller than the background). While the k-means-based identification method (center column, using three clusters and a subgraph threshold of five vertices) typically improves performance over thresholding of the principal eigenvector (first column) for cases where precision is relatively low, it actually hinders performance in cases where precision is high. This shows that a subgraph that separates well along the first eigenvector will not necessarily be equally detectable via k-means, possibly due to spreading in the second dimension.

Since sparse PCA has a much greater computational burden, we carried out a more limited set of experiments on smaller graphs. In each trial, a 512-vertex background graph is generated according to either an R-MAT or ER model. The R-MAT graphs use the same probability matrix as in the previous experiment, and the ER graphs have equal expected volume. In each case, we use an estimated rank-1 expected value, and use the DSPCA software package [47] to solve (19). Detection and identification performance are shown in Fig. 7. These results demonstrate the detection of a 7-vertex, 80% dense subgraph in the R-MAT background or a 5-vertex, 85% dense subgraph in an ER background. Sparse PCA yields markedly superior performance to the three methods used in Fig. 5. By using this more costly technique, much smaller, subtler anomalies can be detected, using the same principles as the less expensive algorithms.

VII. Results on Application Data

Two network datasets were downloaded from the Stanford Network Analysis Project (SNAP) large graph dataset collection (available at http://snap.stanford.edu/data). One dataset consists of product co-purchase records on amazon.com, where each of the 548 552 vertices represents a product, and a directed edge from vertex i to vertex j denotes that when product i is purchased, product j is frequently also purchased [48]. The other dataset has 1 696 415 vertices, representing nodes on the Internet, taken from autonomous system traceroutes in 2005 [49]. The edges in this graph are undirected and represent communication links between nodes. In both cases, the 150 eigenvectors corresponding to the largest positive eigenvalues of the residuals matrices were computed, and subgraphs were analyzed that align with eigenvectors with small L_1 norms.

In the amazon.com co-purchase network, edges are directed, and each vertex has at most five outward edges. We use the symmetrized modularity matrix introduced in [50] as a residuals matrix. As shown on the left in Fig. 8, many of the eigenvectors have small L_1 norms, due to frequent co-purchase of small, relatively isolated sets of products. We consider the two smallest L_1 norms, corresponding to the 23rd and 135th largest eigenvectors. These eigenvectors are concentrated, respectively, on a 53-vertex subgraph with the maximum possible number of internal edges (265) and a 44-vertex subgraph with 215 internal

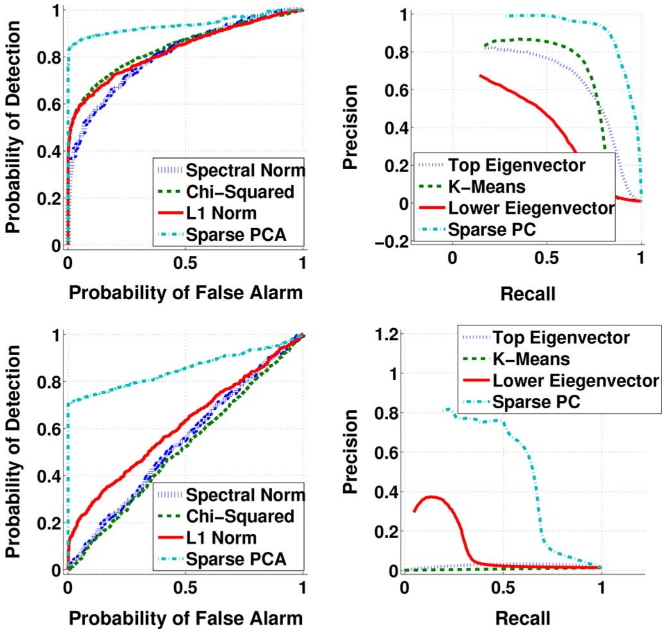

Fig. 7. Detection and identification results using sparse PCA. In both an Erdős-Rényi background (top row) and an R-MAT background (bottom row), sparse PCA significantly outperforms the other algorithms. Similar performance gaps are seen in detection performance (left column) and identification (right column).

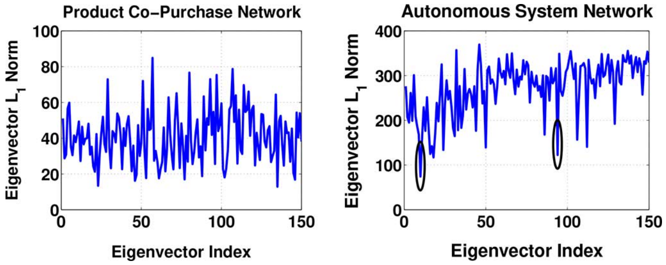

Fig. 8. Eigenvector L_1 norms in application datasets: an amazon.com product co-purchase network (left) and an autonomous system network (right).

edges of a possible 220. Neither subgraph has any outgoing edges, and both have fewer than 20 incoming edges. To compare this to the graph as a whole, we took one million samples of comparable size by performing random walks on the graph. Of all 53-vertex samples, only 609 have average internal degree greater than 4.5, and of those, none has fewer than 20 external edges. Similarly, among the random samples with 44 vertices, 108 have average internal degree greater than 4.4 and fewer than 40 external edges. Each of these 108 samples, however, is primarily outside of the 150-dimensional space spanned by the computed eigenvectors—an indicator vector for the sample vertices in each case is nearly in the null space of the matrix of eigenvectors. Thus, both of these subgraphs are anomalous with respect to random samples of similar size, when considering portions of the graph that are well-represented in the computed subspace.

