TOWARD SIGNAL PROCESSING THEORY FOR GRAPHS AND NON-EUCLIDEAN DATA

Benjamin A. Miller and Nadya T. Bliss

Lincoln Laboratory
Massachusetts Institute of Technology
Lexington, Massachusetts 02420
Email: {bamiller, nt}@ll.mit.edu

Patrick J. Wolfe

Statistics and Information Sciences Laboratory
Harvard University
Cambridge, Massachusetts 02138
Email: wolfe@stat.harvard.edu

ABSTRACT

Graphs are canonical examples of high-dimensional non-Euclidean data sets, and are emerging as a common data structure in many fields. While there are many algorithms to analyze such data, a signal processing theory for evaluating these techniques akin to detection and estimation in the classical Euclidean setting remains to be developed. In this paper we show the conceptual advantages gained by formulating graph analysis problems in a signal processing framework by way of a practical example: detection of a subgraph embedded in a background graph. We describe an approach based on detection theory and provide empirical results indicating that the test statistic proposed has reasonable power to detect dense subgraphs in large random graphs.

***Index Terms*—** Chi-squared test, community detection, graph algorithms, high-dimensional data, signal detection theory

1. INTRODUCTION

A graph $G = (V, E)$ is defined as a set of vertices V and a set of edges E, where each edge connects two vertices. In essence, there is a number of entities (the vertices) with relationships defined between them. Due to their ubiquitous structure, graphs are used in a wide variety of application domains, including the natural sciences, medicine and social network analysis. In biology, graphs have been used to represent interactions between proteins [1, 2] and reproduction within a population in an evolutionary model [3]. Social network analysis, where the data of interest are people and the relationships among them, is another very natural setting for graph processing. Significant work has been done on the detection of communities [4, 5] and influential figures [6] in social networks, frequently using a graph as the primary data structure.

The graph has been an important data structure for the signal processing community. Analysis of graphs derived from radio frequency or image data is common [7, 8], as a graph structure can help discriminate and classify interesting entities. In this context, however, the graphs are typically derived from Euclidean data.

In general graphs are non-Euclidean, which complicates the application of standard signal processing to graph problems. Still, it is natural to seek a framework in which graph processing algorithms can be studied and evaluated in much the same way as classical signal processing methods. Some effort has been made to define signal processing techniques for graphs [9], but this has focused primarily on smoothing (i.e., low-pass filtering) of geometric meshes represented by graphs, and no general theory exists. At a high level, many graph problems can be cast in a signal processing context. For example, the problem of finding a specific subgraph in a larger graph [10] seems naturally coupled with matched filtering for signal detection, and other problems such as detecting a very dense subgraph [11], a frequently-occurring subgraph [12] or a certain behavioral pattern [13] all have a strong signal processing flavor to them.

As an example, there has been a substantial amount of work in the area of anomaly detection in graphs. An algorithm is presented in [14] for finding anomalies that bridge highly connected subgraphs. In [15], the authors use measures of entropy on a graph to determine whether a given subgraph is anomolous. The authors of [16] propose and evaluate several algorithms for detecting anomolous occurrences in graph-based data, using metrics that are common in signal processing. Since graphs are often used for detection of anomolous occurrences or behavior, such problems could be presented in the context of classical detection theory. Indeed, [16–18] present detection problems using graphical data and evaluate their techniques with metrics common in signal processing, such as receiver operating characteristic (ROC) analysis.

These problems all have a similar underlying structure: Given a graph G, we want to find $G_S \subset G$ such that G_S is anomolous, dense or equal to some template. Each problem resembles a classical detection problem, but due to the non-Euclidean nature of graphical data (e.g., the lack of well-defined vector operations), the same theoretical frameworks no longer exist. In most domains, there is some natural ordering of the data, such as a time series or a frequency spectrum. A matched filter, for example, assumes a certain temporal ordering of the data and thus operates in geometric, rather than exponential, time. While the non-Euclidean nature of graphs may prevent anything quite so simple from being applied (since in general subgraph detection is an NP-hard problem), an important eventual goal is to provide analogous theoretical structure.

