Anomaly Detection in Time Series of Graphs using Fusion of Graph Invariants

Youngser Park, Carey E. Priebe, and Abdou Youssef

Abstract

Given a time series of graphs $G(t) = (V, E(t))$, $t = 1, 2, \cdots$, where the fixed vertex set V represents "actors" and an edge between vertex u and vertex v at time t ($uv \in E(t)$) represents the existence of a communications event between actors u and v during the tth time period, we wish to detect anomalies and/or change points. We consider a collection of graph features, or invariants, and demonstrate that adaptive fusion provides superior inferential efficacy compared to naive equal weighting for a certain class of anomaly detection problems. Simulation results using a latent process model for time series of graphs, as well as illustrative experimental results for a time series of graphs derived from the Enron email data, show that a fusion statistic can provide superior inference compared to individual invariants alone. These results also demonstrate that an adaptive weighting scheme for fusion of invariants performs better than naive equal weighting.

Index Terms

Statistical inference on graphs, Time series analysis, Random graphs, Change point detection, Hypothesis testing, Graph Invariants, Fusion.

I. INTRODUCTION

Given a time series of graphs $G(t) = (V, E(t))$, $t = 1, 2, \cdots$, where the vertex set $V = [n] = \{1, \cdots, n\}$ is fixed throughout and the edge sets $E(t) \subset \binom{V}{2}$ are time-dependent, we wish to detect anomalies and/or change points. Let us consider vertices to represent "actors," and an edge between vertex u and vertex v at time t ($uv \in E(t)$) represents the existence of a communications event between

Y. Park and C.E. Priebe are with the Department of Applied Statistics and Mathematics, Johns Hopkins University, Baltimore, MD, 21211. See http://www.cis.jhu.edu/faculty/ for current contact information.

A. Youssef is with the Department of Computer Science, George Washington University, Washington, D.C. 20052.
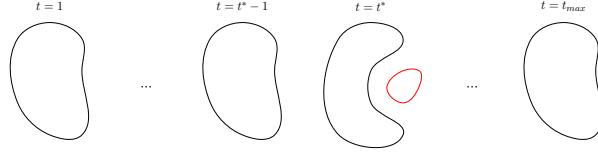
Fig. 1.   Notional depiction of a time series of graphs in which the entire vertex set V behaves in some null state for $t = 1, \cdots, t^* - 1$ and then, at time t^*, a subset of vertices V_A exhibits a change in connectivity behavior.

actors u and v during the tth time period. Thus $E(t)$ represents the collection of (unordered) pairs of vertices which communicate during $(t - 1, t]$. We will not consider directed edges or hyper-graphs (hyper-edges consisting of more than two vertices) or multi-graphs (more than one edge between any two vertices at any time t) or self-loops (an edge from a vertex to itself) or weighted edges, although all of these generalizations of simple graphs may be relevant for specific applications.

The specific anomaly we will consider is the "chatter" alternative – a small (unspecified) subset of vertices with excessive communication amongst themselves during some time period in an otherwise stationary setting, as depicted in Figure 1. This figure notionally depicts the entire vertex set V behaving in some null state for $t = 1, \cdots, t^* - 1$; then, at time t^*, a collection of vertices $V_A \subset V$ ($|V_A| = m, 2 \le m \ll n$) exhibit probabilistically higher connectivity. (The remaining $\binom{n}{2} - \binom{m}{2}$ interconnection probabilities remain in their null state at time t^*.) Our statistical inference task is then to determine whether or not there has emerged a "chatter" group at some time $t = t^*$, as shown in Figure 1.

The latent process model for time series of graphs presented in [1] provides for precisely this temporal structure. Each vertex is governed by a continuous time, finite state stochastic process $\{X_v(t)\}_{v \in V}$, with the state-space given by $\{0, 1, \cdots, K\}$. The probability of edge uv at time t is determined by the inner product of the sub-probability vectors specified by $\int_{t-1}^{t} I\{X_w(\tau) = k\} d\tau$, $k = 1, \cdots, K$, for $w = u, v$. For the scenario depicted in Figure 1, the vertex processes $\{X_v(t)\}_{v \in V_A}$ are stationary until time $t^* - 1$ and then undergo a change point, while the processes $\{X_v(t)\}_{v \in V \setminus V_A}$ remain stationary throughout all time.

