A Spectral Framework for Anomalous Subgraph Detection

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

Abstract—A wide variety of application spaces are concerned with data in the form of relationships or connections between entities, which are commonly represented as graphs. Within these diverse areas, a common problem of interest is the detection of a subset of entities that are 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 spectral framework that enables such analysis. Leveraging tools from community detection, we show that this framework has natural metrics for signal and noise power, and propose several algorithms that leverage these properties. Detection and identification performance are presented for a number of signal and noise models, and the trends observed confirm intuition gleaned from other signal processing areas. We demonstrate the utility of the proposed techniques in detecting small, highly anomalous subgraphs in two real datasets.

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, information regarding connections, relationships and interactions between discrete entities is necessary to increase situational awareness.

When working with relational data, it is common to encode the relationships in 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] and influential figures [4].

As a data structure, graphs have been used extensively by signal processing practitioners. Analysis of graphs derived from radio frequency or image data is common, as a graph structure can help discriminate and classify interesting entities (see, e.g., [5], [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]. 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. To this end, our objective 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. As graphs of interest are often extremely large, the framework should have favorable scaling properties as the number of vertices and edges grow. 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 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 works within in 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 these algorithms are discussed in the sequel.

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, outline open problems, and discuss future avenues of investigation.

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 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 $i$th 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.

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 a vertex. The degree of vertex $v_i$ will be denoted $d_i$, and its expected degree is denoted $d_i$. Note that $k = \sum_{j=1}^N a_{ij}$ and $d_i = \sum_{j=1}^N p_{ij}$.

The vector of the observed and expected degrees will be denoted $k$ and $d$, respectively. The volume of the graph, $\text{Vol}(G)$, 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:

$$\begin{align*}
\mathcal{H}_0 : & \quad G = G_N \\
\mathcal{H}_1 : & \quad G = G_N \cup G_S.
\end{align*}$$

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

There are many flavors of subgraph detection research, not all of which work under the same assumptions. Some work is based on the assumption of common substructures, and tries to detect deviations from this “normative pattern” via methods such as minimum description length [3] or analysis of the graph Laplacian [9]. This work, however, does not address the detection of anomalies in the random graph models described in Section V. Our background graphs may not have such a “normative pattern” that occurs over a significant amount of the graph. Research into anomaly detection in dynamic graphs by Priebe et al. [10] uses the history of a node’s neighborhood to detect anomalous behavior, but we seek a framework for anomalous subgraph detection in both static and dynamic graphs, so this is not directly applicable. Recent work has examined subgraph detection through techniques such as threat propagation [11], [12] and vertex nomination [13]. These techniques are given a tip vertex that is part of the subgraph of interest, and the objective is then to determine the other subgraph vertices. This work is complementary to the work outlined in this paper. The framework presented here does not take a cue vertex as part of the given information. A subgraph detected using our proposed framework could be used as tip vertices for a cued algorithm.

Previous work has considered optimal detection in the same context as we consider in this paper, though in a restricted setting. In [14], 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. [14]). Let $G$ denote the random graph where each possible edge occurs with equal probability

---

1Using this convention, a self-loop only increases a vertex’s degree by 1.
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 [15], 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

$$
\begin{align*}
H_0 : & \text{ each pair of vertices shares an edge with probability } p \\
H_1 : & \text{ an } N_S\text{-vertex subgraph was embedded whose edges were generated with probability } p_S.
\end{align*}
$$

In this situation, the following theorem defines the 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 (2) is given by

$$
\frac{N}{N_S} \left( \frac{1 - \hat{p}}{1 - p} \right) \left( \frac{N_S}{N_S} \right) \sum_{X \subset V} \left[ \frac{\hat{p}(1 - p)}{p(1 - p)} \right]^{E_X},
$$

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 (3) requires, at least, knowing how many $N_S$-vertex subgraphs contain each possible number of edges. In [16], 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 [17]. 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. (This is distinct, it should be noted, from the notion of residual networks when computing network flow [15].) For a random graph $G$, we analyze its residuals matrix

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

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 [18]. The modularity of a partition $C = \{C_1, \cdots, C_n\}$ is defined as

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

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 number of edges with at least one end in $C_i$. Note that $a_i^2$ is the expected number 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). Thus, modularity is a residuals-based quantity.