In this paper we demonstrate the practical and conceptual advantages to be gained by formulating graph analysis problems in a signal processing framework with a concrete example: the detection of a subgraph embedded in a large background graph. In our problem formulation and empirical results, we demonstrate that presenting such problems in the framework of classical signal processing not only provides a basis for algorithm comparison, but also enables mature ideas in signal processing to be directly applied to the new and growing field of graph analysis. The remainder of the paper is organized as follows. Section 2 presents a subgraph detection algorithm including a definition of a test statistic, the noise and signal models, and ROC analysis. In Section 3 we analyze the performance

This work is sponsored by the Department of the Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

of the detection algorithm using different background models and signal models of varying density. Section 4 summarizes the results and highlights future directions.

2. DETECING ANOMOLOUS SUBGRAPHS

As an example of developing and applying signal processing theory for graph data, we focus in the sequel on the problem of detecting an anomalous subgraph in a random graph. To formulate the problem of subgraph detectability in the framework of classical detection, we consider the background "noise" graph G_B to be random, and the "signal" graph G_S to be fixed. Akin to a classical hypothesis testing scenario in a vector space, we may then define a set of null and alternate hypotheses as follows:

$$\begin{cases} H_0: & \text{The observed graph is "noise" } G_B \\ H_1: & \text{The observed graph is "signal+noise" } G_B \cup G_S. \end{cases} \quad (1)$$

We present the algorithm for detecting the presence of a subgraph, define noise and signal models, and analyze its performance.

2.1. Test Statistic

To formulate our detection problem, we consider the spectral decomposition of the modularity matrix described in [4]. The modularity matrix B of an unweighted, undirected graph G is defined as

$$B = A - kk^T/2\,|E|\,,$$

where A is the adjacency matrix of G and k is a column vector whose ith row contains the degree of vertex i. Essentially, it is a matrix of the difference between the actual and expected number of edges between pairs of vertices. Since G is undirected, B will be symmetric and thus will have a spectral decomposition

$$B = U\Lambda U^T$$

that has orthogonal eigenvectors corresponding to distinct eigenvalues, all of which will be real. We will consider B in the space of its two principal eigenvectors—u_1 and u_2, both unit vectors—to examine the statistics of these background models in a low-dimensional space.

Our Chi-squared test statistic is calcuated using a 2×2 contingency table. Considering the two principal eigenvectors as points in a plane, we determine how many of these two-dimensional points (i.e., those defined by the rows of $[u_1\ u_2]$) fall into each quadrant. This yields a 2×2 observation matrix $O = \{o_{ij}\}$, which is then used to compute the expected number of points in each quadrant, resulting in the matrix $M = \{m_{ij}\}$, where

$$m_{ij} = (o_{i1} + o_{i2})(o_{1j} + o_{2j})/\,|V|\,.$$

We then compute the deviation $X = \{x_{ij}\}$ from the expected value as

$$x_{ij} = \frac{(o_{ij} - m_{ij})^2}{m_{ij}}.$$

The resulting test statistic $\chi^2([u_1\ u_2]^T) = \sum_i \sum_j x_{ij}$ is then maximized with respect to rotation in the plane, i.e., for each graph we compute

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

We determine the presence of a "signal" graph by comparing $\chi^2_{\max}$ to a threshold.

2.2. Distribution of the Test Statistic: Noise Model

In signal processing, noise is typically modeled as a stochastic process where the distribution may or may not be i.i.d. For our "background" graph models, i.e., the *noise* in the system, we consider two random graph models. In addition to the random graphs defined by Erdös and Rényi [19], which are reminiscent of an i.i.d. process since each edge occurs with equal independent probability, recent work has focused on alternative models that exhibit phenomena frequently seen in real-world graphs, such as power-law degree distributions. As an example of the latter type, we consider here the R-MAT graph model [20], which uses a recursion on Kronecker products to formulate edge probabilities that can yield heavy-tailed degree distributions. We will thus use both Erdös-Rényi (E-R) and R-MAT as canonical models for our background graph, G_B.

The distribution of test statistics for 10000 1024-vertex graphs generated using the R-MAT method is shown in Fig. 1(a). A Gamma probability density function with shape parameter 2, which appears to be a good fit for the distribution, has been fit to the test statistics and overlayed on the histogram. The distribution of test statistics for 10000 E-R graphs is shown in Fig. 1(b). Again, the distribution resembles the overlayed Gamma process.

2.3. Distribution of the Test Statistic: Signal Model

The "foreground" (or *signal*) model in our graph signal processing problem is our subgraph of interest, G_S. In signal processing, the weaker the signal is, or the more it resembles the background, the more difficult it is to process, i.e., when the signal-to-noise ratio (SNR) is low, signals are harder to detect, estimate, and classify. Our intuition tells us that a similar property exists in a graphical setting, and our initial investigation has confirmed this.