In [1], the model produces a dependent time series of graphs $G(t)$, each of which is itself a latent position model with conditionally independent edges given $\{X_v(\tau)\}_{v \in V, \tau \le t}$. The model allows two simplifying approximations; a second-order (central limit theorem) approximation with temporally independent random graphs each of which is itself a random dot product ([2], [3], and Section 16.4 in [4]) latent position model [5], and a first-order (law of large numbers) approximation with temporally independent
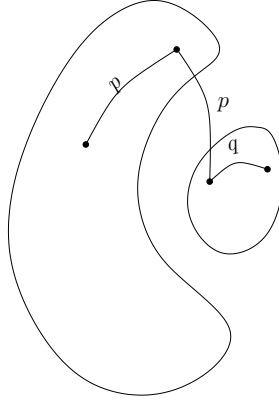
Fig. 2. The "kidney-egg" random graph model, denoted $\kappa(n, p, m, q)$. The small "egg" represents the m vertices (V_A) that exhibit chatter (each edge occurring with probability q). The "kidney" is the population of $n-m$ vertices which are not exhibiting chatter (each edge occurring with probability $p < q$). Edges between a vertex in the kidney and a vertex in the egg occur with probability p. When $m = 0$ or $q = p$, this model degenerates to $ER(n, p)$.

random graphs each of which is itself an independent edge random graph model [6].

The simplicity of the first-order approximation, depicted in Figure 2 for the special case of *homogeneity* vs. *kidney-egg*, provides a useful framework for description. If the vertex processes $\{X_v(t)\}_{v \in V}$ are independent and identical, with stationary probability vector $\pi_0 = [\pi_{0,0}, \pi_{0,1}, \cdots, \pi_{0,K}]'$, then the first-order approximation produces a temporally independent series of homogeneous independent edge Erdös-Rényi random graph (denoted by $ER(n, p)$) with $p = \langle \overline{\pi}_0, \overline{\pi}_0 \rangle$, where $\overline{\pi}_0 = [\pi_{0,1}, \cdots, \pi_{0,K}]'$. The vertex processes $\{X_v(t)\}_{v \in V_A}$ change at time $t^* - 1$, taking on stationary probability vector π_A, so that $G(t^*)$ is a kidney-egg independent edge $\kappa(n, p, m, q)$ random graph with $q = \langle \overline{\pi}_A, \overline{\pi}_A \rangle$. The idea that the change point consists of a small collection of vertices exhibiting *excessive* interconnection probability results in the restriction of this model to the case $q > p$. (Here we have assumed, for simplicity, that the geometry provides $\langle \overline{\pi}_0, \overline{\pi}_A \rangle = p$.)[1]

In [7], the scan statistic graph invariants are introduced and applied to the problem of detecting "chatter" anomalies in time series of Enron graphs. In [8], various graph invariants (size, maximum degree, etc.)

---

[1]If $\langle \overline{\pi}_0, \overline{\pi}_A \rangle = p' \geq p$, then we have $\mathbb{E}[deg(v)] = mq + (n-m) \times p'$ for a $v \in egg$, and $\mathbb{E}[deg(v)] = (n-m) \times p + m \times p'$ for a $v \in kidney$. The difference between these two expected degrees is then $m \times (q - p') + (n - m) \times (p' - p)$. If m is of order $o(n)$, we see that the above expression is minimized over $p' \geq p$ when $p' = p$, which indicates that the most *difficult* scenario is when $p' = p$.

are considered for their power as test statistics in testing $H_0 : ER(n, p)$ vs. $H_A : \kappa(n, p, m, q)$. It is demonstrated that no single invariant is uniformly most powerful. See also [9].