The eigenvector L_1 norms in the autonomous system graph generally follow a trend, getting larger as the eigenvalues get smaller (indices increasing). The two vectors highlighted in the figure—the 10th and 94th—were considered for further investigation, since they have the largest local deviations. The 10th eigenvector is aligned with a 70-vertex subgraph with over 99% of its possible edges, and the 94th eigenvector is aligned with a 28-vertex subgraph with over 81% of its possible edges. These subgraphs consisted of primarily high-degree vertices, with average external degrees of about 957 and 577 for the 70- and 28-vertex subgraphs, respectively. We took one million random samples from among the vertices with degree greater than 500, with sizes commensurate with the number of high-degree vertices in each subgraph (68 of 70 and 17 of 28). Among the three 68-vertex samples with density greater than 80%, all share at least 55 vertices with the detected subgraph. Of the 17-vertex samples, 713 are at least 75% dense and have fewer than 16 000 external edges (the 17-vertex subset is 93% dense and has about 12 500 external edges). Of these 713 samples, all are significantly aligned with eigenvectors 10 and 18, both of which also have extremely small L_1 norms as shown in the figure. Thus, the only subgraphs among the samples with similar densities and external degrees would be detected through analysis of eigenvector L_1 norms.

VIII. Conclusion

In this paper, we present a spectral framework for the uncued detection of small anomalous signals within large, noisy background graphs. This framework is based on analysis of graph residuals in their principal eigenspace. We propose the spectral norm as a power metric, and several algorithms are outlined, with varying degrees of complexity. In simulation, we demonstrate the utility of the algorithms for detection and identification of two foregrounds within three background models, with the more computationally complex methods providing better detection performance. In two real networks, subgraphs detected via one of the algorithms are shown to be anomalous with respect to random samples of the background.

The framework presented in this paper demonstrates the utility of considering the anomalous subgraph detection problem in a signal processing context. There are myriad avenues of investigation from this point. Recent work has focused on extending this framework to time-varying graphs [51], [52] and attributed graphs [53]. Non-spectral statistics have also been of interest, in particular for detecting anomalously sparse (rather than anomalously dense) subgraphs [54], though this complicates the analysis since embedding the signal involves subtracting edges rather than adding them. Another interesting area is detection using supervised learning based on subgraph features, as in [55]. Performance bounds in spectral detection of cliques and communities have recently been studied [11], [56], as have computational limits of detection [57], [58]. Also, while the presented framework relies on analysis of residuals, considering *normalized* residuals may improve detection for subgraphs where the edges are extremely unlikely [30], [59]. This analysis, however, may be intractable for more complicated graph models, since it requires normalizing each observed vertex pair and may not allow the computational tricks mentioned in Section IV.A. As the detection of anomalous behavior in relational datasets continues to be a problem of interest, the field of signal processing for graphs will continue to pose a rich set of challenges for the research community.

Appendix A
Proof of Theorem 2

Under $\mathcal{H}_0$—the hypothesis that the observed graph was generated by an Erdős-Rényi process—the likelihood of the observed graph is given by

$$\mathcal{L}(G;\mathcal{H}_0,p)=p^{|E|}(1-p)^{\binom{N}{2}-|E|}. \tag{24}$$

Under the alternative hypothesis, an N_S-vertex subset was selected uniformly at random to serve as the subgraph. Suppose that $V_S \subset V$, $|V_S| = N_S$, was chosen as the subset. Each pair of vertices within V_S still has probability p of sharing an edge due to background activity. If there is no edge in the background, however, an edge will be added with probability p_S. Thus, the probability of an edge occurring between a given pair of vertices both in V_S is

$$\hat{p}=p+(1-p)p_S=p+p_s-p\cdot p_S. \tag{25}$$

All other vertex pairs still have probability p of sharing an edge. Therefore, we have

$$\mathcal{L}(G;\mathcal{H}_1,p,V_S,p_S)=\hat{p}^{|E_S|}(1-\hat{p})^{\binom{N_S}{2}-|E_S|} \cdot p^{|E|-|E_S|}(1-p)^{\binom{N}{2}-|E|-\binom{N_S}{2}+|E_S|}. \tag{26}$$

Note that $\binom{N}{2}-|E|-\binom{N_S}{2}+|E_S|$ is the number of "non-edges" that are not within the subgraph vertices. Since only one vertex subset is chosen for the signal embedding, the likelihood of G under the alternative hypothesis is

$$\sum_{V_S\subset V,|V_S|=N_S}\mathcal{L}(G;\mathcal{H}_1,p,V_S,p_S)\Pr[V_S\text{ is chosen}]. \tag{27}$$

Each of the $\binom{N}{N_S}$ possible subsets is equally likely, so the likelihood ratio is

$$\frac{\binom{N}{N_S}^{-1}\sum_{V_S\subset V,|V_S|=N_S}\mathcal{L}(G;\mathcal{H}_1,p,V_S,p_S)}{\mathcal{L}(G;\mathcal{H}_0,p)} \tag{28}$$

or, equivalently,

$$\binom{N}{N_S}^{-1} \sum_{V_S \subset V, |V_S| = N_S} \frac{\mathcal{L}(G; \mathcal{H}_1, p, V_S, p_S)}{\mathcal{L}(G; \mathcal{H}_0, p)}. \tag{29}$$