Given a foreground G_S, we create a "signal + noise" model using the union operation on the two edge sets, i.e., $G = G_S \cup G_B$ with $G = (V, E),\ E = E_B \cup E_S$. We are interested in detecting subgraphs that are highly interconnected, where detectability seems intuitively apparent given the foreground's anomolous structure in a sparse, random background. In a 1024-vertex graph, we choose 12 vertices at random to comprise V_S and and select a substantial fraction of the $\binom{12}{2} = 66$ possible edges to use as E_S. After creating G, we perform the same statistical analysis on its modularity matrix as we did with G_B. Fig. 1(c) demonstrates the markedly different distribution in the test statistics (again for 10000 randomly-generated graphs) when there is a highly-connected embedding, in this case containing all possible edges. Using an E-R background, where such a dense subgraph is much more unlikely since the model is less structured, we observe even greater separation of the test statistics between G_B and G, as shown in Fig. 1(d).

3. DETECTION PERFORMANCE

A Monte Carlo simulation was run in which we evaluated the performance of the detection algorithm. For each case, the background graph G_B has 1024 vertices and an average degree of approximately 12. The signal graph G_S, as mentioned in Section 2.3, has 12 vertices and we evaluated detection performance as the subgraph density was increased from 70% to 100% in increments of 5%. We used both E-R and R-MAT backgrounds in this experiment. For each combination of background model and signal density, we generated 10000 different background graphs, each time choosing 12 different vertices at random to comprise the signal graph. Based on the desired subgraph density, we then randomly selected the edges