In the community detection literature, numerous algorithms exist to maximize $Q$ for a given number of communities. In [8], an algorithm is proposed by casting modularity maximization as optimization of a vector with respect to a matrix. The modularity matrix $B$ is given as the observed minus the expected adjacency matrices, i.e., a matrix of the form in [4]. To divide the graph into two partitions in which modularity is maximized, we can solve

$$
\hat{s} = \arg \max_{s \in \{0, 1\}^N} s^T \left( A - \frac{1}{\text{Vol}(G)} kk^T \right) s,
$$

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 2 communities. Since this is a hard problem, it is suggested that the principal eigenvector of $B = A - \frac{1}{\text{Vol}(G)} kk^T$ 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 [8], [19]–[21] and global anomalies anomalies [2]. [22], [23]. In the framework described in this paper, we specifically aim to detect small subsets that are distinct from the background activity by exploiting these same properties.

IV. DETECTION FRAMEWORK

A. Framework Overview

The subgraph detection framework we propose is based on the analysis of graph residuals, as expressed by (4). 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 [24]. 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., [25]). 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 (1). 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 [5] 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 [26]. 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 [27]). 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. Estimating the expected degree as simply the observed degree is, in fact, the maximum likelihood estimator for the version of this model where each possible edge is a Poisson random variable [28]. For small edge probabilities, this is a good approximation for Bernoulli random variables. Finally, this model has demonstrated utility for inter-community 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 [29]).

### 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 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 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 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

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

where the rows and columns have been permuted so that the subgraph vertices are all in the upper-lefthand corner. 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 $\|A\| > \|B_N\| + \|B_S\|$, the components of $u$ associated with only the signal vertices, denoted $u_S$, is bounded below as $\|u_S\|_F^2 \geq 1 - \varepsilon$ where

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

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 (i.e., $\|\hat{A}\|_F - \|B_N\|_F$) becomes much larger than the noise power involving subgraph vertices (i.e., $\|B_S\|_F + \|B_{SN}\|_F$), 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 $\|\hat{A}\|_F$ may be significantly smaller than $\|A_S\|_F$, and $\|B_S\|_F$ 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 $\|B_S\|_F$ and $\|B_{SN}\|_F$), 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 (15) 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.

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 [30]–[33], 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 2 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 righthand 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 $2 \times 2$ contingency table, where the table contains the number of vertices projected into 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 $2 \times 2$ 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, $M = \{m_{ij}\}$, where

$$m_{ij} = (o_{11} + o_{22})(o_{1j} + o_{2j})/N.$$ 

The chi-squared statistic is then calculated as

$$\chi^2([u_1 u_2]) = \sum_i \sum_j (o_{ij} - m_{ij})^2/m_{ij},$$

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

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

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 2-dimensional space, we compute $k$ clusters, and declare the smallest cluster with at least a minimum number of vertices to be the subgraph of interest.

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 [34].

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. From measuring cases with no embedding present, we obtain the mean $\mu_i$ and standard deviation $\sigma_i$ for the $L_1$ norm of the $i$th eigenvector. For each of the eigenvectors $u_i$, $1 \leq i \leq m$, 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

$$\min_{1 \leq i \leq m} (\|u_i\|_1 - \mu_i)/\sigma_i. \quad (7)$$

The occurrence of tightly connected subgraphs highly aligned with eigenvectors was documented independently in [35]. 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 the test statistic (7), the corresponding eigenvector is then thresholded to determine the vertices of interest.

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 stability of 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 compute the eigenvectors of the residuals matrix and find 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) [36]. This technique has been used in the statistics literature to find high variance in the space of a limited number of variables. We utilize the method 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, it is desired to solve

$$\hat{x} = \arg \max_{x} x^T B x \quad \|x\|_2 = 1$$

subject to $\|x\|_0 \leq N_S$,

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:

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

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 [36], to achieve a semidefinite program that can be solved using well-documented techniques:

$$\hat{X} = \arg \max_{X \in S_n} \text{tr}(B X) - \lambda \|X\|_1$$

subject to $\text{tr}(X) = 1$, \quad (8)

where $\text{tr}(\cdot)$ denotes the matrix trace 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 [37]. 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 [36], has a running time that is $O(N^4 \sqrt{\log N}/\varepsilon)$, where $\varepsilon$ 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 3 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 [38]. In this model, given a vertex set $V$ and a number $p \in (0, 1)$, an edge occurs between any two vertices $u, v \in V$ with probability $p$. In matrix form, $p_{ij} = p$ for all $i$ and $j$. This model is a subset of the model for a random graph with a given expected degree sequence assumed by equation (5), 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 [20]. Similarly to the dynamic preferential attachment model of [39], 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 \( p_{ij} = (d_i d_j) / \sum_{t=1}^{V} d_t \), yielding a rank-1 probability matrix

\[
P = \frac{1}{\sum_{t=1}^{V} d_t} d d^T.
\]