The ratio in (29) can be further simplified as

$$\frac{\mathcal{L}(G; \mathcal{H}_1, p, V_S, p_S)}{\mathcal{L}(G; \mathcal{H}_0, p)} = \left(\frac{1-\hat{p}}{1-p}\right)^{\binom{N_S}{2}} \left[\frac{\hat{p}(1-p)}{p(1-\hat{p})}\right]^{|E_S|}. \tag{30}$$

Replacing the ratio in (29) with the expression in (30), and moving the non-subgraph-dependent portion outside of the summation, yields the expression in (5). This completes the proof.

Appendix B
Proof of Theorem 3

Let u_1 be the (unit-normalized) principal eigenvector of $\hat{A}$. Since u is the eigenvector corresponding to the largest eigenvalue of $B + \hat{A}$, we have

$$u_1^T (B + \hat{A}) u_1 = \|\hat{A}\| + u_1^T B u_1 \le u^T (B + \hat{A}) u. \tag{31}$$

Since u_1 only has nonzero entries in rows corresponding to subgraph vertices, we can bound this quantity below by $\|\hat{A}\| - \|B_S\|$.

The vector u can be decomposed as $u = u_S + u_B$, where the only nonzero components of u_S correspond to the signal subgraph vertices and u_B may only be nonzero in the rows corresponding to $V \setminus V_S$. Let $\delta = \|u_S\|_2^2$. Since u has unit L_2 norm, and u_S and u_B are orthogonal, we have $0 \le \delta \le 1$ and $\|u_B\|_2^2 = 1 - \delta$. The largest eigenvalue of the residuals matrix is then given by

$$\begin{aligned} u^T (\hat{A} + B) u = {} & u_S^T \hat{A} u_S + 2 u_S^T \hat{A} u_B + u_B^T \hat{A} u_B \\ & + u_S^T B u_S + 2 u_S^T B u_B + u_B^T B u_B. \end{aligned} \tag{32}$$

Both terms that include $\hat{A} u_B$ are zero, since $\hat{A}$ is only nonzero within the subgraph vertices. To get an upper bound for this quantity, we bound each term in (32), yielding

$$\begin{aligned} u^T (\hat{A} + B) u \le {} & \delta \|\hat{A}\| + \delta \|B_S\| \\ & + 2\sqrt{\delta(1-\delta)} \|B_{SN}\| + (1-\delta) \|B_N\|. \end{aligned} \tag{33}$$

For convenience, let $\alpha = \|\hat{A}\| + \|B_S\| - \|B_N\|$, $\beta = 2\|B_{SN}\|$, and $\gamma = \|B_N\| + \|B_S\| - \|\hat{A}\|$. Combining the upper bound in (33) with the lower bound yields

$$\beta \sqrt{\delta(1-\delta)} \ge -(\gamma + \alpha\delta). \tag{34}$$

We can verify that, for $\beta \ge 0$ and $-\alpha \le \gamma < 0$, (34) will achieve equality at the lesser of the two roots of the parabola obtained by squaring both sides of the expression. Therefore, (34) holds whenever

$$\delta \ge \frac{\beta^2 - 2\alpha\gamma - \sqrt{\beta^4 - 4\beta^2 \gamma(\alpha + \gamma)}}{2(\alpha^2 + \beta^2)}. \tag{35}$$

Using the triangle inequality to remove the radical in (35) and substituting the matrix norms back into the equation yields the bound in (12). This completes the proof.

Appendix C
Concentration of Eigenvectors on Subgraph Vertices

Here we provide an example of an embedding on which a single eigenvector will be concentrated. Consider a subgraph $\hat{A}$ that is regular, i.e., each vertex has the same degree d_S. Such a subgraph will have a spectral norm $\|\hat{A}\| = d_S$, and the principal eigenvector will be a vector in which all components on subgraph vertices are equal. Let x be a unit-normalized indicator vector for the subgraph, i.e., a vector where the ith component is $1/\sqrt{N_S}$ if i corresponds to a subgraph vertex and is 0 otherwise. Further consider $x^T (B + \hat{A}) x$ and $x^T (B + \hat{A})^2 x$. We have

$$x^T (B + \hat{A}) x = d_S + x^T B x = d_S + X, \tag{36}$$

where

$$X = \frac{1}{N_S} \sum_{i,j \in V_S} (a_{ij} - p_{ij}) \tag{37}$$

is a random variable whose mean is 0 and variance is less than $\frac{2}{N_S^2} \mathbb{E}[|E \cap (V_S \times V_S)|]$, that is, the expected fraction of possible edges between the subgraph vertices that exist in the background. If the embedding occurs on vertices where the expected connectivity is low, then X will likely be very small. We also have

$$\begin{aligned} x^T (B + \hat{A})^2 x &= d_S^2 + 2 d_S X + x^T B^2 x \\ &= d_S^2 + 2 d_S X + Y. \end{aligned} \tag{38}$$

