GOODNESS-OF-FIT STATISTICS FOR ANOMALY DETECTION IN CHUNG–LU RANDOM GRAPHS

Benjamin A. Miller[1]*, Lauren H. Stephens*[2] *and Nadya T. Bliss*[1]

[1]MIT Lincoln Laboratory, Lexington, MA, 02420
{bamiller, nt}@ll.mit.edu
[2]Massachusetts Institute of Technology, Cambridge, MA, 02139
lhs@mit.edu

ABSTRACT

Anomaly detection in graphs is a relevant problem in numerous applications. When determining whether an observation is anomalous with respect to the model of typical behavior, the notion of "goodness of fit" is important. This notion, however, is not well-understood in the context of graph data. In this paper, we propose three goodness-of-fit statistics for Chung–Lu random graphs, and analyze their efficacy in discriminating graphs generated by the Chung–Lu model from those with anomalous topologies. In the results of a Monte Carlo simulation, we see that the most powerful statistic for anomaly detection depends on the type of anomaly, suggesting that a hybrid statistic would be the most powerful.

***Index Terms*—** Graph theory, signal detection theory, anomaly detection, goodness of fit, probabilistic models

1. INTRODUCTION

A graph $G = (V, E)$ is defined as a set of vertices (V) that are interconnected by a set of edges (E). Graphs are useful in many applications in which relationships (edges) between entities (vertices) are of interest. As they are used in a wide variety of disciplines, the problem of anomaly detection in graphs has gained significant interest in the past several years (see, e.g., [1]). Detecting anomalies in a graph's topology is relevant to a host of applications, such as the detection of strange or malicious behavior in computer or social networks.

Recent work has focused on developing a statistical detection framework for graphs, akin to that for Euclidean data [2, 3]. The central tool of this framework is the modularity matrix [4]. The modularity matrix B of an unweighted, undirected graph G is defined as

$$B = A - \frac{kk^T}{2|E|}.$$

Here A is the adjacency matrix of G, where the entry in the ith row and jth column of A is 1 if $\{v_i, v_j\} \in E$ and is 0 otherwise; and k is the degree vector of G, with the ith entry in k being the number of edges adjacent to vertex i. In a random graph in which the probability of an edge between two vertices is proportional to the degrees of the associated vertices, $kk^T/(2|E|)$ is the expected value of A. Thus, we consider the modularity matrix a graph "residuals" matrix, representing the difference between the observed and expected adjacency matrices.

This interpretation broaches an important topic in signal detection: goodness of fit. Given an observation, it is useful to understand how well the data fit the assumed model of "normal" behavior. While this notion is well understood for linear models, it is not at all mature in the context of graphs. The purpose of this paper is to investigate the use of goodness-of-fit statistics in a specific random graph model to determine whether or not an observed graph was generated by that model, i.e., to reject the hypothesis that the graph was generated under the model if the data do not properly fit.

The remainder of this paper is organized as follows. In Section 2 we describe our problem model, including our null model, the Chung–Lu random graph; and several anomalous alternatives. Section 3 introduces the goodness-of-fit statistics we use to test our observations. Section 4 presents empirical results, demonstrating detection performance for each statistic paired with each alternative model. We discuss interesting phenomena observed in the results in Section 5, and in Section 6 we summarize and outline future work.

2. PROBLEM MODEL

2.1. The Chung–Lu Model

In this work, we focus on determining whether or not an observed graph is generated by the Chung–Lu random graph model [5]. Under this model, an unweighted, undirected graph G is created according to the following process. Each vertex $v \in V$ is given an expected degree d_v. The probability that an edge occurs between two vertices $v, u \in V$ is equal to

$$\frac{d_v d_u}{\sum_{x \in V} d_x},$$

that is, it is proportional to the product of the vertices' expected degrees. (It is easily verified that the probabilities of edges adjacent to v sum to d_v.) We allow self-loops (edges from a vertex to itself), as it simplifies our analysis, although an extension to the case without these edges is possible. This model requires that the largest expected degree of any vertex be at most the square root of the sum of all expected degrees, so that no probability is greater than 1.

In an alternative formulation, we assign each vertex v a weight $w_v \in [0, 1]$. The two definitions are equivalent, with the change in parameters given by

$$w_v = \frac{d_v}{\sqrt{\sum_{u \in V} d_u}} \text{ and } d_v = w_v \sum_{u \in V} w_u.$$

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.