Using the observed degree as the expected degree—shown to be an approximately asymptotically unbiased estimator in [40]—the standard formulation of the modularity matrix 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 [41]. In this model, a base probability matrix

\[
P_b = \begin{bmatrix}
a & b \\
c & d
\end{bmatrix}
\]

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 \( \hat{P} = \{\hat{p}_{ij}\} = \otimes_{i=1}^{n} P_b \). This results in matrices with \( 2^n \) vertices. The graph is generated by an iterative method where 1 edge is added at each iteration with probabilities defined by \( \hat{P} \). If the total number of iterations is \( t \), the edge probabilities are defined as

\[
p_{ij} = 1 - (1 - \hat{p}_{ij})^t. \tag{9}
\]

If the base probability matrix has rank 1, this graph will have a similar structure to the CL model. As shown in [41], however, this model creates graphs with mild community structure, thereby presenting a more challenging noisy background for our subgraph detection framework.

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 V-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 12\( N \) iterations, resulting in an average degree of approximately 12. The graph is unweighted, and undirected via the “clip-and-flip” procedure as in [41], 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 as in (9). 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. 3. 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 that 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 \( N_x \times N_x \) matrix where every entry is \( p_s \), and thus has spectral norm \( p_s N_x \). 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

\[
\mathbb{E}[A_S] = \begin{bmatrix}
0_{N_1 \times N_1} & p_s 1_{N_1 \times N_2} \\
p_s 1_{N_2 \times N_1} & 0_{N_2 \times N_2}
\end{bmatrix},
\]

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 demonstrate 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 \( H_0 \) and \( H_1 \). Residuals matrices are formed using either the exact expected value, or a rank-1 approximation based on the observed degrees, as in (5). 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 less than 5.

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. 4. 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 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 least 5, 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.

The identification results on the righthand side of the figure 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 (middle row, using $k = 3$ and a subgraph threshold of 5 vertices) typically improves performance over thresholding of the principal eigenvector (top row) 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, perhaps because of 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

---

**Fig. 4.** A summary of detection and identification performance. The equal error rate (EER) for each background and foreground is shown as the average 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 ($\Box$), R-MAT graphs with an estimated rank-1 expected value (+), CL graphs with given expected degrees (○) and CL graphs using observed degrees (×). Performance improves as the test statistic goes from the spectral norm (top row), to the chi-squared statistic (middle row) to the largest deviation in $L_1$ norm (bottom row). Detection performance with the $L_1$ norm-based statistic improves when the subgraph is embedded on low-degree vertices (second column), rather than choosing the vertices uniformly at random (first column). The same performance trends typically hold for the vertex identification algorithms (uniform random embedding in third column, degree-biased embedding in fourth column), shown here in terms of precision at a 35% recall rate.
each case, we use an estimated rank-1 expected value, and use the DSPCA software package \cite{42} to solve \eqref{8}. Detection and identification performance are shown in Fig. 5 \cite{45}. 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. 4 \cite{45}. 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 \cite{43}. 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 (up to 5 directed edges depart from a given vertex) \cite{44}. The other dataset has 1,696,415 vertices, representing nodes on the Internet, taken from autonomous system traceroutes in 2005 \cite{45}. 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 5 outward edges. We use the symmetrized modularity matrix introduced in \cite{46} as a residuals matrix. As shown on the left in Fig. 6 \cite{47} many of the eigenvectors have small $L_1$ norms, due to frequent co-purchase of small, relatively isolated sets of products. We consider the 2 smallest $L_1$ norms, corresponding to the 23rd and 135th largest eigenvectors. These eigenvectors are concentrated, respectively, on a 53-vertex subgraph with all possible internal edges (265) and a 44-vertex subgraph with 215 of its 220 possible internal edges. 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 1 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 1 million random samples from among the subgraph of vertices with degree greater than 500, with sizes commensurate with the number of vertices in the subgraph within the high-degree vertex set (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 less than 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 are detectable through analysis of eigenvector $L_1$ norms.
VIII. Conclusion

In this paper, we present a spectral framework for the 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 [47, 48] and attributed graphs [49]. Non-spectral statistics have also been of interest, in particular for detecting anomalously sparse (rather than anomalously dense) subgraphs [50], though this complicates the analysis since the signal involves subtracting edges rather than adding them. Performance bounds in spectral detection of cliques and communities have recently been studied [51, 52], as well as complexity limits of sparse PCA [53]. Also, while the presented framework relies on analysis of residuals, considering normalized residuals may improve detection for subgraphs where the edges are extremely unlikely [26, 54]. This analysis, however, may be intractable for more complicated graph models, since it requires normalizing each observation and may not allow the computational tricks mentioned in Section V. 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 \(H_0\), the hypothesis that the observed graph was generated by an Erdős–Rényi process, the likelihood of the observed graph is given by

\[
\mathcal{L}(G; H_0, p) = p^{|E|}(1 - p)^{\binom{N}{2} - |E|}.
\]