Note that $Y = \|Bx\|_2^2 = x^T B^2 x$, which can be rewritten as

$$\begin{aligned} x^T B^2 x &= \frac{1}{N_S} \sum_{i=1}^{N} \left[\sum_{j \in V_S} (a_{ij} - p_{ij}) \right]^2 \\ &= \frac{1}{N_S} \sum_{i=1}^{N} \sum_{j,k \in V_S} (a_{ij} a_{ik} - a_{ij} p_{ik} - a_{ik} p_{ij} + p_{ij} p_{ik}). \end{aligned} \tag{39}$$

For $j \ne k$, the expectation of the summand is 0. Considering only $j = k$, we have

$$\mathbb{E}[x^T B^2 x] = \frac{1}{N_S} \sum_{i=1}^{N} \sum_{j \in V_S} p_{ij} (1 - p_{ij}) \tag{40}$$

$$< \frac{1}{N_S} \sum_{i=1}^{N} \sum_{j \in V_S} p_{ij}, \tag{41}$$

where the upper bound is the average expected degree of the subgraph vertices before the embedding occurs. Again, if the subgraph is embedded on vertices with low expected degree, this quantity is likely to be small.

Let $U\Lambda U^T = B+\hat{A}$ be the eigendecomposition of the residuals matrix, with λ_i denoting the ith eigenvector ($\lambda_i \geq \lambda_j$ for $i < j$), and let $z = U^T x$. We have

$$\begin{aligned}(x^T(B+\hat{A})x)^2 &= \left(\sum_{i=1}^{N}\lambda_i z_i^2\right)^2\\ &= (d_S + X)^2\\ &= d_S^2 + 2d_S X + X^2 \qquad (42)\end{aligned}$$

and

$$x^T(B+\hat{A})^2 x = \sum_{i=1}^{N}\lambda_i^2 z_i^2 = d_S^2 + 2d_S X + Y. \qquad (43)$$

If the quantities in (42) and (43) were the same, then x would be an eigenvector of $B+\hat{A}$. Since their difference is very small (i.e., assuming Y and X^2 are small, as they are in expectation), then x may be highly correlated with a single eigenvector. That is, for some i, z_i^2 may be quite large, so that x concentrates most of its magnitude on the ith eigenvector. Let λ_m be the eigenvalue closest to $d_S + X$, and $\delta = d_S + X - \lambda_m$. Then we have

$$\begin{aligned}d_S + X &= \lambda_m + \delta\\ &= \sum_{i=1}^{m-1}\lambda_i z_i^2 + \lambda_m z_m^2 + \sum_{i=m+1}^{N}\lambda_i z_i^2. \qquad (44)\end{aligned}$$

For $i \neq m$, let $\Delta_i = \lambda_i - \lambda_m$. For convenience, define the following substitutions:

$$a = \sum_{i=1}^{m-1} z_i^2 \qquad (45)$$

$$b = z_m^2 \qquad (46)$$

$$c = \sum_{i=m+1}^{N} z_i^2 \qquad (47)$$

$$\varepsilon_1^+ = \frac{\sum_{i=1}^{m-1}\Delta_i z_i^2}{a} \qquad (48)$$

$$\varepsilon_1^- = \frac{\sum_{i=m+1}^{N}\Delta_i z_i^2}{c}. \qquad (49)$$

Thus, $\lambda_m + \varepsilon_1^+$ and $\lambda_m + \varepsilon_1^-$ are convex combinations of the eigenvalues greater than λ_m and less than λ_m, respectively. We can then express (44) as

$$d_S + X = a(\lambda_m + \varepsilon_1^+) + b\lambda_m + c(\lambda_m + \varepsilon_1^-). \qquad (50)$$

Similarly, letting

$$\varepsilon_2^+ = \frac{\sum_{i=1}^{m-1}\Delta_i^2 z_i^2}{a} \qquad (51)$$

and

$$\varepsilon_2^- = \frac{\sum_{i=m+1}^{N}\Delta_i^2 z_i^2}{c}, \qquad (52)$$

(43) can be rewritten as

$$\begin{aligned}d_S^2 + 2Xd_S + Y = a\left(\lambda_m^2 + 2\varepsilon_1^+\lambda_m + \varepsilon_2^+\right) + b\lambda_m^2\\ +c\left(\lambda_m^2 + 2\varepsilon_1^-\lambda_m + \varepsilon_2^-\right). \qquad (53)\end{aligned}$$

Combining (50) and (53) and performing some algebraic manipulation yields the system of equations

$$\delta = a\varepsilon_1^+ + c\varepsilon_1^- \qquad (54)$$

$$\delta^2 + Y - X^2 = a\varepsilon_2^+ + c\varepsilon_2^- \qquad (55)$$

$$1 = a + b + c, \qquad (56)$$

which, solving for b, gives us

$$\begin{aligned}b &= 1 - \frac{(\delta^2 + Y - X^2)(\varepsilon_1^+ - \varepsilon_1^-) - \delta(\varepsilon_2^+ - \varepsilon_2^-)}{\varepsilon_1^+\varepsilon_2^- - \varepsilon_1^-\varepsilon_2^+}\\ &> 1 - \frac{\delta^2 + Y - X^2}{\min(\varepsilon_2^+, \varepsilon_2^-)} - \frac{|\delta|}{\min(\varepsilon_1^+, -\varepsilon_1^-)}. \qquad (57)\end{aligned}$$

If the eigenvalues around λ_m are spread far apart, then ε_1^+, $-\varepsilon_1^-$, ε_2^+, and ε_2^- will be relatively large, the fractions in (57) will be small, and x will be heavily concentrated on a single eigenvector. This is supported by the empirical results in Section VI, where embedding clusters onto vertices with low expected degree yields separation in a single eigenvector.

