

# Anomaly Detection for Random Graphs using Distributions of Vertex Invariants

Nash Borges<sup>\*†</sup>, Glen A. Coppersmith<sup>\*†</sup>, Gerard G. L. Meyer<sup>\*†</sup>, and Carey E. Priebe<sup>\*‡</sup>

Johns Hopkins University

<sup>\*</sup>Human Language Technology Center of Excellence

<sup>†</sup>Department of Electrical and Computer Engineering

<sup>‡</sup>Department of Applied Mathematics and Statistics

Email: {nashborges,coppersmith,gglmeyer,cep}@jhu.edu

**Abstract**—Anomaly detection is a longstanding problem with many applications in signal processing. We consider anomaly detection on graphs, a subject which has not previously had treatment in such depth. Our approach is inspired largely by previous work [1]–[3], where anomaly detection in an acoustic signal is accomplished by measuring and comparing the distribution of localized measurements to those available from a non-anomalous signal. In similar spirit, we proceed by comparing distributions of vertex invariants to those obtained from non-anomalous graphs. Specifically, we consider homogeneous Erdős-Rényi random graphs (where each vertex is connected independently with equal probability  $p$ ) to be non-anomalous, and compare them to four classes of heterogeneous alternatives (where a subset of the vertices are connected according to a different process). Our contributions are (1) a novel method of incorporating information from vertex invariants for anomaly detection on graphs, (2) a principled approach to fusing information from an arbitrary number of such statistics, and (3) evaluation on several types of anomalous graphs. We demonstrate superior performance to available state-of-the-art approaches against the specific type of anomalies optimized for, and further demonstrate superior generalization to an entire class of anomalies.

## I. INTRODUCTION

Graphs serve as useful representations of many naturally occurring phenomena, such as communication networks [4], [5], human interactions and turn-taking behavior [6], neocortical maps [7], and biological processes [8]. Therefore, detecting anomalies in such graphs could serve as a useful tool for determining when changes have occurred in these natural phenomena.

Previously, Pao, et al. [9] analyzed the inferential capability of graph invariants when differentiating homogeneous graphs from heterogeneous “chatter” alternatives, where a subset of the vertices are overly-connected. Their results indicate that there is no uniformly most powerful summary statistic across the space of “chatter” alternatives, although the maximum locality statistic has significantly more power than the rest over large regions of the alternative parameter space. The methodology of using graph invariants as test statistics to differentiate homogeneous graphs from “chatter” alternatives is important when deciding how to approach anomaly detection in this setting, but it does not speak to the robustness of using a single graph invariant for the detection of a wider variety of anomalous graphs.

In this work, we propose a methodology of detecting anomalous graphs when a subset of the vertices can be either under- or overly-connected. Our approach is inspired by our previous work in acoustic anomaly detection [1]–[3], where we compare the distributions of local measurements in order to perform the desired global inference task. When performing anomaly detection on graphs, we examine distributions of vertex invariants instead of using a single graph invariant. We demonstrate a computationally efficient method of combining an arbitrary number of vertex measurements when assessing graph abnormality and show significant performance improvements compared to using a single graph invariant for various types of anomalies.

## II. RANDOM GRAPHS

Consider graph  $G = (V, E)$  from the space of simple graphs  $\mathcal{G}$ , where  $V$  is the set of  $n = \text{order}(G) = |V|$  vertices and  $E$  is the set of edges with  $\text{size}(G) = |E|$ . We denote an edge between  $u, v \in V$  as  $uv \in E$  and only consider undirected edges, so  $uv = vu$ . The adjacency matrix  $A$  of graph  $G$  is the matrix in which entry  $a_{ij} = 1$  if  $v_i v_j \in E$ , otherwise  $a_{ij} = 0$ . For brevity, we exclude a general treatment of the subject of graph theory and refer interested readers to [10].

### A. Null Hypothesis

Our null hypothesis ( $H_0$ ) is that the observed graph is drawn from the class of Erdős-Rényi random graphs,  $ER(n, p)$ , with  $n$  vertices where each of the  $\binom{n}{2}$  possible edges exist independently with probability  $p$ .