In [10] the principal eigenvector of a matrix based on the graph is tracked over time, and an anomaly is declared to be present if its direction changes by more than some threshold. Researchers in [11] have addressed problems in dynamic network analysis such as detection of anomalies or distinct subgraphs in large, noisy background in signal processing fields. Recently, [12] proposed a methodology of detecting anomalous graphs by examining distributions of vertex invariants instead of using a single graph invariant. They used a simple non-time series of simulated ER random graph models. In [13], a locality statistic using a generalized likelihood ratio test statistic (they call this a scan statistic) has been applied for an online network intrusion detection. Other notable recent efforts in this direction include [14]–[16].

In this paper, we consider the problem of detecting "chatter" anomalies in time series of graphs using combinations of invariants. We present experimental results for anomaly detection on time series of simulated data from the model in [1], as well as an investigation of a time series of graphs extracted from the Enron email corpus, to demonstrate that a statistic which combines multiple invariants can provide superior inference compared to individual invariants alone. We further demonstrate an adaptive weighting scheme for fusion of invariants that performs better than naive equal weighting.

Section II presents the graph features (invariants, used as statistics) considered herein, Section III introduces our adaptive fusion, and Section IV presents results with simulated data as well as Enron email data. We conclude with discussion in Section V.

II. GRAPH FEATURES

We investigate a collection of nine graph features similar to that considered in [8]: size, maximum degree, maximum average degree (eigenvalue approximation), scan statistic (scale 1,2,3), number of triangles, clustering coefficient, and (negative) average path length. In all cases, a large value of the feature F is an evidence in favor of *excessive* interconnection probability.

A. Invariants

1) Size: The size of a graph is the number of edges in the graph, given by

$$F_1(G) = \texttt{size}(G) = |E(G)|.$$

This is the simplest global graph statistic.

2) Maximum Degree: The maximum degree $\Delta(G)$ of a graph is given by

$$F_2(G) = \Delta(G) = \max_{v \in V} deg(v)$$

where $deg(v)$ is the degree of vertex v. This is the simplest localized graph feature.

3) Maximum Average Degree: The maximum average degree of a graph is the maximum over all subgraphs H of G of the average degree of H. If $deg(v)$ is the degree of vertex v, then the average degree of a graph $G = (V, E)$ is given by

$$\bar{d}(G) = \frac{1}{|V|} \sum_{v \in V} deg(v) = \frac{2 \times \texttt{size}(G)}{\texttt{order}(G)}$$

where $\texttt{order}(G) = |V|$, the number of vertices. Thus the maximum average degree is given by

$$\texttt{MAD}(G) = \max_{H \subset G} \bar{d}(H)$$

where the maximum is over all (induced) subgraphs H of G.

Since $\texttt{MAD}(G)$ is difficult to compute exactly [17], we resort to an eigenvalue approximation. $\texttt{MAD}(G)$ is bounded above by the largest eigenvalue of the adjacency matrix of G, denoted $\texttt{MAD}_e(G)$, and we use

$$F_3(G) = \texttt{MAD}_e(G).$$

As demonstrated in [8], the eigenvalue method appears to be strictly better at detecting increased local activity than the greedy approximation method of [17] (Problem 5.7.2, page 90).

4) Scan Statistic: Scan statistics [7] are graph features based on local neighborhoods of the graph. We will consider the scan statistic $\texttt{SS}_k(G)$ to be the maximum number of edges over all k^{th} order neighborhoods, where the kth order neighborhood of a vertex v, $N_k[v]$, is the set of vertices whose graph shortest path distance from v is less than equal to k. We will consider $k = \{1, 2, 3\}$, where $\texttt{SS}_k(G)$ is given by

$$F_{3+k}(G) = \texttt{SS}_k(G) = \max_{v \in V} \texttt{size}(\Omega(N_k[v])),$$

where $\Omega(N_k[v])$ denotes the induced subgraph.

5) Number of Triangles: We consider the total number of triangles in G. If A is the adjacency matrix for the graph G, then the number of triangles is given by

$$F_7(G) = \tau(G) = \frac{\texttt{trace}(A^3)}{6}.$$

The trace is zero if and only if the graph is triangle-free.