Appendix D
Change in Modularity Due to Subgraph Embedding

When using observed degree to estimate expected degree, the difference in the expected value terms caused by the signal is as follows. If no embedding occurs, the estimated expected value is $\|k\|_1^{-1}kk^T$, where k is the observed degree vector resulting from the background noise. If an anomalous subgraph is embedded into the background, the degree vector is changed by $\hat{k} = \hat{A}\mathbf{1}$. Since $\hat{A}$ consists of only edges within the subgraph that do not appear due to noise, the degree vector after embedding is $k + \hat{k}$, and the volume is $\|k+\hat{k}\|_1 = \|k\|_1 + \|\hat{k}\|_1$. Thus, the difference between the modularity matrix with estimated expected degrees under $\mathcal{H}_0$ and $\mathcal{H}_1$ is

$$\begin{aligned}\Delta K &= \frac{kk^T}{\|k\|_1} - \frac{(k+\hat{k})(k+\hat{k})^T}{\|k\|_1 + \|\hat{k}\|_1}\\ &= \frac{\|\hat{k}\|_1 kk^T - \|k\|_1(k\hat{k}^T + \hat{k}k^T + \hat{k}\hat{k}^T)}{\|k\|_1(\|k\|_1 + \|\hat{k}\|_1)}. \qquad (58)\end{aligned}$$

To bound the strength of ΔK, we will bound the spectral norm of each summand in the numerator of (58) and ignore the $\|\hat{k}\|_1$ in the denominator, yielding

$$\|\Delta K\| \leq \frac{\|\hat{k}\|_1\|k\|_2^2 + 2\|k\|_1\|k\|_2\|\hat{k}\|_2 + \|k\|_1\|\hat{k}\|_2^2}{\|k\|_1^2}. \qquad (59)$$

To show that the strength of this quantity will grow more slowly than the signal strength, given certain conditions, we will show that $\|\Delta K\|/\|\hat{A}\|$ is $o(1)$, i.e., that

$$\frac{\left(\|\hat{k}\|_1\|k\|_2^2 + 2\|k\|_1\|k\|_2\|\hat{k}\|_2 + \|k\|_1\|\hat{k}\|_2^2\right)}{\|k\|_1^2\|\hat{A}\|} \to 0. \qquad (60)$$

Since $N_S \ll N$, we will ignore the $\|k\|_1\|\hat{k}\|_2^2$ term, as the other terms will dominate it. Thus, we must bound

$$\begin{aligned}\frac{\|\hat{k}\|_1\|k\|_2^2 + 2\|k\|_1\|k\|_2\|\hat{k}\|_2}{\|k\|_1^2\|\hat{A}\|} &= \frac{2\|k\|_2}{\|k\|_1}\cdot\frac{\|\hat{k}\|_2}{\|\hat{A}\|}\\ &\quad + \left(\frac{\|k\|_2}{\|k\|_1}\right)^2\frac{\|\hat{k}\|_1}{\|\hat{A}\|}. \qquad (61)\end{aligned}$$

In many applications, the graphs of interest have degree sequences that follow a power law; i.e., the number of vertices with degree i is approximately $\alpha i^{-\beta}$ for constants $\alpha, \beta > 0$. Using this model, we can analyze the ratio of L_1 and L_2 norms in graphs with a realistic growth pattern. Let $k_{\max}$ be the largest degree in the graph. Then the squares of the L_1 and L_2 norms of k can be approximated as

$$\|k\|_1^2 \approx \left(\sum_{i=1}^{k_{\max}} i \cdot \alpha i^{-\beta} \right)^2 = \left(\sum_{i=1}^{k_{\max}} \alpha i^{1-\beta} \right)^2 \tag{62}$$

and

$$\|k\|_2^2 \approx \sum_{i=1}^{k_{\max}} i^2 \cdot \alpha i^{-\beta} = \sum_{i=1}^{k_{\max}} \alpha i^{2-\beta}, \tag{63}$$

respectively. Their ratio is then approximated, assuming β does not exactly equal 1 or 2, as

$$\begin{aligned} \left(\frac{\|k\|_2}{\|k\|_1} \right)^2 &\approx \frac{\sum_{i=1}^{k_{\max}} \alpha i^{2-\beta}}{\left(\sum_{i=1}^{k_{\max}} \alpha i^{1-\beta} \right)^2} \\ &< \frac{1}{\alpha} \frac{\int_1^{k_{\max}+1} x^{2-\beta} dx}{\left(\int_2^{k_{\max}} x^{1-\beta} dx \right)^2} \\ &= \frac{1}{\alpha} \frac{\frac{1}{3-\beta}[(k_{\max}+1)^{3-\beta} - 1]}{\left(\frac{1}{2-\beta} \left[k_{\max}^{2-\beta} - 2^{2-\beta} \right] \right)^2} \\ &= \frac{(2-\beta)^2}{\alpha(3-\beta)} \frac{(k_{\max}+1)^{3-\beta} - 1}{k_{\max}^{4-2\beta} - 2(2k_{\max})^{2-\beta} + 2^{4-2\beta}}. \end{aligned} \tag{64}$$