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

\[
\hat{p} = p + (1 - p)p_S = p + p_S - p \cdot p_S.
\]

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

\[
\mathcal{L}(G; H_1, p, V_S, p_S) = \hat{p}^{|E_S|}(1 - \hat{p})^{\binom{N_S}{2} - |E_S|} \\
\cdot p^{|E| - |E_S|}(1 - p)^{\binom{N}{2} - |E| - \binom{N_S}{2} + |E_S|}.
\]

Note that \(\binom{N}{2} - |E| - \binom{N_S}{2} + |E_S|\) is the number of “non-edges” not within the subgraph. Since only one subgraph is chosen, the likelihood of \(G\) under the alternative hypothesis is

\[
\sum_{V_S \subset V, |V_S| = N_S} \mathcal{L}(G; H_1, p, V_S, p_S) \Pr[|V_S| = N_S].
\]

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

\[
\frac{\binom{N}{N_S}^{-1} \sum_{V_S \subset V, |V_S| = N_S} \mathcal{L}(G; H_1, p, V_S, p_S)}{\mathcal{L}(G; H_0, p)} = \left( \frac{N}{N_S} \right)^{-1} \sum_{V_S \subset V, |V_S| = N_S} \frac{\mathcal{L}(G; H_1, p, V_S, p_S)}{\mathcal{L}(G; H_0, p)}. \tag{10}
\]

The ratio in (10) can be further simplified as

\[
\frac{p^{E_S}|(1 - \hat{p})^{\binom{N_S}{2} - |E_S|} - p^{|E|} - |E_S| |(1 - p)^{\binom{N}{2} - |E| - \binom{N_S}{2} + |E_S|} - |E| |(1 - p)^{\binom{N}{2} - |E|} - |E_S|}| - |E|}{p^{|E|}(1 - p)^{\binom{N_S}{2} - |E_S|} - |E|}
\]

\[
= \left( 1 - \frac{\hat{p}}{1 - p} \right)^{\binom{N_S}{2}} \left( \frac{p(1 - p)}{p(1 - \hat{p})} \right)^{|E_S|}. \tag{11}
\]

Replacing the ratio in (10) with the expression in (11), and moving the non-subgraph-dependent portion outside of the summation, yields the expression in (3). This completes the proof.

APPENDIX B

PROOF OF THEOREM [3]

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

\[
u_1^T (B + \tilde{A}) u_1 = \| \tilde{A} \| + u_1^T Bu_1 \leq u^T (B + \tilde{A}) u.
\]

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

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 \(\delta = \| u_S \|^2\)). Since \(u\) has unit \(L_2\) norm, and \(u_S\) and \(u_B\) are orthogonal, we have \(0 \leq \delta \leq 1\) and \(\| u_B \|^2 = 1 - \delta\). The largest eigenvalue of the residuals matrix is then given by

\[
u_1^T (\tilde{A} + B) u = u_1^T \tilde{A} u_S + 2 u_1^T \tilde{A} u_B + u_1^T B u_B
\]

\[
+ u_S^T B u_S + 2 u_S^T B u_B + u_B^T B u_B. \tag{12}
\]

Both terms that include \(\tilde{A} u_B\) are zero, since \(\tilde{A}\) is only nonzero within the subgraph vertices. To get an upper bound for this quantity, we bound each term in (12), yielding

\[
u_1^T (\tilde{A} + B) u \leq \delta \| \tilde{A} \| + \delta \| B_S \|
\]

\[
+ 2 \sqrt{\delta(1 - \delta)} \| B_S\| + (1 - \delta) \| B_N \|. \tag{13}
\]

For convenience, let \(\alpha = \| \tilde{A} \| + \| B_S \| - \| B_N \|\), \(\beta = 2\| B_{SN}\|\), and \(\gamma = \| B_S \| + \| B_N \| - \| \tilde{A} \|\). Combining the upper bound in (13) with the lower bound yields