6) Clustering Coefficient: We consider the global clustering coefficient (CC) in G, given by

$$F_8(G) = \mathtt{CC}(G) = \frac{ct(G)}{ot(G)},$$

where ct is the number of closed triplets (a subgraph with three vertices and three edges) and ot is the number of open triplets (a subgraph with three vertices and at least two edges). This measures the probability that the adjacent vertices of a vertex are connected. This is sometimes called the *transitivity* of a graph.

7) Average Path Length: The average path length (APL) is given by

$$\mathtt{APL}(G) = \frac{\sum_{u,v} s(u, v)}{n(n - 1)},$$

where $s(u, v)$ is the shortest path between vertices u and v. This measures how many steps are required to access every other vertex from a given vertex, on average. Unlike our other invariants, a *small* value of the average path length is an evidence in favor of *excessive* interconnection probability, so we use the negated value

$$F_9(G) = -\mathtt{APL}(G)$$

in this work. (If no path exists between u and v, we use $s(u, v) = 2 \max s(u', v')$, where the maximum is taken over all pairs of vertices that have an existing path between them.)[2]

B. Temporal Normalization

The purpose of our inference is to detect a local (temporal) behavior change in the time series of graphs. In particular, we wish to consider as our alternative hypothesis that a small (unspecified) collection of vertices (the "egg") increases their within-group activity at some time t^* as compared to recent past while the majority of vertices (the "kidney") continue with their normal behavior. The null hypothesis, then, is a form of temporal homogeneity – no probabilistic behavior changes in terms of graph features. See Figure 3.

As mentioned in [7], the raw features $F_i(G(t))$ are standardized using a quantity computed from the recent past:

$$S_i(t) = \frac{F_i(G(t)) - \widetilde{\mu}_{i,\ell}(t)}{\widetilde{\sigma}_{i,\ell}(t)},$$

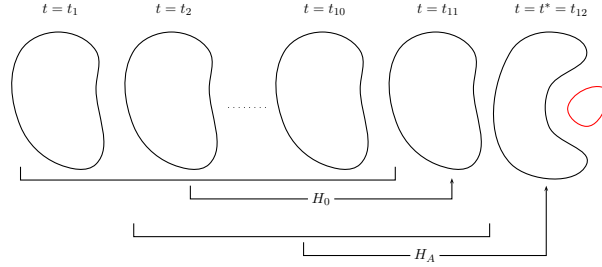[2]In fact, the average path length (APL) is inappropriate for sparse (highly disconnected) graphs.

Fig. 3. H_0 at $t = t^* - 1$ and H_A at $t = t^*$. The H_0 state compares previous many (10 in this case) null graphs to a null graph, $G(t = t_{11})$ and the H_A state compares many null graphs to an alternative graph, $G(t = t^* = t_{12})$.

where $\widetilde{\mu}_{i,\ell}(t)$ and $\widetilde{\sigma}_{i,\ell}(t)$ are the running mean and standard deviation estimates of F_i based on the most recent ℓ time steps; that is,

$$\widetilde{\mu}_{i,\ell}(t) = \frac{1}{\ell} \sum_{t'=t-\ell}^{t-1} F_i(G(t'))$$

and

$$\widetilde{\sigma}_{i,\ell}^2(t) = \frac{1}{\ell - 1} \sum_{t'=t-\ell}^{t-1} (F_i(G(t')) - \widetilde{\mu}_{i,\ell}(t))^2.$$

Then, a detection at time t is obtained when $S_i(t)$ is large. (Note that for the localized statistics (maximum degree, maximum average degree, and the scan statistics) we must first perform *vertex standardization*, as in [7] Section 6, so that, for an *inhomogeneous* collection of stationary null vertex processes, the most active vertices do not dominate these statistics.)

C. Simulation

Our general algorithm for implementing the time series of random dot product graphs is presented in Algorithm II.1. The only difference among our three models in [1] occurs in line 3, where the probability vectors for vertices are obtained; the first approximation uses fixed (non-random or deterministic) probability vectors π_0 and π_A so that $\langle \overline{\pi}_0, \overline{\pi}_0 \rangle = \langle \overline{\pi}_0, \overline{\pi}_A \rangle = p$ and $\langle \overline{\pi}_A, \overline{\pi}_A \rangle = q$ while the second approximation and the exact models use random probability vectors [1].