In practice, β is typically greater than 1 and less than 3 (see, e.g., [60]), so the constant $(2-\beta)^2/(3-\beta)$ will be positive. As $k_{\max}$ increases, the ratio on the right will tend to $k_{\max}^{\beta-1}$. If we let the maximum degree increase, however, α should be allowed to increase as well, since this controls the number of vertices with a given degree. Assume $k_{\max}$ is a degree that will probably not occur in the graph. Specifically, for a small, constant threshold t, let $k_{\max} = \inf\{i \mid \alpha i^{-\beta} < t\}$. Since this means that

$$\alpha(k_{\max} - 1)^{-\beta} \geq t, \tag{65}$$

we have

$$\frac{1}{\alpha} k_{\max}^{\beta-1} \leq \frac{1}{\alpha} (\sqrt[\beta]{\alpha/t} + 1)^{\beta-1} = O(\alpha^{-1/\beta}). \tag{66}$$

Using the approximation in (64), the ratio of the L_2 and L_1 norms of k is approximately $O(1/\sqrt[2\beta]{\alpha})$.

To bound the term dependent on the subgraph, we have

$$\frac{\|\hat{k}\|_2}{\|\hat{A}\|} = \frac{\sqrt{\mathbf{1}^T \hat{A}^2 \mathbf{1}}}{\|\hat{A}\|} \leq \frac{\sqrt{N_S \|\hat{A}\|^2}}{\|\hat{A}\|} = \sqrt{N_S}. \tag{67}$$

This upper bound can be achieved if the subgraph is a clique or a star. Noting that $\|\hat{k}\|_1 \leq \sqrt{N_S} \|\hat{k}\|_2$, we substitute (66) and (67) into (61) to obtain

$$\frac{\|\Delta K\|}{\|\hat{A}\|} \approx O\left(\sqrt{N_S / \sqrt[\beta]{\alpha}} + N_S / \sqrt[\beta]{\alpha} \right), \tag{68}$$

meaning that $\|\Delta K\|$ is $o(\|\hat{A}\|)$ if N_S is $o(\sqrt[\beta]{\alpha})$. Using (65) as a lower bound for α, this implies that $\|\Delta K\|/\|\hat{A}\|$ will vanish as the graph grows if N_S grows more slowly than $k_{\max}$.

ACKNOWLEDGMENT

The authors would like to thank Dr. B. Johnson and the Lincoln Laboratory Technology Office for supporting this work, and R. Bond, Dr. J. Ward, and D. Martinez for their managerial support. We would also like to thank N. Singh, for his early work on the method in Section V.C. Finally, we would like to thank Dr. R. S. Caceres, Dr. R. J. Crouser, D. Goodwin, Prof. A. O. Hero III, Dr. S. Kelley, Dr. A. Reuther, Dr. M. C. Schmidt, Dr. M. M. Wolf, and the anonymous referees for many helpful comments on this paper.

REFERENCES

[1] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, and H. Lu *et al.*, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[2] T. Idé and H. Kashima, "Eigenspace-based anomaly detection in computer systems," in *Proc. ACM Int. Conf. Knowl. Discov. Data Min.*, 2004, pp. 440–449.

[3] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, no. 3, 2006, Article ID 036104.

[4] K. S. Xu and A. O. Hero, III, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE J. Sel. Topics Signal Process.*, pp. 552–562, Aug. 2014.

[5] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sept. 1999.

[6] K. Chen, C. Huo, Z. Zhou, and H. Lu, "Unsupervised change detection in SAR image using graph cuts," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2008, vol. 3, pp. 1162–1165.

[7] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, pp. 1644–1656, Apr. 2013.

[8] D. I. Shuman, S. K. 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, pp. 83–98, May 2013.

[9] T. L. Mifflin, C. Boner, G. A. Godfrey, and J. Skokan, "A random graph model for terrorist transactions," in *Proc. IEEE Aerosp. Conf.*, 2004, pp. 3258–3264.

[10] N. Alon, M. Krivelevich, and B. Sudakov, "Finding a large hidden clique in a random graph," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, 1998, pp. 594–598.

[11] R. R. Nadakuditi, "On hard limits of eigen-analysis based planted clique detection," in *Proc. IEEE Statist. Signal Process. Workshop*, 2012, pp. 129–132.

[12] E. Arias-Castro and N. Verzelen, "Community detection in random networks," 2013, preprint: arXiv.org:1302.7099 [Online]. Available: http://arxiv.org/abs/1302.7099

[13] N. Verzelen and E. Arias-Castro, "Community detection in sparse random networks," 2013, preprint: arXiv:1308.2955 [Online]. Available: http://arxiv.org/abs/1308.2955

[14] W. Eberle and L. Holder, "Anomaly detection in data represented as graphs," *Intell. Data Anal.*, vol. 11, no. 6, pp. 663–689, Dec. 2007.

[15] D. B. Skillicorn, "Detecting anomalies in graphs," in *Proc. IEEE Intell. Secur. Informatics*, 2007, pp. 209–216.

[16] S. T. Smith, S. Philips, and E. K. Kao, "Harmonic space-time threat propagation for graph detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 3933–3936.