\[
\alpha \delta + \beta \sqrt{\delta(1 - \delta)} + \gamma \geq 0 \tag{14}
\]
We can verify that, for $\beta \geq 0$ and $-\alpha \leq \gamma < 0$, the expression of $[14]$ will achieve equality at the lesser of the two roots of the parabola obtained by squaring the expression. Therefore, $[14]$ holds whenever
\[
\delta \geq \frac{\beta^2 - 2\alpha \gamma - \sqrt{\beta^4 - 4\beta^2 \alpha \gamma}}{2(\alpha^2 + \beta^2)}.
\] (15)
Substituting the matrix norms back into the equation yields the bound in (6). 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 are the same. Let $x$ be a unit-normalized indicator vector for the subgraph, i.e., a vector where the $i$th component is $\frac{1}{\sqrt{N_S}}$ if $i$ corresponds to a subgraph vertex and is 0 otherwise. Further consider $x^T (B + \hat{A}) x$ and $x^T (B + A)^2 x$. We have
\[
x^T (B + \hat{A}) x = d_S + x^T B x = d_S + X,
\]
where $X = \frac{1}{N_S} \sum_{i,j \in V_S} (a_{ij} - p_{ij})$ is a random variable whose mean is 0 and variance is less than $\frac{\delta^2}{N_S} (E \cap (V_S \times V_S))$, that is, the fraction of possible edges in the subgraph 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
\[
x^T (B + \hat{A})^2 x = d_S^2 + 2d_S X + x^T B^2 x = d_S^2 + 2d_S X + Y.
\]

Note that $\| B x \|^2 = x^T B^2 x$, which can be rewritten as
\[
x^T B^2 x = \frac{1}{N_S} \sum_{i=1}^{N} \left( \sum_{j \in V_S} (a_{ij} - p_{ij}) \right)^2
= \frac{1}{N_S} \sum_{i=1}^{N} \sum_{j,k \in V_S} (a_{ij} a_{ik} - a_{ij} p_{ik} - a_{ik} p_{ij} + p_{ij} p_{ik}).
\]
For all $i$, the expectation of the summand is $\sum_{j \in V_S} p_{ij} (1 - p_{ij})$, so we have
\[
\mathbb{E} [x^T B^2 x] < \frac{1}{N_S} \sum_{i=1}^{N} \sum_{j \in V_S} p_{ij},
\]
which 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 $U A U^T = B + \hat{A}$ be the eigendecomposition of the residuals matrix, with $\lambda_i$ denoting the $i$th eigenvector ($\lambda_i \geq \lambda_j$ for $i < j$), and let $z = U^T x$. We have
\[
\left( x^T (B + \hat{A}) x \right)^2 = \left( \sum_{i=1}^{N} \lambda_i z_i^2 \right)^2 = (d_S + X)^2
= d_S^2 + 2d_S X + X^2
\]
and
\[
x^T (B + \hat{A})^2 x = \sum_{i=1}^{N} \lambda_i^2 z_i^2 = d_S^2 + 2d_S X + Y.
\] (16)

If these quantities 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 $i$th eigenvector. Let $\lambda_m$ be the eigenvalue closest to $d_S + X$, and $\delta = d_S + X - \lambda_m$. Then we have
\[
d_S + X = \lambda_m + \delta = \sum_{i=1}^{m-1} \lambda_i z_i^2 + \lambda_m z_m^2 + \sum_{i=m+1}^{N} \lambda_i z_i^2.
\] (17)

For $i \neq m$, let $\Delta_i = \lambda_i - \lambda_m$. For convenience, let
\[
a = \sum_{i=1}^{m-1} z_i^2, \quad b = z_m^2, \quad c = \sum_{i=m+1}^{N} z_i^2,
\]
\[
\varepsilon_i^+ = \frac{\sum_{i=1}^{m-1} \lambda_i z_i^2}{a}, \quad \text{and} \quad \varepsilon_i^- = \frac{\sum_{i=m+1}^{N} \lambda_i z_i^2}{c}.
\]

Thus, $\lambda_m + \varepsilon_i^+$ and $\lambda_m + \varepsilon_i^-$ are convex combinations of the eigenvalues greater than $\lambda_m$ and less than $\lambda_m$, respectively. We can then express (17) as
\[
d_S + X = a (\lambda_m + \varepsilon_i^+) + b \lambda_m + c (\lambda_m + \varepsilon_i^-).
\]

Similarly, letting
\[
\varepsilon_2^+ = \frac{\sum_{i=1}^{m-1} \lambda_i z_i^2}{a} \quad \text{and} \quad \varepsilon_2^- = \frac{\sum_{i=m+1}^{N} \lambda_i z_i^2}{c},
\]
(16) can be rewritten as
\[
d_S^2 + 2X d_S + Y = a (\lambda_m^2 + 2 \varepsilon_i^+ \lambda_m + \varepsilon_i^2) + b \lambda_m^2 + c (\lambda_m^2 + 2 \varepsilon_i^- \lambda_m + \varepsilon_i^-).
\] (19)