In matrix form, let $w \in [0,1]^{|V|}$ be a vector of weights. Then the expected value of the adjacency matrix A of a graph generated by the Chung–Lu model with weights w is given by $E[A] = P = ww^T$. We will refer to P as the probability matrix for the graph, which is equal to the expected value since each edge is the result of a Bernoulli trial with a 0-or-1 outcome.

Given this probability structure, certain subgraphs or topologies are unlikely to occur. We now discuss the alternative models that we wish to reject as being non-Chung–Lu.

2.2. Alternative Models

Two alternatives start with a Chung–Lu background graph, and connect it to an anomalous subgraph, i.e., a subgraph that violates the assumptions of the Chung–Lu model. In both cases, the probability matrix of the graph has the form

$$P = \begin{pmatrix} p_{in}\mathbf{1}_{N_s}\mathbf{1}_{N_s}^T & p_{out}\mathbf{1}_{N_s}\mathbf{1}_{N_b}^T \\ p_{out}\mathbf{1}_{N_b}\mathbf{1}_{N_s}^T & w_b w_b^T \end{pmatrix}.$$

Here $\mathbf{1}_N$ is a column vector of N ones, N_b is the number of background vertices, N_s is the number of anomalous subgraph vertices, and $w_b \in [0,1]^{N_b}$ is the vector of weights for the Chung–Lu background. The values for p_{in} and p_{out} are, respectively, the probability of an edge between two subgraph vertices and the probability of an edge between a subgraph vertex and a background vertex. Note that the expected degree of a vertex in the subgraph is

$$p_{in}N_s + p_{out}N_b, \tag{1}$$

and the expected degree of vertex i in the background is

$$p_{out}N_s + w_b(i)\|w_b\|_1, \tag{2}$$

where $w_b(i)$ is the ith entry in w_b.

Manipulating these values, we can create a topology that is unlikely to occur under a Chung–Lu model. If we let p_{in} be large and p_{out} be small, then the subgraph will be a tightly-connected cluster with little connectivity to the background. Conversely, if we let p_{out} be large and p_{in} be small, then the subgraph will consist of high-degree vertices that are unlikely to be connected to each other. Both of these phenomena are anomalous under the Chung–Lu model. Indeed, considering an extreme case in which $p_{in} = 1$ and $p_{out} = 0$, the subgraph vertices will have degree N_s, but will never be connected to any background vertices, regardless of their expected degree. At the opposite extreme, we may set $p_{in} = 0$ and adjust p_{out} so that, under the Chung–Lu model, the probability of an edge between subgraph vertices is arbitrarily close to 1. We will refer to the high p_{in}, low p_{out} case as a *cluster* anomaly, and the high p_{out}, low p_{in} case as a *hubs* anomaly.

We are also interested in cases where the graph is anomalous throughout its topology, and for this purpose we use the R-MAT Kronecker graph [6]. An R-MAT graph is generated by an iterative procedure where at each iteration, an edge is selected according to a probability matrix defined by the n-fold Kronecker product of a 2×2 probability matrix. This procedure continues for a fixed number of iterations or until the graph has a certain number of edges. (Since we are dealing with unweighted, undirected graphs, we do not increase the weight if an edge is chosen multiple times, and we use the "clip-and-flip" procedure from [6] to undirect the graph.) The probability matrix for this alternative does not have the rank-1 structure of a Chung–Lu graph, and this should create a topology unlikely under the Chung–Lu model.

3. TEST STATISTICS

As discussed in [7], while goodness of fit is an important concept in statistical modeling, this concept is underdeveloped in the context of graphs. In this section we propose 3 goodness-of-fit statistics for Chung–Lu graphs that we will use in our experiments.

3.1. Spectral Norm (SN)

Our first test statistic measures how far the observed graph is from its expected value. We use the spectral norm of the difference between the adjacency matrix and its expected value, i.e.,

$$\|A - E[A]\| = \|A - ww^T\|.$$

This is the maximum eigenvalue (in terms of absolute value) of the matrix consisting of the observed minus expected edges in the graph. If the observed degree vector k is equal to the expected degree vector d, then this is the same as the maximum eigenvalue of the graph's modularity matrix.