### B. Alternative Hypotheses

The alternative hypothesis ( $H_A$ ) is that the observed graph is not drawn from the class of Erdős-Rényi random graphs. We test this by generating graphs where a subset of vertices are anomalous and are connected according to a different process than the majority of the vertices. Let the set of anomalous vertices be  $\mathcal{M}$  (of order  $m$ ) and the set of non-anomalous vertices be  $V \setminus \mathcal{M}$  (of order  $n - m$ ). In all cases, the  $\binom{n-m}{2}$  possible edges connecting vertices in  $V \setminus \mathcal{M}$ , exist independently with probability  $p$ , as in  $H_0$ . Moreover, the  $m(n - m)$  possible edges between vertices in  $\mathcal{M}$  and  $V \setminus \mathcal{M}$  also exist independently with probability  $p$ , as in  $H_0$ . The

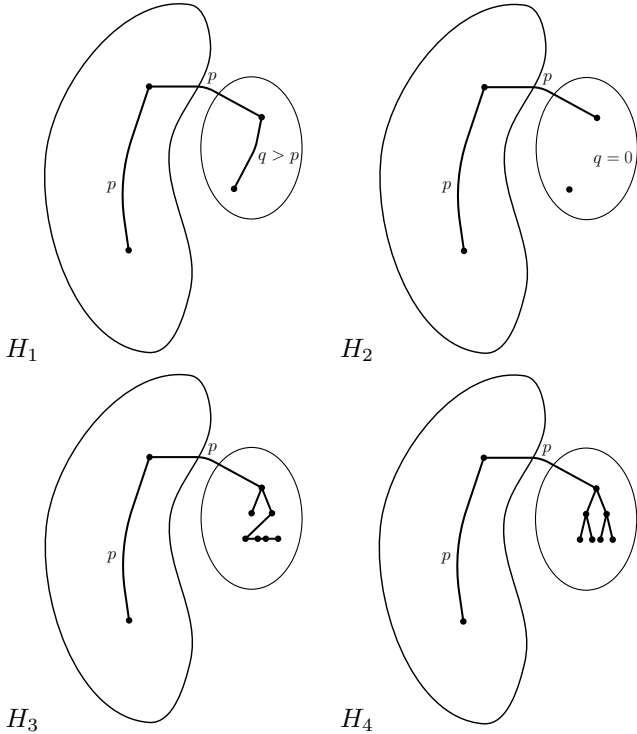


Fig. 1. Depictions of the anomalous graphs investigated.  $H_1 : \kappa(n, p, m, q)$  where  $q > p$ , the *kidney-egg* graph.  $H_2 : \kappa_c(n, p, m, b = 0)$ , the *disappearing-egg* graph.  $H_3 : \kappa_c(n, p, m, b = 1)$ , the *kidney-line* graph.  $H_4 : \kappa_c(n, p, m, b = 2)$ , the *kidney-tree* graph.

four types of anomalous graphs treat the  $\binom{m}{2}$  possible edges between vertices in  $\mathcal{M}$  differently (shown in Figure 1).

- **Increased Connectivity**  $H_1$ : A *kidney-egg* graph,  $\kappa(n, p, m, q)$ , where connections between the  $m$  anomalous vertices exist independently with probability  $q$  where  $q > p$ . This condition is directly comparable to [9].
- **Decreased Connectivity**  $H_{\{2,3,4\}}$ : A *constrained kidney-egg* graph,  $\kappa_c(n, p, m, b)$ , where the  $m$  vertices are either not connected to each other or form a tree with branching factor  $b$  (each vertex connected to at most  $b$  children in  $\mathcal{M}$  and one parent<sup>1</sup>). We investigated  $H_2: b = 0$ , where there are no connections between the vertices in  $\mathcal{M}$ ;  $H_3: b = 1$ , where  $\mathcal{M}$  form a path; and  $H_4: b = 2$ , where  $\mathcal{M}$  form a binary tree.

We also evaluate detection performance when distinguishing Erdős-Rényi graphs from a pooled test set of these anomalous graphs, where half have a local region of increased connectivity ( $H_1$ ) and half have a local region of decreased connectivity ( $H_{\{2,3,4\}}$ ).

### III. METHODS

We consider two classes of test statistics and compare their relative performance for anomaly detection on graphs. Graph invariants previously investigated [9] are unnormalized scalars,  $T : \mathcal{G} \rightarrow \mathbb{R}$ , summarizing either local or global connectivity.

<sup>1</sup>For  $b > 0$ , this means there are exactly  $m - 1$  edges present between vertices in  $\mathcal{M}$ ; equivalent values of  $q$  are extremely small.

We compare this approach to test statistics measuring the abnormality of the joint distribution of multiple vertex invariants.

In this work, we use a training set  $\mathcal{N} = \{G_1, \dots, G_R\}$  of  $R = 1000$  known homogeneous graphs, generated according to  $ER(n = 1000, p = 0.1)$ . We use these graphs to construct test statistics,  $T_{\mathcal{N}} : \mathcal{G} \rightarrow \mathbb{R}$ , and we reject  $H_0$  for large values representing a deviation from normality. Crucially, we use no knowledge of the potential space of anomalies in our test procedure in the hopes of maximizing generalization. In the outlier detection literature [11], this approach is often referred to as novelty detection, since it requires only normal exemplars.