Combining equations (18) and (19) and performing some algebraic manipulation yields the system of equations
\[
\delta = a \varepsilon_2^+ + c \varepsilon_2^- = a \varepsilon_1 + c \varepsilon_1^- \quad \text{and} \quad \delta^2 + Y - X^2 = a \varepsilon_1^2 + c \varepsilon_2^- 1 = a + b + c,
\]
which, solving for $b$, gives us
\[
b \geq 1 - \frac{(\delta^2 + Y - X^2)(\varepsilon_1^+ - \varepsilon_1^-) + (\varepsilon_1^- \varepsilon_2^- - \delta (\varepsilon_1^2 - \varepsilon_1^-) - (\varepsilon_1^2 - \varepsilon_2^-))^2}{\varepsilon_1^- \varepsilon_2^- - \varepsilon_1^2 \varepsilon_2^-} > 1 - \frac{\delta^2 + Y - X^2}{\min (\varepsilon_1^2, \varepsilon_2^-)}.
\]

If the eigenvalues around $\lambda_m$ are spread far apart, then $\varepsilon_2^+$ and $\varepsilon_2^-$ will be large, and $x$ will be heavily concentrated on a single eigenvector. This is supported by the empirical results in Section I 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 degree vector resulting from the background noise. If an anomalous subgraph is embedded into the background, the degree vector is changed by \( \hat{k} = A1 \). Since \( A \) 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 \( H_0 \) and \( H_1 \) is

\[
\Delta K = \frac{kk^T}{\| k \|_1} - \frac{(k + \hat{k})(k + \hat{k})^T}{\| k \|_1 + \| \hat{k} \|_1} = \frac{\| \hat{k} \|_1 kk^T - \| k \|_1 (k^T k + \hat{k}^T k + \hat{k}^T \hat{k})}{\| k \|_1 (\| k \|_1 + \| \hat{k} \|_1)} \tag{20}
\]

To bound the strength of \( \Delta K \), we will bound the spectral norm of each summand in the numerator of (20) and ignore the \( \| \hat{k} \|_1 \) in the denominator, yielding

\[
\| \Delta K \| \leq \frac{\| \hat{k} \|_1 \| k \|_2^2 + 2\| k \|_1 \| k \|_2 \| \hat{k} \|_2 + \| k \|_1 \| \hat{k} \|_2^2}{\| k \|_1^2} \tag{21}
\]

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

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

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

\[
\frac{\| \hat{k} \|_1 \| k \|_2^2 + 2\| k \|_1 \| k \|_2 \| \hat{k} \|_2 + \| k \|_1 \| \hat{k} \|_2^2}{\| k \|_1^2 \| \hat{A} \|} = \frac{2\| k \|_1 \| k \|_2 \| \hat{k} \|_2 + \| k \|_1 \| \hat{k} \|_2^2}{\| k \|_1 \| \hat{A} \|} \tag{22}
\]

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 \( k_{\text{max}} \) be the largest degree in the graph. Then the squares of the \( L_1 \) and \( L_2 \) norms of \( k \) can be approximated as

\[
\| k \|_1^2 \approx \left( \sum_{i=1}^{k_{\text{max}}} \alpha i^{-\beta} \right)^2 = \left( \sum_{i=1}^{k_{\text{max}}} \alpha i^{-\beta} \right)^2
\]

and

\[
\| k \|_2^2 \approx \sum_{i=1}^{k_{\text{max}}} i^2 \cdot \alpha i^{-\beta} = \sum_{i=1}^{k_{\text{max}}} \alpha i^{2-\beta},
\]

respectively. Their ratio is then approximated, assuming \( \beta \) does not exactly equal 1 or 2, as

\[
\left( \frac{\| k \|_2}{\| k \|_1} \right)^2 \approx \frac{\sum_{i=1}^{k_{\text{max}}} \alpha i^{2-\beta}}{\left( \sum_{i=1}^{k_{\text{max}}} \alpha i^{-\beta} \right)^2} < \frac{1}{\alpha} \int_{1}^{k_{\text{max}}+1} x^{2-\beta} \, dx \frac{\sum_{i=1}^{k_{\text{max}}+1} \alpha i^{-\beta}}{\left( \sum_{i=1}^{k_{\text{max}}+1} \alpha i^{-\beta} \right)^2} = 1 + \frac{1}{\alpha} \left( \frac{\sum_{i=1}^{k_{\text{max}}+1} \alpha i^{-\beta}}{\left( \sum_{i=1}^{k_{\text{max}}+1} \alpha i^{-\beta} \right)^2} \right) \tag{23}
\]