3.2. Least Squares Coefficient (LSC)

Another metric for the difference between the observed and expected graph is the least squares coefficient, that is, the coefficient that optimally fits the observed graph to the expectation in a least squares sense. This statistic is expressed as

$$\arg\min_{\gamma} \|A - \gamma ww^T\|_F,$$

with the optimal value given by

$$\gamma_{min} = \frac{w^T A w}{\|w\|_2^4}.$$

One convenient property of this statistic is that it is relatively easy to analyze. It is not difficult to show that, under the Chung–Lu model, the expected value of γ_{min} is 1, and its variance is

$$\mathrm{Var}(\gamma_{min}) = \frac{2\|w\|_3^6 - 2\|w\|_4^8 - \|w\|_6^6 + \|w\|_8^8}{\|w\|_2^8}. \tag{3}$$

3.3. Minimum Neighborhood Likelihood (MNL)

The final statistic is derived from the probability that a given graph G would be created by a Chung-Lu model with a weight vector w. The likelihood of an observed graph under the Chung–Lu model is expressed as

$$\prod_{i=1}^{|V|}\prod_{j=i}^{|V|}((A_{ij}w_iw_j) + (1 - A_{ij})(1 - w_iw_j))),$$

where A_{ij} is the entry in the ith row and jth column of the adjacency matrix. For the types of non-Chung–Lu behavior we examined, however, the graph likelihood was ineffective at distinguishing between the Chung-Lu and alternative models. We refined this statistic in an attempt to improve its detection power. Rather than the likelihood of the entire graph, we use the least likely 1-hop vertex neighborhood. The minimum neighborhood likelihood for an observed graph is

$$\min_i \prod_{j=1}^{|V|} \left((A_{ij}w_iw_j) + (1 - A_{ij})(1 - w_iw_j)\right).$$

	True Weights						Estimated Weights					
Alternative	SN		LSC		MNL		SN		LSC		MNL	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
R-MAT-1	0.561	0.538	**0.555**	**0.541**	0.483	0.489	0.558	0.537	**0.785**	**0.713**	0.483	0.487
R-MAT-2	0.758	0.688	**0.981**	**0.931**	0.730	0.668	0.716	0.657	**1.000**	**1.000**	0.724	0.659
R-MAT-3	**0.998**	**0.983**	0.547	0.532	0.440	0.447	**0.999**	**0.985**	0.957	0.886	0.426	0.448
Cluster	**0.981**	**0.943**	0.509	0.506	0.496	0.497	**0.977**	**0.933**	0.597	0.568	0.495	0.497
Hubs	0.685	0.635	**0.882**	**0.799**	0.711	0.656	0.656	0.614	**0.999**	**0.987**	0.717	0.660

Table 1. Equal-error rate and area under the curve performance for the 3 test statistics, using given and estimated weights. For each alternative model, the most powerful statistic in terms of EER is highlighted.

4. EXPERIMENTAL RESULTS

In each of the following experiments, we ran a 10,000-trial Monte Carlo simulation under the null model and one of the alternatives, and computed our test statistics for each graph. For the cluster and hub alternatives, we first generate a 1024-vertex R-MAT graph with average degree 10, and use the degree vector of this graph to compute the background weights w_b, yielding a background with a powerlaw degree distribution. For the cluster anomaly, we use 8 vertices with $p_{in} = 0.9375$ and $p_{out} = 2/(8 \cdot 1024)$, resulting in a subgraph with an average internal degree of 7.5 and 2 edges, in expectation, between the background and the subgraph. For the hubs anomaly, we add 5 vertices with $p_{in} = 1/25$ and $p_{out} = 0.065$, yielding expected external degrees of over 66 and 0.6 expected internal edges. In both cases, we compare the resulting alternative models to a Chung–Lu graph whose expected degree vector is computed by equations (1) and (2). We are, thus, comparing the test statistics of two random graphs with the *same* expected degree vector.