#### A. Graph Invariants

Graph invariants can be used as test statistics [12] and are considered extensively in [9]. Here, we use the total number of edges, maximum degree, maximum locality statistic, total number of triangles, average clustering coefficient, average path length, and two approximations of maximum average degree (one greedy algorithm and one based on the largest eigenvalue of  $A$ ).

#### B. Normalization

Since we are going to test graph homogeneity against heterogeneous alternatives where a subset of vertices have increased or decreased activity, we prefer that the test statistics represent the degree of abnormality of the observation. To do this, we require knowledge of typical values for each graph invariant, which we acquire using our set  $\mathcal{N}$  of known Erdős-Rényi graphs. We adjust each graph invariant,

$$t_{\text{norm}} = \frac{|t - \mu_{t_{\mathcal{N}}}|}{\sigma_{t_{\mathcal{N}}}}, \quad (1)$$

to measure the normalized distance from its mean. This allows us to perform a one-tailed test whether the statistic is abnormally large or small, either of which is indicative of an anomalous graph in the pooled test condition.

#### C. Vertex Invariants

Global and extremum graph invariants aggressively summarize local information, so we consider the following vertex invariants and investigate methods of estimating the abnormality of their distribution. Several are referred to as centrality measures which assess the relative importance of each vertex in the graph.

- **Degree**: The simplest local vertex invariant is its degree or number of incident edges. When used as a measure of local centrality independent of graph order  $n$ , it is often normalized [13],

$$D(v) = \frac{\text{deg}(v)}{n - 1}. \quad (2)$$

- **Betweenness**: The betweenness of a vertex  $v$ ,

$$B(v) = \frac{\sum_{v \neq s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}}{(n - 1)(n - 2)}, \quad (3)$$

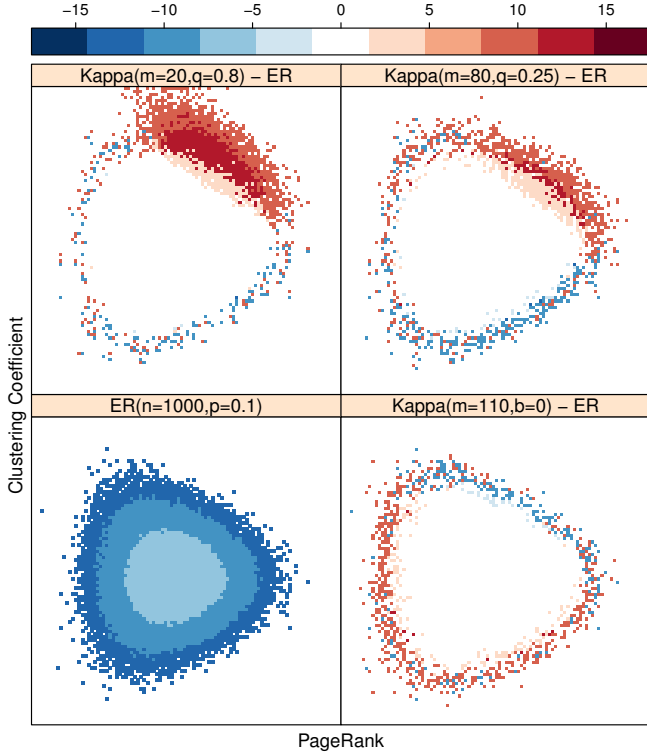


Fig. 2. Log probability histogram of clustering coefficient and PageRank vertex invariants. Lower left panel shows their distribution for  $H_0$ : Erdős-Rényi graphs with  $n = 1000$  vertices where each edge exists independently with probability  $p = 0.1$ . Upper panels show log differences between histograms of invariants from  $H_0$  and  $H_1$  (*kidney-egg*) where  $H_1 = \kappa(n = 1000, p = 0.1, m = 20, q = 0.8)$  on the left and  $H_1 = \kappa(n = 1000, p = 0.1, m = 80, q = 0.25)$  on the right. Lower right panel shows log differences between histograms of invariants from  $H_0$  and  $H_2$  (*disappearing-egg*) where  $H_2 = \kappa_c(n = 1000, p = 0.1, m = 110, b = 0)$ . The histograms shown here have 100 bins per axis resulting in the best performance of the distributional methods for the pooled test condition. The  $\kappa$  parameters were chosen to show different conditions leading to approximately the same AUC of 0.99 when using cross entropy for histogram comparison.

is a measure of global centrality where  $\sigma_{st}$  is the number of shortest paths between vertices  $s$  and  $t$  and  $\sigma_{st}(v)$  is the number of shortest paths between vertices  $s$  and  $t$  that pass through  $v$ .