In practice, \( \beta \) is typically greater than 1 and less than 3 (see, e.g., [55]), so the constant \( (2-\beta)^2/(3-\beta) \) will be positive. As \( k_{\text{max}} \) increases, the ratio on the right will tend to \( k_{\text{max}}^{\beta-1} \). If we let the maximum degree increase, however, \( \alpha \) should be allowed to increase as well, since this controls the number of vertices with a given degree. Assume \( k_{\text{max}} \) is a degree that will probably not occur in the graph. Specifically, for a small, constant threshold \( t \), let \( k_{\text{max}} = \inf \{ i | \alpha i^{-\beta} < t \} \). Since this means that \( \alpha(k_{\text{max}} - 1)^{-\beta} \geq t \), we have

\[
\frac{1}{\alpha} k_{\text{max}}^{\beta-1} \leq \int_{t}^{\sqrt[\alpha]{t+1}} x^{\beta-1} \, dx = O \left( \frac{1}{\sqrt[\alpha]{t+1}} \right)
\]

Using the approximation in (21), the ratio of the \( L_2 \) and \( L_1 \) norms of \( k \) is approximately \( O \left( 1/\sqrt[\alpha]{\hat{A}} \right) \). To bound the term dependent on the subgraph, we have

\[
\left( \frac{\| k \|_2}{\| \hat{A} \|} \right)^2 \leq \frac{\| \hat{A} \|^2}{\| \hat{A} \|} = \frac{\sqrt{N_S} \| \hat{A} \|}{\| \hat{A} \|} = \sqrt{N_S} \tag{24}
\]

This upper bound can be achieved if the subgraph is a clique or a star. Noting that \( \| k \|_1 \leq \sqrt{N_S} \| k \|_2 \), we substitute (22) and (24) into the expression to be bounded to obtain

\[
\| \Delta K \| / \| \hat{A} \| \approx O \left( \sqrt{N_S/\sqrt[\alpha]{\hat{A} + N_S/\sqrt[\alpha]{\hat{A}}}} = O \left( N_S/\sqrt[\alpha]{\hat{A}} \right) \right), \tag{25}
\]

meaning that \( \| \Delta K \| / \| \hat{A} \| \) is \( o(\| \hat{A} \|) \) if \( N_S \) is \( o(\sqrt[\alpha]{\hat{A}}) \). As demonstrated in [56], the average degree of a graph tends to increase as a sublinear polynomial with the number of vertices, denoted \( N^\rho, 0 < \rho < 1 \). Since the volume of the graph must grow at least as fast as \( \alpha \), \( \alpha \) must be \( O(N^{1+\rho}) \), meaning that if \( N_S \) is \( O(N^{(1+\rho - \varepsilon)/\beta}) \), then \( \| \Delta K \| / \| \hat{A} \| \) will vanish as the graph grows.

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 VII. Finally, we would like to thank Dr. R. S. Caceres, Dr. R. J. Crouser, Dr. S. Kelley, Dr. A. Reuther, Dr. M. C. Schmidt, and Dr. M. M. Wolf for many helpful comments on this paper.
REFERENCES

[1] D. Bu 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. Int. Conf. Knowledge Discovery and Data Mining, 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.

[4] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” J. ACM, vol. 46, no. 5, pp. 604–632, September 1999.

[5] K. Chen, C. Huo, Z. Zhou, and H. Lu, “Unsupervised change detection in SAR image using graph cuts,” in IEEE Int. Geoscience and Remote Sensing Symp., vol. 3, July 2008, pp. 1162–1165.

[6] A. Krivaneck and M. Sonka, “Ovarian ultrasound image analysis: Follicle segmentation,” IEEE Trans. Med. Imag., vol. 17, December 1998.

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

[8] W. Eberle and L. Holder, “Anomaly detection in data represented as graphs,” Intelligent Data Analysis, vol. 11, no. 6, pp. 663–689, December 2007.

[9] D. B. Skillicorn, “Detecting anomalies in graphs,” in Proc. IEEE Intelligence and Security Informatics, 2007, pp. 209–216.

[10] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, “Scan statistics on Enron graphs,” Computational & Mathematical Organization Theory, vol. 11, no. 3, pp. 229–247, 2005.

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

[12] S. T. Smith et al., “Bayesian discovery of threat networks,” 2013, preprint: arXiv:1311.5552.

[13] G. A. Coppersmith and C. E. Priebe, “Vertex nomination via content and context,” 2012, preprint: arXiv:1201.4118v1.

[14] T. L. Mifflin, C. Boner, G. A. Godfrey, and J. Skokan, “A random graph model for terrorist transactions,” in Proc. IEEE Aerospace Conf., 2004, pp. 3258–3264.