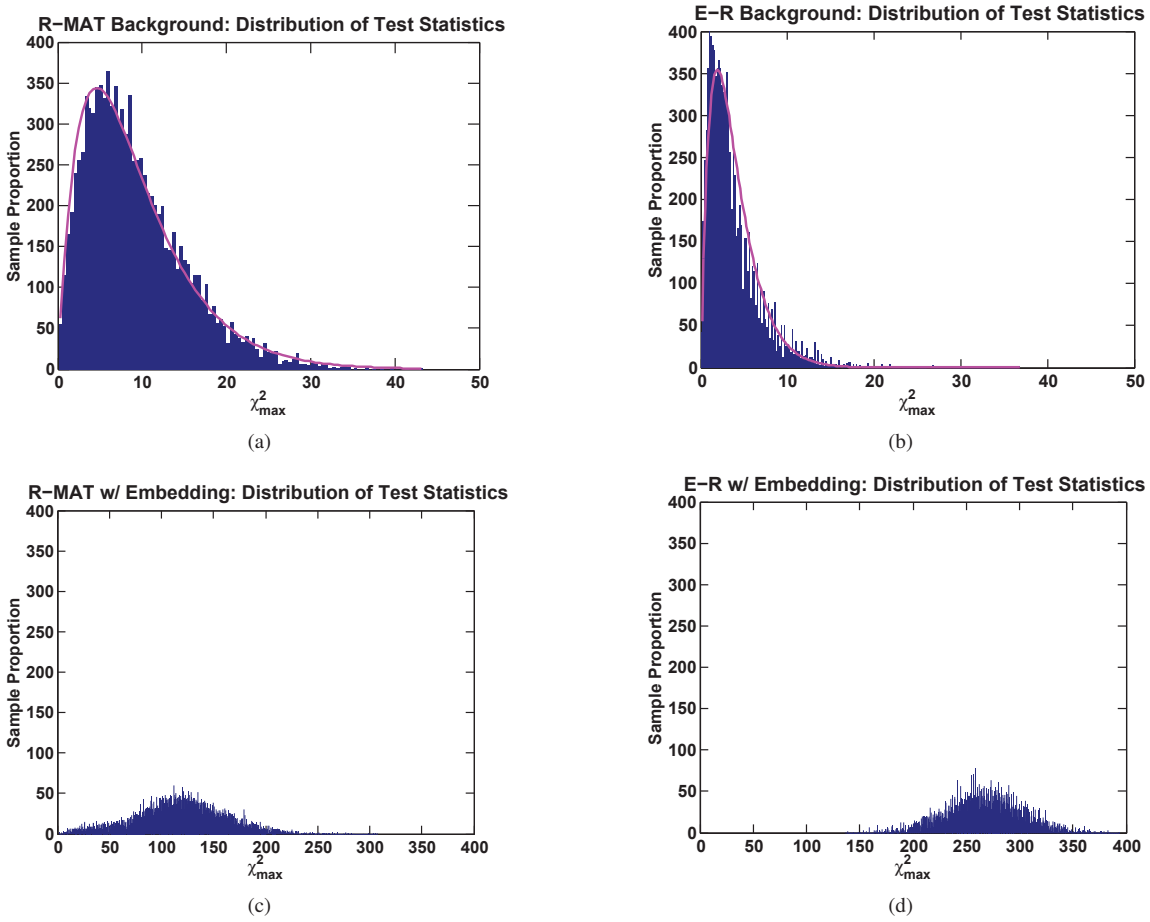

Fig. 1. Distribution of the test statistic $\chi^2_{\max}$ of equation (2) shown under the null and alternate models (top and bottom rows, respectively), for cases of E-R and R-MAT background graphs (left and right columns, respectively). Note that the horizontal axes are scaled differently for the null and alternate models.

for the signal model, E_S, such that $|E_S|$ is the subgraph density times $\binom{12}{2}$.

After creating the 10000 background and signal models, we computed the test statistic from Section 2.1 for each case. This resulted in distributions similar to those in Fig. 1, with clearer separation between the null and alternate hypotheses as the signal graphs get more dense. Considering a range of thresholds to declare a detection, we demonstrate with ROC curves the detectability of subgraphs with varying density using this algorithm.

As demonstrated in Fig. 2(a), when the 12-vertex subgraph has only 70% of all possible edges, it is undetectable. By increasing the number of edges in the subgraph, detection performance increases until our subgraph becomes a 12-vertex clique, where we achieve near-perfect detection performance. The distributions of test statistics for both the "noise" and "signal+noise" for the case of a complete subgraph are displayed on the same plot in Fig. 2(b). The two distributions are highly seperable, with an equal-error rate of 2.12% achieved by setting the threshold to 24.716. When using an E-R background with average degree of 12, perfect detection is achieved for false alarm probabilities greater than zero for all subgraph densities of 70% and higher. While we used backgrounds with the same average degree, the more structured R-MAT model creates a background in which clustering is less anomolous, and detection is more difficult.

4. SUMMARY

In this article we have demonstrated some of the practical and conceptual advantages to be gained by formulating the graph processing problem of subgraph detection in a classical signal processing framework. This formulation provides not only a basis for the performance comparison of various algorithms, by way of comparative ROC curve analysis, but also a means of relating new data types and problem domains to the more mature setting of signal processing in linear vector spaces.

In the case at hand, we provided empirical results indicating that the test statistic we propose has reasonable power to detect dense subgraphs in large random graphs. More broadly, this problem scenario can be viewed as a proxy for more general tasks involving high-dimensional data sets that patently do not conform to the classical Euclidean setting. We are encouraged by the initial successes documented in this article, and hope they will similarly encourage others to join in the challenge of developing a more general signal processing theory for graphs and other non-Euclidean data.

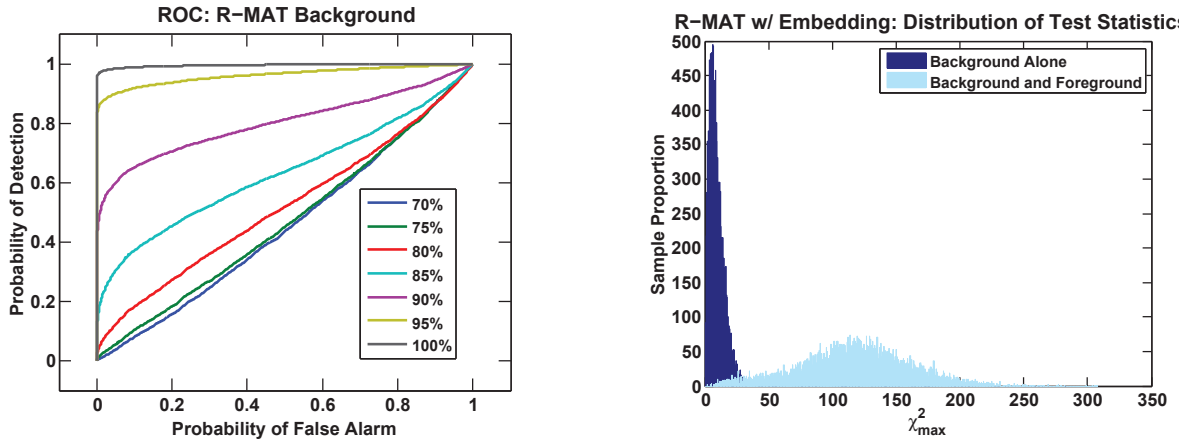

Fig. 2. Operating characteristics of the subgraph detection test, shown for various subgraph densities (left), with empirical sampling distributions of the test statistic shown for the case of a 12-vertex clique (right).