- **Closeness**: The closeness of vertex  $v$  is the reciprocal of the average distance to other reachable vertices,

$$C(v) = \frac{n-1}{\sum_{u \in V \setminus v} d_p(u, v)}, \quad (4)$$

where  $d_p(u, v)$  is the shortest path distance between vertices  $u$  and  $v$ .

- **Eigenvector Centrality**: The eigenvector centrality [14] of vertex  $v$  is proportional to the sum of scores of all adjacent vertices,

$$EV(v) = \frac{1}{\lambda} \sum_{u \in V, uv \in E} EV(u). \quad (5)$$

It is so named because it is the solution to the eigen-

vector equation of the adjacency matrix  $A$ . In general, many solutions exist, but the additional constraint that  $EV(v) > 0$  for all  $v \in V$  restricts us to the eigenvector corresponding to the largest eigenvalue  $\lambda$ .

- **PageRank**: Originally defined on directed graphs representing the World Wide Web [15], PageRank is a modified version of eigenvector centrality which can be reformulated for undirected graphs as the solution to the recursive equation,

$$PR(v) = (1 - df) + df \sum_{u \in V, uv \in E} \frac{PR(u)}{\deg(u)}, \quad (6)$$

where  $df = 0.85$  is a commonly used damping factor.

- **Triangles**: We denote the number of triangles (cycles of length 3) involving vertex  $v$  as  $\tau(v)$ .
- **Clustering Coefficient**: The local clustering coefficient of vertex  $v$  is defined as

$$CC(v) = \frac{2\tau(v)}{\deg(v)(\deg(v) - 1)}, \quad (7)$$

which measures how close its neighbors are to being fully-connected [16].

- **Locality Statistic**: The first order locality statistic [4] of vertex  $v$  is

$$L(v) = \text{size}(\Omega(N[v])), \quad (8)$$

where  $N$  is the first order neighborhood of  $v$  and  $\Omega$  is the induced subgraph. The first order scan statistic of graph  $G$  is  $S(G) = \max_{v \in V} L(v)$  also investigated here as a graph invariant.

#### D. Divergence of Vertex Invariant Distributions

For each vertex  $v$  in a graph  $G$ , we can fuse the information from  $\mathcal{D}$  vertex invariants into  $\psi(v, G) \in \mathbb{R}^{\mathcal{D}}$ . For notational convenience, we will occasionally drop the operands and refer only to  $\psi$ . Given a set of graphs  $\mathbb{G} = \{G_1, \dots, G_s\}$ , each of order  $n$ , we estimate the joint probability density function  $p_{\mathbb{G}}(\psi)$  of vertex invariants using  $\{\psi(v_1, G_1), \dots, \psi(v_n, G_s)\}$ . Our general approach is to measure the divergence,

$$D(p_{\{\mathbb{G}\}}(\psi) || p_{\mathcal{N}}(\psi)), \quad (9)$$

between density estimates of vertex invariants for graph  $G$  and the training set  $\mathcal{N}$  of non-anomalous graphs.

1) *Histograms*: The oldest, simplest, and most popular form of nonparametric density estimation is the histogram which dates back as far as 1662 to mortality tables in the age of the plague [17]. Histograms provide a consistent estimate of the true underlying probability density function [18] while not making parametric assumptions about its form. Kernel density estimates converge to the true distribution faster than histograms, but this can come at considerable computational and storage costs [18] especially for large sample sizes in a multivariate setting. Adaptive histograms with variable bin widths offer an intriguing compromise, but finding an *optimal* adaptive grid is difficult in practice, and *ad hoc* methods that are easier to implement “need not be better and in fact can be much worse” [19].

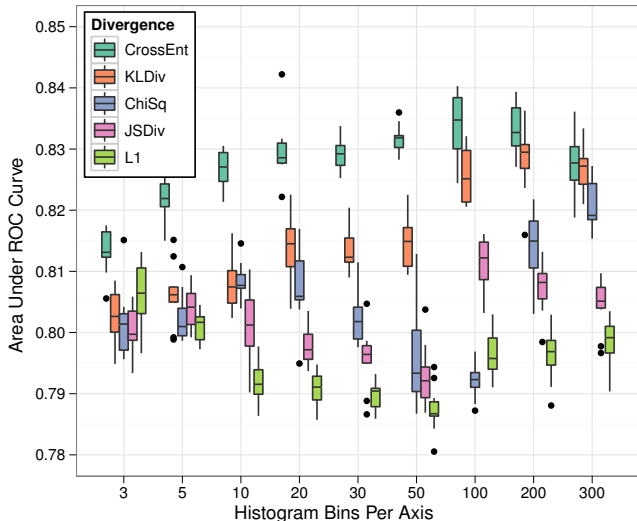


Fig. 3. Performance of several methods of histogram comparison while varying the number of bins per axis when using the clustering coefficient and PageRank vertex invariants. Boxplots depict the interquartile range of the area under ROC curves for 10 resampling experiments of the pooled test condition where half of the anomalous graphs are sampled from those with a region of increased connectivity ( $H_1$ ) and half are sampled from those with a region of decreased connectivity ( $H_{\{2,3,4\}}$ ).