[15] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, Introduction to Algorithms. MIT Press, 1990.

[16] E. Arias-Castro and N. Verzelen, “Community detection in random networks,” 2013, preprint: arXiv:1307.7099.

[17] N. Verzelen and E. Arias-Castro, “Community detection in sparse random networks,” 2013, preprint: arXiv:1308.2955.

[18] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Phys. Rev. E, vol. 69, no. 2, 2004.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman, “The dynamics of blog authorship networks,” in Proc. 2nd Workshop on Link Mining, 2007, pp. 148–156.

[20] 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 and Signal Process., 2010, pp. 5414–5417.

[21] ———, “Subgraph detection using eigenvector L1 norms,” in Advances in Neural Inform. Process. Syst. 23, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 1633–1641.

[22] S. T. Smith, B. A. Miller, T. T. Bliss, and P. J. Wolfe, “Anomalous subgraph detection via sparse principal component analysis,” in Proc. IEEE Statistical Signal Process. Workshop, 2011, pp. 485–488.

[23] B. A. Miller, N. T. Bliss, P. J. Wolfe, and M. S. Beard, “Detection theory for graphs,” Lincoln Laboratory J., vol. 20, no. 1, 2013.

[24] D. Donoho, “Compressed sensing,” IEEE Trans. Inf. Theory, vol. 52, no. 4, pp. 1289–1306, 2006.

[25] B. A. Prakash et al., “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. Radvindran, and V. Pudi, Eds. Springer, 2010, pp. 435–448.

[26] A. d’Aspremont, L. E. Ghaoui, M. J. Jordan, and G. R. G. Lanckriet, “A direct formulation for sparse PCA using semidefinite programming,” SIAM Review, vol. 49, no. 3, pp. 434–448, 2007.

[27] R. Lethoucq and D. Sorensen, “Implicitly restarted Lanczos method,” in Templates for the Solution of Algebraic Eigenvalue Problems: A Prac- tical Guide, Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Eds. Philadelphia: SIAM, 2000, ch. 4.5.

[28] A. Erdős and A. Rényi, “On random graphs,” Publicationes Mathematicae Debrecen, vol. 6, pp. 290–297, 1959.

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

[30] N. Arcolano et al., “Moments of parameter estimates for Chung–Lu random graph models,” in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process., 2012, pp. 3961–3964.

[31] D. Chakrabarti, Y. Zhan, and C. Faloutsos, “R-MAT: A recursive model for graph mining,” in Proc. SIAM Int. Conf. Data Mining, vol. 6, 2004, pp. 442–446.

[32] R. Luss, A. d’Aspremont, and L. E. Ghaoui, “DSPCA: Sparse PCA using semidefinite programming,” December 2008, version 0.6. [Online]. Available: http://www.di.ens.fr/~aspremon/DSPCA.html

[33] J. Leskovec, “Stanford network analysis project,” 2009. [Online]. Available: http://snap.stanford.edu

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

[35] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graphs over time: Den- sification laws, shrinking diameters and possible explanations,” in Proc. Int. Conf. Knowledge Discovery and Data Mining, 2005, pp. 177–187.

[36] E. A. Leicht and M. E. J. Newman, “Community structure in directed networks,” Phys. Rev. Lett., vol. 100, pp. 118703–(1–4), Mar 2008.

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

[38] B. A. Miller and N. T. Bliss, “Toward matched filter optimization for subgraph detection in dynamic networks,” in Proc. IEEE Statistical Signal Process. Workshop, 2012, pp. 113–116.

[39] B. A. Miller, N. Arcolano, and N. T. Bliss, “Efficient anomaly detection in dynamic, attributed graphs,” in Proc. IEEE Intelligence and Security Informatics, 2013, pp. 179–184.

[40] 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 and Signal Process., 2012, pp. 3265–3268.

[41] R. R. Nadakuditi, “On hard limits of eigen-analysis based planted clique detection,” in Proc. IEEE Statistical Signal Process. Workshop, 2012, pp. 129–132.

[42] 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. 188701–1–5, 2012.

[43] Q. Berthet and P. Rigollet, “Complexity theoretic lower bounds for sparse PCA using semidefinite programming,” in Proc. 26th Conference on Learning Theory, 2013, ser. Proceedings of the 26th Conference on Learning Theory, JMLR W&CP, S. Shalev-Shwartz and I. Steinwart, Eds., 2013, vol. 50, pp. 1046–1066.

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

[45] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the Internet topology,” in Proc. SIGCOMM, 1999.

[46] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, 2007.