For the R-MAT alternatives, we use 3 base probability matrices,

$$p_1 = \begin{pmatrix} 0.3 & 0.238 \\ 0.238 & 0.224 \end{pmatrix}, \; p_2 = \begin{pmatrix} 0.2 & 0.2 \\ 0.2 & 0.4 \end{pmatrix},$$

$$\text{and } p_3 = \begin{pmatrix} 0.35 & 0.1625 \\ 0.1625 & 0.325 \end{pmatrix},$$

and we will refer to the models resulting from these as R-MAT-1, R-MAT-2 and R-MAT-3, respectively. In each case, we used the 10-fold Kronecker product to define the edge probability matrix (resulting in a 1024-vertex graph), and ran the algorithm for 10240 iterations. Letting $\hat{p}_{ij}$ be the proabability of an edge between vertex i and vertex j being added in a single iteration, the probability of an edge occurring between these vertices before the algorithm terminates is

$$p_{ij} = 1 - (1 - \hat{p}_{ij})^{10240}.$$

For the Chung–Lu graphs we compare to the R-MAT graphs, we use the expected degree vector given by $d_i = \sum_j p_{ij}$, again making the expected degree vectors the same for the null and alternative models.

Since, in an anomaly detection problem, we may not have access to the true model parameters, we evaluate performance with both given and estimated weights. For an estimated weight vector, we use the simple, closed-form estimator

$$\hat{w} = k/\sqrt{2|E|},$$

i.e., we substitute the observed degree for the expected degree.

Receiver operating characteristic (ROC) curves for these simulations are shown in Fig. 1, with performance summarized in terms of equal-error rate (EER) and area under the curve (AUC) in Table 1. Using the R-MAT-1 alternative, which has the most balanced probability matrix, detection performance is little better than chance except using LSC with *estimated* weights. For R-MAT-2, which is much more biased toward one corner of the probability matrix, all statistics have better-than-chance detection performance, with LSC yielding near-perfect detection with estimated and given weights (again, better performance with estimated weights). One interesting observation is that while MNL is somewhat effective at discriminating Chung–Lu from R-MAT graphs in this case, the minimum likelihood is actually *higher* (i.e., less unlikely) under the alternative (we account for this in Table 1). For MNL and SN, performance with given and estimated weights is similar. For R-MAT-3, in which the probability matrix is much more concentrated on the diagonal, the spectral norm is the most powerful statistic. The least squares coefficient with estimated weights is a close second, with the other cases not much better than chance.

For the other two alternatives, we see similar trends. For the cluster anomaly, like the R-MAT-3 case, we get excellent detection performance with SN, followed by LSC with estimated weights, and near-chance for all others. For hubs we see a simimlar behavior to R-MAT-2, with near-perfect detection using LSC with estimated weights, good performance for LSC with given weights, and lower, but still better-than-chance, performance for the other statistics.

5. DISCUSSION

One particularly intriguing phenomenon in the results is the significant increase in anomaly detection performance using the estimated weights rather than given weights for the LSC statistic. On closer inspection, the distribution of test statistics (under both the null and alternative models) using estimated weights has about the same mean as when the true weights are used, but the variance is much tighter; in all cases at least a factor of 7 lower. Using the true weights, we confirm that the sample variance in the simulation data is very close to (3). The variance of the LSC statistic with estimated weights is expressed as

$$E\left[\left(\frac{\sum_i \sum_j \sum_\ell A_{ij} k_i k_j k_\ell}{\sum_i \sum_j k_i^2 k_j^2}\right)^2\right] - E\left[\frac{\sum_i \sum_j \sum_\ell A_{ij} k_i k_j k_\ell}{\sum_i \sum_j k_i^2 k_j^2}\right]^2.$$

Analysis of this quantity is complicated and beyond the scope of this paper, but an additional Monte Carlo simulation confirms that, for the degree distributions in these experiments, it is much smaller than the variance expressed in (3).

Considering the two cases in which SN is the most powerful statistic, it is notable that R-MAT-3 is the R-MAT with the most clustering (i.e., with the 2×2 probability matrix that puts most entries in

the lower right or upper left quarters). Due to this probability structure, the resulting alternative graph will be partitioned, like the cluster anomaly, in a way such that connections within the two partitions are much more likely than those across the partition. Eigenspace techniques have long been associated with graph partitioning, and the same underlying phenomena may be at work here.

Finally, considering the MNL statistic under R-MAT-2, we would like to understand why the lowest neighborhood likelihood is larger under the alternative than under the null model. Looking at the probability matrices for the null and alternative models, we see that, around certain high-degree vertices, the R-MAT model further biases the graph to have edges with high likelihood under Chung–Lu. Since high-degree vertices tend to have the lowest likelihoods (due to the probability of occurrence for any particular edge being rather small), this causes higher minimum likelihoods under the alternative, resulting in the behavior we see in Fig. 1b.