Given a set of graphs,  $\mathbb{G} = \{G_1 = (V_1, E_1), \dots, G_s = (V_s, E_s)\}$ , each of order  $n$ , let a frequency histogram of  $\mathcal{D}$  vertex invariants be defined as a set  $h_{\mathbb{G}} = \{(b_1, c_1), \dots, (b_k, c_k)\}$  of bin centroids  $b_{\gamma} \in \mathbb{R}^D$  with corresponding counts,

$$c_{\gamma} = \sum_{i=1}^s \sum_{v \in V_i} \mathbb{I} \left( \arg \min_j d(b_j, \psi(v, G_i)) = \gamma \right), \quad (10)$$

using the indicator function  $\mathbb{I}$ . We convert this to a density estimate using add-one smoothing,

$$\hat{p}_{\mathbb{G}}(\psi) = \frac{c_{\rho} + \frac{1}{k}}{ns + 1}, \quad (11)$$

where  $\rho = \arg \min_j d(b_j, \psi)$ . This allows for an arbitrary number of vertex invariants, although the sparsity of their joint distribution could arise for large  $\mathcal{D}$ . In this work, we chose fixed-bin histograms to model the joint distribution of each pair of vertex invariants.

2) *Cross Entropy*: To measure the abnormality of an observed graph  $G^* = (V^*, E^*)$  of order  $n$ , we can efficiently compute the negative average log likelihood of its vertex invariants under the model of  $\mathcal{N}$  normal graphs,

$$-\frac{1}{n} \sum_{v \in V^*} \log \hat{p}_{\mathcal{N}}(\psi(v, G^*)) \quad (12)$$

$$\approx -\sum_{i=1}^k \hat{p}_{\{G^*\}}(b_i) \log \hat{p}_{\mathcal{N}}(b_i) \quad (13)$$

$$= H(\hat{p}_{\{G^*\}}, \hat{p}_{\mathcal{N}}) \quad (14)$$

using the equivalent cross entropy between histograms.

3) *Other Methods of Histogram Comparison*: We tested several other methods of histogram comparison [2] including the Kullback-Liebler divergence,  $\chi^2$  test statistic, Jensen-Shannon divergence, and  $L_1$  norm, but were unable to outperform cross entropy (Figure 3).

#### IV. MONTE CARLO EXPERIMENTS

Since we only consider random graphs in this work, the asymptotic distributions of some vertex invariants can be found analytically, especially for  $H_0$  [9]. However, invariant distributions of finite graphs are typically known only for an extremely small number of vertices and those for  $H_{\{1,2,3,4\}}$  would be even more complex. Thus, we estimate performance via Monte Carlo simulation.

We explore anomaly detection against five compound alternatives, namely each of  $H_{\{1,2,3,4\}}$ , along with their pooled combination. We first generate the set  $\mathcal{N}$  of  $R = 1000$  graphs according to  $H_0 : ER(n = 1000, p = 0.1)$ . Another set of graphs are generated according to the same process for testing. We also generate a large set of anomalous graphs according to  $H_A$ , which depends on the test condition described below. We conduct 10 trials, each time randomly sampling 1,000 graphs from  $H_0$  and 10,000 graphs from  $H_A$ . For each method of anomaly detection we compute the test statistic  $T_{\mathcal{N}} : \mathcal{G} \rightarrow \mathbb{R}$  of each graph and reject the null for large values. We compute the area under the receiver operating characteristic (ROC) curve (AUC) to assess performance across the range of possible thresholds.

In order to assess the performance of detecting *kidney-egg* anomalies, we randomly sample graphs for  $H_A$  from  $H_1 : \kappa(n = 1000, p = 0.1, m, q)$  with  $m$  drawn uniformly from  $\{5, 10, \dots, 100\}$  and  $q$  from  $\{0.15, 0.20, \dots, 1.00\}$ . When assessing the detection of each type of *constrained kidney-egg* graph, we randomly sample graphs for  $H_A$  from  $\kappa_c(n = 1000, p = 0.1, m, b)$  with  $m$  drawn uniformly from  $\{5, 10, \dots, 200\}$  and  $b = 0$  for  $H_2$ ,  $b = 1$  for  $H_3$ , and  $b = 2$  for  $H_4$ .