[17] S. T. Smith, E. K. Kao, K. D. Senne, G. Bernstein, and S. Philips, "Bayesian discovery of threat networks," *IEEE Trans. Signal Process.*, vol. 62, pp. 5324–5338, Oct. 2014.

[18] G. A. Coppersmith and C. E. Priebe, "Vertex nomination via content and context," 2012, preprint: arXiv.org:1201.4118v1 [Online]. Available: http://arxiv.org/abs/1201.4118

[19] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, pp. 75–174, Feb. 2010.

[20] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, 2004, Article ID 026113.

[21] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, "Scan statistics on Enron graphs," *Comput. Math. Organiz. Theory*, vol. 11, no. 3, pp. 229–247, 2005.

[22] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 1990.

[23] J. Ruan and W. Zhang, "An efficient spectral algorithm for network community discovery and its applications to biological and social networks," in *Proc. IEEE Int. Conf. Data Min.*, 2007, pp. 643–648.

[24] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proc. SIAM Int. Conf. Data Min.*, 2005, pp. 274–285.

[25] D. Fasino and F. Tudisco, "An algebraic analysis of the graph modularity," *SIAM. J. Matrix Anal. Appl.*, vol. 35, no. 3, pp. 997–1018, 2014.

[26] Q. Ding and E. D. Kolaczyk, "A compressed PCA subspace method for anomaly detection in high-dimensional data," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7419–7433, Nov. 2013.

[27] S. Hirose, K. Yamanishi, T. Nakata, and R. Fujimaki, "Network anomaly detection based on eigen equation compression," in *Proc. ACM Int. Conf. Knowl. Discov. Data Min.*, 2009, pp. 1185–1193.

[28] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: Amer. Math. Soc., 1997.

[29] S. J. Young and E. R. Scheinerman, "Random dot product graph models for social networks," in *Algorithms and Models for the Web-Graph*, ser. Lecture Notes in Computer Science, A. Bonato and F. R. K. Chung, Eds. New York, NY, USA: Springer, 2007, vol. 4863, pp. 138–149.

[30] F. Chung, L. Lu, and V. Vu, "The spectra of random graphs with given expected degrees," in *Proc. Nat. Acad. Sci. USA*, 2003, vol. 100, no. 11, pp. 6313–6318.

[31] B. A. Miller, N. Arcolano, M. S. Beard, J. Kepner, M. C. Schmidt, N. T. Bliss, and P. J. Wolfe, "A scalable signal processing architecture for massive graph analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5329–5332.

[32] P. O. Perry and P. J. Wolfe, "Null models for network data," 2012, preprint: arXiv:1201.5871v1 [Online]. Available: http://arxiv.org/abs/1201.5871

[33] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, "Stochastic blockmodels with a growing number of classes," *Biometrika*, vol. 99, no. 2, pp. 273–284, 2012.

[34] B. A. Miller, N. T. Bliss, and P. J. Wolfe, "Toward signal processing theory for graphs and non-Euclidean data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5414–5417.

[35] B. A. Miller, N. T. Bliss, and P. J. Wolfe, "Subgraph detection using eigenvector L1 norms," in *Proc. Adv. Neural Inf. Process. Syst.*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, vol. 23, pp. 1633–1641.

[36] N. Singh, B. A. Miller, N. T. Bliss, and P. J. Wolfe, "Anomalous subgraph detection via sparse principal component analysis," in *Proc. IEEE Statist. Signal Process. Workshop*, 2011, pp. 485–488.

[37] B. A. Miller, N. T. Bliss, P. J. Wolfe, and M. S. Beard, "Detection theory for graphs," *Lincoln Lab. J.*, vol. 20, no. 1, pp. 10–30, 2013.

[38] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[39] B. A. Prakash, A. Sridharan, M. Seshadri, S. Machiraju, and C. Faloutsos, "EigenSpokes: Surprising patterns and scalable community chipping in large graphs," in *Advances in Knowledge Discovery and Data Mining*, ser. LNCS, M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, Eds. New York, NY, USA: Springer, 2010, vol. 6119, ch. 14, pp. 435–448.

[40] L. Wu, X. Wu, A. Lu, and Z.-H. Zhou, "A spectral approach to detecting subtle anomalies in graphs," *J. Intell. Inf. Syst.*, vol. 41, no. 2, pp. 313–337, 2013.

[41] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, 2007.

[42] R. Lehoucq and D. Sorensen, "Implicitly restarted Lanczos method," in *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Eds. Philadelphia, PA, USA: SIAM, 2000, ch. 4.5.

[43] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[44] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[45] N. Arcolano, K. Ni, B. A. Miller, N. T. Bliss, and P. J. Wolfe, "Moments of parameter estimates for Chung-Lu random graph models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 3961–3964.

[46] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-MAT: A recursive model for graph mining," in *Proc. SIAM Int. Conf. Data Min.*, 2004, pp. 442–446.

[47] R. Luss, A. d'Aspremont, and L. E. Ghaoui, DSPCA: Sparse PCA using semidefinite programming, ver. 0.6, Dec. 2008 [Online]. Available: http://www.di.ens.fr/~aspremon/DSPCA.html

[48] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, pp. 1–39, May 2007.

[49] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shinking diameters and possible explanations," in *Proc. Int. Conf. Knowl. Discov. Data Min.*, 2005, pp. 177–187.