5. REFERENCES

[1] Dongbo Bu, Yi Zhao, Lun Cai, Hong Xue, Xiaopeng Zhu, Hongchao Lu, Jingfen Zhang, Shiwei Sun, Lunjiang Ling, Nan Zhang, Guojie Li, and Runsheng Chen, "Topological structure analysis of the protein–protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[2] Nizar N. Batada, Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz, Laurence D. Hurst, and Mike Tyers, "Stratus not altocumulus: A new view of the yeast protein interaction network," *PLoS Biology*, vol. 4, no. 10, pp. 1720–1731, 2006.

[3] Erez Lieberman, Christoph Hauert, and Martin A. Nowak, "Evolutionary dynamics on graphs," *Nature*, , no. 433, pp. 312–316, 20 January 2005.

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

[5] Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu, "Community detection in large-scale social networks," in *Int'l Conf. on Knowledge Discovery and Data Mining*, 2007, pp. 16–25.

[6] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, September 1999.

[7] Keming Chen, Chunlei Huo, Zhixin Zhou, and Hanqing Lu, "Unsupervised change detection in SAR image using graph cuts," in *IEEE Int'l Geoscience and Remote Sensing Symposium*, July 2008, vol. 3, pp. 1162–1165.

[8] Anthony Krivanek and Milan Sonka, "Ovarian ultrasound image analysis: Follicle segmentation," *IEEE Trans. on Medical Imaging*, vol. 17, no. 6, 6 December 1998.

[9] Gabriel Taubin, Tong Zhang, and Gene H. Golub, "Optimal surface smoothing as filter design," in *Proc. European Conference on Computer Vision*, 1996, pp. 283–292.

[10] Boaz Gelbord, "Graphical techniques in intrusion detection systems," in *Proc. Int'l Conf. on Information Networking*, 2001, pp. 253–258.

[11] Yuichi Asahiro, Refael Hassin, and Kazuo Iwama, "Complexity of finding dense subgraphs," *Discrete Applied Mathematics*, vol. 121, no. 1–3, pp. 15–26, 2002.

[12] Mukund Deshpande, Michihiro Kuramochi, Nikil Wale, and George Karypis, "Frequent substructure-based approaches for classifying chemical compounds," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 8, pp. 1036–1050, August 2005.

[13] Thayne R. Coffman and Sherry E. Marcus, "Pattern classification in social network analysis: A case study," in *Proc. IEEE Aerospace Conf.*, 2004, pp. 3162–3175.

[14] Jimeng Sun, Juiming Qu, Deepayan Chakrabarti, and Christos Faloutsos, "Neighborhood formation and anomaly detection in bipartite graphs," in *Proc. IEEE Int'l. Conf. on Data Mining*, Nov. 2005.

[15] Caleb C. Noble and Diane J. Cook, "Graph-based anomaly detection," in *Proc. ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, 2003, pp. 631–636.

[16] William Eberle and Lawrence Holder, "Anomaly detection in data represented as graphs," *Intelligent Data Analysis*, vol. 11, no. 6, pp. 663–689, December 2007.

[17] Tung Le and Christoforos N. Hadjicostis, "Graphical inference for multiple intrusion detection," *IEEE Trans. on Information Forensics and Security*, vol. 3, no. 3, pp. 370–380, September 2008.

[18] Hsun-Hsien Chang, José M. F. Moura, Yijen L. Wu, and Chien Ho, "Early detection of rejection in cardiac MRI: A spectral graph approach," in *Proc. 3rd IEEE Int'l Symp. on Biomedical Imaging*, April 2006, pp. 113–116.

[19] Paul Erdös and Alfréd Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[20] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos, "R-MAT: A recursive model for graph mining," in *Proc. Fourth SIAM Int'l Conference on Data Mining*, 2004, vol. 6, pp. 442–446.