6. SUMMARY

In this paper we investigate the use of goodness-of-fit statistics to reject the hypothesis that an observed graph was generated by a particular model, specifically the Chung–Lu random graph model. We propose 3 goodness-of-fit statistics and analyze their power to discriminate between Chung–Lu graphs and several alternative models. Simulation results demonstrate that the spectral norm of the graph residual performs the best when there is a partitioning of the nodes in which internal connectivity is much more likely than connectivity across the partition, while a least squares fitting coefficient is the most powerful statistic in other cases. Future work will include a deeper theoretical study of the test statistic distributions, as well as the integration of several statistics to optimize detection power.

7. ACKNOWLEDGMENT

The authors wish to thank Nicholas Arcolano, Karl Ni, Matthew Schmidt and Patrick Wolfe for many helpful comments and conversations.

8. REFERENCES

[1] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *Proc. KDD*, 2003, pp. 631–636.

[2] B. A. Miller, N. T. Bliss, and P. J. Wolfe, "Toward signal processing theory for graphs and non-Euclidean data," in *Proc. ICASSP*, 2010, pp. 5414–5417.

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

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

[5] F. Chung, L. Lu, and V. Vu, "The spectra of random graphs with given expected degrees," *PNAS*, vol. 100, no. 11, pp. 6313–6318, 2003.

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

[7] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*, Springer, 2009.

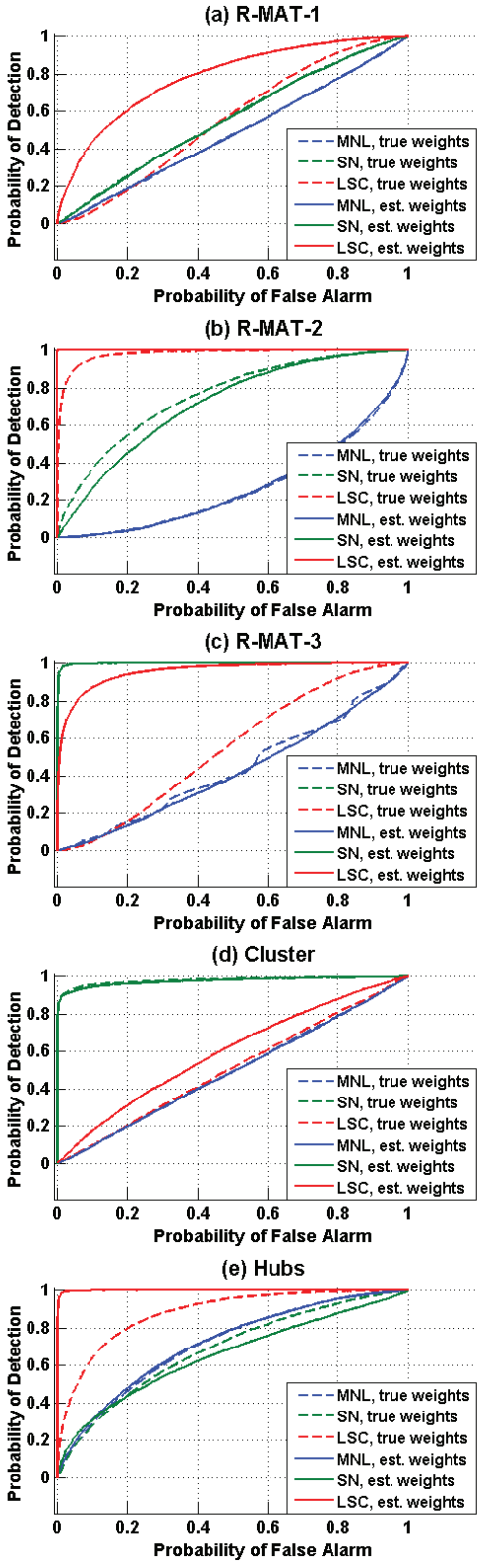

Fig. 1. ROC curves demonstrating the power of the 3 test statistics to discriminate between graphs generated by a Chung–Lu model and those generated by one of the 5 anomalous alternatives.