Our primary goal is to find a test statistic that can robustly reject the null hypothesis when presented with a graph from any of the four types of anomalies. To assess this, we compute the AUC when graphs are sampled with equal probability from  $H_1$  and the set  $H_{\{2,3,4\}}$ . This is referred to as the pooled test condition.

#### V. RESULTS

As a baseline, we evaluate the performance of eight graph invariants when detecting anomalies from  $H_{\{1,2,3,4\}}$  along with their pooled combination. We compare this to our family of methods comprising the Cartesian product of (a) all  $\binom{8}{2}$  pairs of vertex invariants using (b)  $\{3, 5, 10, 20, 30, 50, 100, 200, 300\}$  bins per axis when comparing histograms via (c) cross entropy, Kullback-Leibler divergence,  $\chi^2$  test statistic, Jensen-Shannon divergence, and  $L_1$  norm. We jointly optimized these three parameters and show the interquartile range of AUC for the top performing systems along with that of the graph invariants in Figure 4.

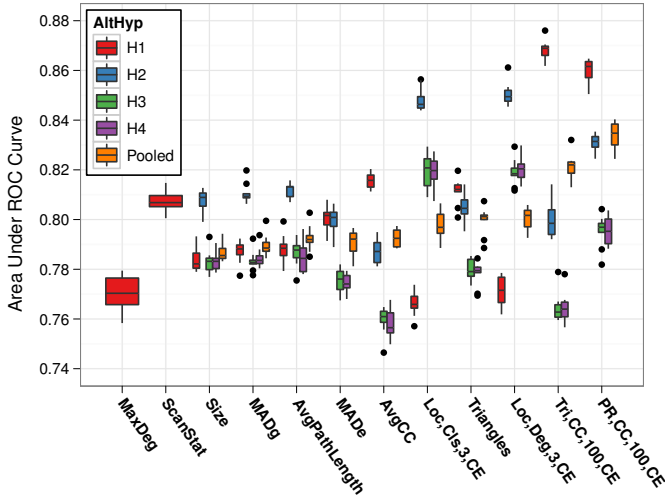


Fig. 4. Performance of graph invariants and the best systems using distributions of vertex invariants. Boxplots depict the interquartile range of the area under ROC curves for 10 resampling experiments with each type of anomaly. The anomalous graphs present in the  $H_1$  test condition are  $\kappa(n = 1000, p = 0.1, m, q)$  graphs with  $m$  drawn uniformly from  $\{5, 10, \dots, 100\}$  and  $q$  from  $\{0.15, 0.20, \dots, 1.0\}$ . The anomalous graphs in the  $H_{\{2,3,4\}}$  test conditions are  $\kappa_c(n, p, m, b)$  graphs with  $m$  drawn uniformly from  $\{5, 10, \dots, 200\}$  and  $b = 0, b = 1$ , and  $b = 2$  for  $H_2, H_3$ , and  $H_4$ , respectively.

TABLE I  
MEDIAN OF AREAS UNDER 10 ROC CURVES

Alternative Hypothesis		$H_1$	$H_2$	$H_3$	$H_4$	Pooled
Graph Invariant	MaxDegree	0.770	0.515	0.519	0.516	0.644
	Size	0.782	0.809	0.783	0.783	0.786
	AvgPathLength	0.788	0.813	0.788	0.784	0.792
	MADg	0.788	0.809	0.783	0.784	0.789
	MADe	0.802	0.801	0.776	0.774	0.792
	ScanStat	0.807	0.562	0.553	0.555	0.681
	Triangles	0.812	0.805	0.779	0.779	0.801
	AvgCC	0.816	0.787	0.761	0.756	0.792
	Distrib.	Tri,CC,100,CE	<b>0.869</b>	0.798	0.763	0.764
Loc,Deg,3,CE		0.772	<b>0.849</b>	<b>0.818</b>	<b>0.820</b>	0.802
Loc,Cls,3,CE		0.766	<b>0.846</b>	<b>0.821</b>	<b>0.820</b>	0.797
PR,CC,100,CE		<b>0.862</b>	<b>0.831</b>	<b>0.797</b>	<b>0.795</b>	<b>0.835</b>

We also report the median AUC for these systems across all test conditions in Table I, emphasizing the performance of distributional systems in bold that perform significantly better ( $p < 0.05$ ) than each graph invariant using a two-sided Wilcoxon paired-sample signed rank test for the 10 trials.

Amongst the vertex invariants investigated here (a), clustering coefficient is the least correlated with measures of vertex centrality in Erdős-Rényi graphs. When optimizing for the detection of increased activity in  $H_1$ , clustering coefficient therefore provides complementary information to the number of triangles involving each vertex. However, when optimizing for the detection of decreased activity in  $H_{\{2,3,4\}}$ , locality is chosen along with another measure of vertex centrality (degree or closeness) even though they are highly correlated.