[50] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Phys. Rev. Lett.*, vol. 100, pp. 118703-1–118703-4, Mar. 2008.

[51] B. A. Miller, M. S. Beard, and N. T. Bliss, "Matched filtering for subgraph detection in dynamic networks," in *Proc. IEEE Statist. Signal Process. Workshop*, 2011, pp. 509–512.

[52] B. A. Miller and N. T. Bliss, "Toward matched filter optimization for subgraph detection in dynamic networks," in *Proc. IEEE Statist. Signal Process. Workshop*, 2012, pp. 113–116.

[53] B. A. Miller, N. Arcolano, and N. T. Bliss, "Efficient anomaly detection in dynamic, attributed graphs," in *Proc. IEEE Intell. Secur. Inform.*, 2013, pp. 179–184.

[54] B. A. Miller, L. H. Stephens, and N. T. Bliss, "Goodness-of-fit statistics for anomaly detection in Chung-Lu random graphs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 3265–3268.

[55] S. Pan and X. Zhu, "Graph classification with imbalanced class distributions and noise," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1586–1592.

[56] R. R. Nadakuditi and M. E. J. Newman, "Graph spectra and the detectability of community structure in networks," *Phys. Rev. Lett.*, vol. 108, no. 18, pp. 188701-1–188701-5, 2012.

[57] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Conf. Learn. Theory*, S. Shalev-Shwartz and I. Steinwart, Eds., 2013, vol. 30, pp. 1046–1066, ser. JMLR W&CP.

[58] Y. Chen and J. Xu, "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices," 2014, preprint arXiv:1402.1267 [Online]. Available: http://arxiv.org/abs/1402.1267

[59] R. R. Nadakuditi and M. E. J. Newman, "Spectra of random graphs with arbitrary expected degrees," *Phys. Rev. E*, vol. 87, no. 1, pp. 012803-1–012803-12, 2013.

[60] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," in *Proc. SIGCOMM*, 1999.

Benjamin A. Miller (M'10–SM'15) received the B.S. degree (with highest honors) and the M.S. degree in computer science in 2005 from the University of Illinois at Urbana-Champaign. In 2005, he joined Lincoln Laboratory at the Massachusetts Institute of Technology as an Associate Staff member in the Embedded Digital Systems (later Embedded and High Performance Computing) group. In this role, he developed novel algorithms for real-time linearization of radio-frequency electronics, researched methods and models for signal recovery in multi-sensor compressed sensing, and developed efficient spectral techniques for the detection of anomalies in large graphs. Since 2012, he has been a Technical Staff member at Lincoln Laboratory, currently in the Cyber Analytics and Decision Systems group, where he continues to focus his research on the theoretical and computational aspects of anomaly detection in large networks and other dynamic, combinatorial structures. Mr. Miller is a member of the IEEE Signal Processing Society, the Association for Computing Machinery, and the Society for Industrial and Applied Mathematics. He holds 6 patents and is author or co-author of 36 peer-reviewed conference and journal papers on nonlinear signal processing, compressive sensing, and detection and estimation theory for graph-based data.

Michelle S. Beard is a member of the technical staff in the Product Assurance division at Draper Laboratory. Previously, she was a research staff member in the Computing and Analytics Group at MIT Lincoln Laboratory. She received her bachelor's degree in computer science at Bryn Mawr College in 2010 and is currently pursuing her Master's degree in computer science at Tufts University. Her technical experience includes software engineering and test, visual analytics, and HCI.

Patrick J. Wolfe (S'96–M'03–SM'08) holds chairs in statistics and computer science at University College London (UCL), where his research is focused on statistical theory and methods for network data analysis and signal processing. He received the B.S.E.E. degree from the University of Illinois at Champaign-Urbana in 1998 and the Ph.D. degree from Cambridge University in 2003, after which he joined the faculty of Harvard University, receiving the Presidential Early Career Award from the White House in 2008 for contributions to signal and image processing. In addition to serving as founding Executive Director of the UCL Big Data Institute, he is currently a U.K. Royal Society Research Fellow and an EPSRC Established Career Fellow in the Mathematical Sciences, and leads several major research initiatives in network modeling and inference.

Nadya T. Bliss (M'08–SM'10) received the Bachelor of Science degree in computer science from Cornell University, Ithaca, NY in 2002, Master of Engineering degree in computer science from Cornell University in 2002, and the Ph.D. degree in applied mathematics for the life and social sciences (complex adaptive systems science) from Arizona State University, Tempe, AZ, USA in 2015.

From 2002 to 2012, she was with MIT Lincoln Laboratory, most recently as the Group Leader of the Computing and Analytics Group. Currently, she holds the following appointments at Arizona State University: Director, Global Security Initiative; Professor of Practice, School of Computing, Informatics, and Decision Systems Engineering; and Senior Sustainability Scientist, Julie Ann Wrigley Global Institute of Sustainability. She leads interdisciplinary research teams addressing global challenges in cyber security and digital identity, mitigation and adaptation to climate change, and human security. Her personal research both at MIT Lincoln Laboratory and ASU has been focused on analysis of large networks for wide range of applications.

Dr. Bliss was awarded the inaugural MIT Lincoln Laboratory Early Career Technical Achievement award recognizing her work in parallel computing, computer architectures, and graph processing architectures and her leadership in anomaly detection in graph-based data (presented to 2 employees under 35) in 2011.