While clustering coefficient is not used by the systems

optimized for detecting decreased activity, it is used in the top performing system for the pooled test condition along with PageRank. These vertex invariants provide complimentary information about local neighborhood connectivity and global centrality, both of which are useful when detecting anomalies that can have regions of increased or decreased activity (Figure 2). When computing the cross entropy between distributions of these invariants using 100 histogram bins per axis, this system achieves an overall median AUC of 0.835 for the pooled test condition, which is significantly greater ( $p = 0.002$ ) than each graph invariant. It also significantly outperforms the graph invariants when testing against each anomaly type separately. When detecting  $H_1$  graphs, the median AUC of 0.862 was significantly greater ( $p = 0.002$ ) than the top performing graph invariant, average clustering coefficient (median AUC = 0.816). Performance improvements were also significant for  $H_2$  (median AUC = 0.831,  $p = 0.002$ ),  $H_3$  (median AUC = 0.797,  $p = 0.014$ ), and  $H_4$  (median AUC = 0.795,  $p = 0.006$ ) when compared to average path length, which was the top performing graph invariant with median AUCs of 0.813, 0.788, and 0.784, respectively.

Considerable work has been done to optimize histogram bins ( $b$ ) for a given amount of data, especially for a single dimension [18], [20]. In the multivariate setting, optimal bin size also depends on the correlation coefficient between variables [21] along with the measure of divergence between histograms when performing anomaly detection [2]. We thus left it as another free variable in the joint optimization and found that two highly correlated centrality measures, like locality and degree, are best utilized with as few as 3 histogram bins when detecting anomalies with decreased activity ( $H_{\{2,3,4\}}$ ). When using uncorrelated vertex invariants like clustering coefficient and PageRank, 100 bins per axis yields the top performing overall system. While this is contrary to accepted theory where more bins are required to track the diagonal distributions of highly correlated variables, our goal is to optimize anomaly detection performance, not distributional accuracy.

For each test condition, we find that cross entropy is the best method of histogram comparison of those investigated (c) as demonstrated in Figure 3. We attribute this to anomalous vertex invariants being in typically low density regions of the space making the likelihood-equivalent, cross entropy, a natural choice for measuring anomalousness. Also, the disparity in the amount of training and test data ( $10^6$  and  $10^3$  measurements, respectively) is well accommodated by the cross entropy computation where the logarithmic emphasis is not performed on the poorly estimated  $\hat{p}_{\{G^*\}}$ .

While most of our investigation explores compound alternatives to assess overall performance, we show the median AUC for each simple alternate hypothesis in Figures 5 and 6. For the “chatter” alternative, the tradeoff between the  $m$  vertices involved in the anomaly and their increased connectivity  $q$  is well covered in [9]. We display the AUC results here (Figure 5) to confirm that this performance metric is highly correlated with statistical power and to demonstrate that our new methods experience similar trends across this space of parameterized

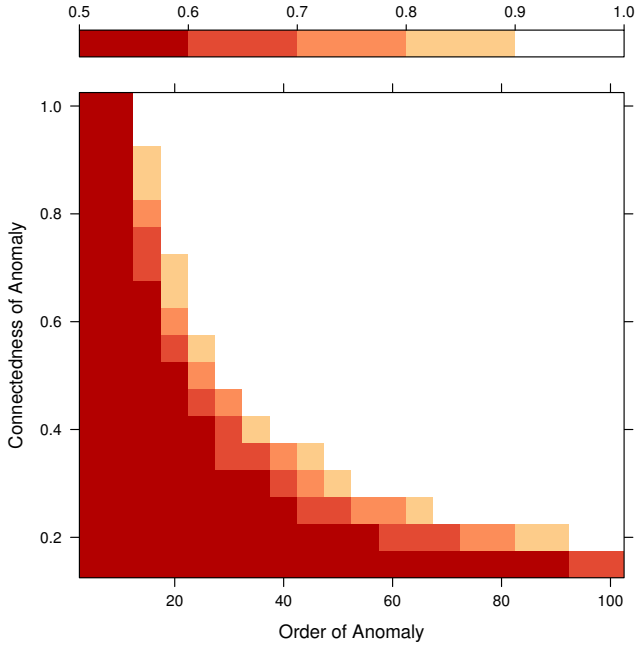


Fig. 5. Area under the ROC curve as a function of the anomalous subgraph's order  $m$  and connectivity  $q$  when using histograms with 100 bins per axis to compare the joint distribution of clustering coefficient and PageRank vertex invariants between  $ER(n = 1000, p = 0.1)$  and  $\kappa(n = 1000, p = 0.1, m, q)$  random graphs.

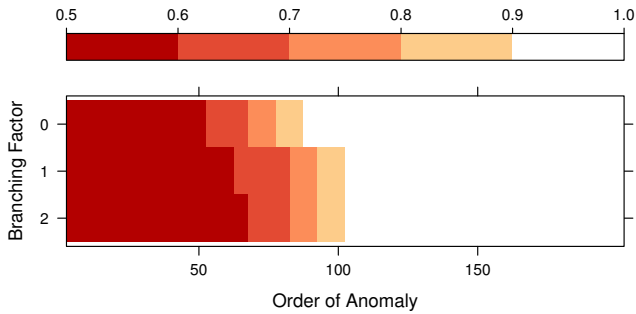


Fig. 6. Area under the ROC curve when varying the anomalous subgraph's order  $m$  and branching factor  $b$  when using histograms with 100 bins per axis to compare the joint distribution of clustering coefficient and PageRank vertex invariants between  $ER(n = 1000, p = 0.1)$  and  $\kappa_c(n = 1000, p = 0.1, m, b)$  random graphs.

alternate hypotheses.

For the anomalous graphs with regions of decreased activity (Figure 6), performance tends to be better for  $b = 0$  compared to  $b = \{1, 2\}$ . This is further demonstrated for the other methods in Figure 4 when comparing  $H_2$  to  $H_{\{3,4\}}$ . This is likely due to the greater degree of connectivity change when  $q = 0$  for  $b = 0$  compared to  $q \approx \frac{m-1}{\binom{m}{2}}$  for  $b = \{1, 2\}$ .

## VI. CONCLUSION

By estimating the joint distribution of two vertex invariants using histograms and assessing its divergence from normality, we significantly outperform all available graph invariants when detecting anomalies with a local region of increased or decreased connectivity. We demonstrate that clustering coefficient and PageRank provide complementary information about vertices, and modeling their distribution makes for a robust system capable of detecting multiple types of anomalous graphs. While we chose these features for their performance in the pooled test condition, they also outperform all available graph invariants when constraining the problem to the detection of each of four anomalous graph types.

## REFERENCES

- [1] N. Borges and G. G. L. Meyer, "Unsupervised distributional anomaly detection for a self-diagnostic speech activity detector," in *CISS*, 2008.
- [2] —, "Coping with training contamination in unsupervised distributional anomaly detection," in *CISS*, 2009.
- [3] —, "Trimmed KL divergence between Gaussian mixtures for robust unsupervised acoustic anomaly detection," in *Interspeech*, 2009.
- [4] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, "Scan statistics on Enron graphs," *Computational and Mathematical Organization Theory*, vol. 11, pp. 229–247, 2005.
- [5] C. E. Priebe, Y. Park, D. J. Marchette, J. M. Conroy, J. Grothendieck, and A. L. Gorin, "Statistical inference on attributed random graphs: Fusion of graph features and content: An experiment on time series of Enron graphs," *Computational Statistics and Data Analysis*, vol. 54, pp. 1766–1776, 2010.
- [6] J. Grothendieck, A. Gorin, and N. Borges, "Social correlates of turn-taking behavior," in *ICASSP*, 2009.
- [7] T. Binzegger, R. J. Douglas, and K. A. C. Martin, "Topology and dynamics of the canonical circuit of cat V1," *Neural Networks*, vol. 22, pp. 1071–1078, 2009.
- [8] E. M. Schmid and H. T. McMahon, "Integrating molecular and network biology to decode endocytosis," *Nature*, vol. 448, pp. 883–888, 2007.
- [9] H. Pao, G. A. Coppersmith, and C. E. Priebe, "Statistical inference on random graphs: Comparative power analyses via Monte Carlo," *Journal of Computational and Graph Statistics*, 2010.
- [10] D. B. West, *Introduction to Graph Theory*. Upper Saddle River, NJ: Prentice Hall, 2001.
- [11] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.
- [12] C. E. Priebe, G. A. Coppersmith, and A. Rukhin, "You say graph invariant, I say test statistic," *ASA Sections on Statistical Computing Statistical Graphics SCGN Newsletter*, vol. 21, no. 2, 2010.
- [13] L. C. Freeman, "Centrality in social networks: conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.
- [14] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.
- [15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *International Conference on the World Wide Web*, 1998, pp. 107–117.
- [16] D. J. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [17] H. Westergaard, *Contributions to the History of Statistics*. P. S. King, 1932.
- [18] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.
- [19] —, *Multivariate Density Estimation*. Wiley, 1992.
- [20] M. P. Wand, "Data-based choice of histogram bin width," *The American Statistician*, vol. 51, no. 1, pp. pp. 59–64, 1997.
- [21] H. A. Sturges, "The choice of a class interval," *Journal of the American Statistical Association*, vol. 21, no. 153, pp. 65–66, 1926.