Density estimates of $S_i(t)$ for all nine features are presented in Figure 4 (using $\ell = 5$). Black denotes $H_0 : S_i(t^* - 1)$ and red denotes $H_A : S_i(t^*)$. As we can see from this figure, all features have mean zero and variance one (approximately) for H_0. It is our goal to measure the performance of each individual graph feature, and then compare these results with the effectiveness of combining features, on our statistical inference task.

Algorithm II.1 Time Series of Random Dot Product Graph

Require: n, π_0, π_A, t_{max}

1: **for all** time t such that $0 < t \leq t_{max}$ **do**

2:   initialize the $n \times n$ adjacency matrix A_t with zeros

3:   $vp \leftarrow$ calculate probability vectors for all vertices using (π_0, π_A)

4:   **for all** vertex u such that $1 \leq u \leq n$ **do**

5:     **for all** vertex v such that $1 \leq v \leq n$ **do**

6:       **if** $u > v$ **then**

7:         $e \leftarrow \langle vp_u, vp_v \rangle$ {vector dot product}

8:         $A_t[u, v] \leftarrow A_t[v, u] \leftarrow Bernoulli(e)$ {draw an edge}

9:       **end if**

10:     **end for**

11:   **end for**

12:   $A[t] \leftarrow A_t$

13: **end for**

14: **return** A, time series of graph

Comparative power results for the individual features are depicted in Figure 5, with a cumulative color bar for each feature. For the most subtle case (when q is small, in blue) the power for each feature is relatively low, while higher power is achieved as q increases. These results agree qualitatively with the results presented in [8].

III. FUSION OF GRAPH FEATURES

We will consider two weighting methods for fusion of our graph features introduced in Section II. Our fusion test statistic is given by

$$S^w(t) = \sum_{i=1}^{d} w_i(t) S_i(t),$$

where d is the number of graph features ($d = 9$, for our investigations).

A. Weighting

The naive equal weighting scheme is given by
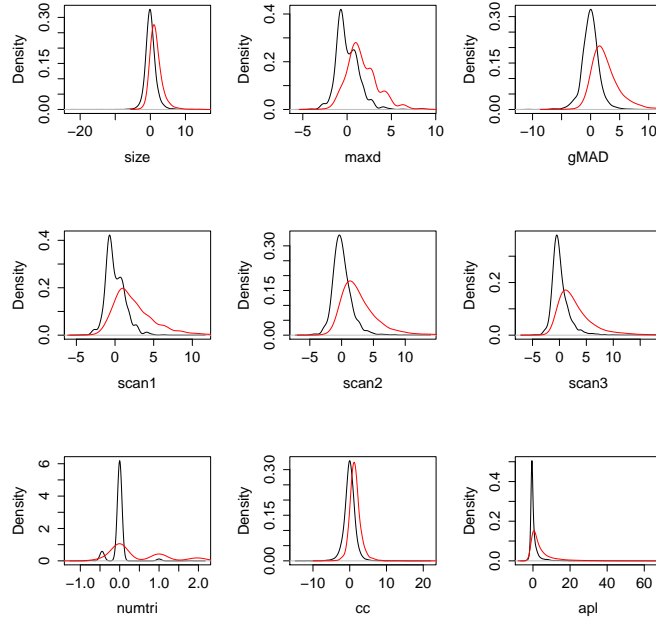
$$w_i(t) = 1/d$$

Fig. 4. Density estimates for $M = 10,000$ Monte Carlo replicates of $S_i(t)$ in the first approximation model. $G(t) = ER(n = 50, p = 0.01)$ for $t = 1, \cdots, t^* - 1$ and $G(t^*) = \kappa(n = 50, p = 0.01, m = 6, q = 0.3)$. For each invariant, black denotes $H_0 : S_i(t^* - 1)$ and red denotes $H_A : S_i(t^*)$.

for all i, and t.

Our *adaptive* weighting scheme uses

$$w_i(t) = \frac{|S_i(t) - \mu_i(t)|}{\sigma_i(t)} \approx |S_i(t)|,$$

where $\mu_i(t)$ and $\sigma_i(t)$ are the mean and the standard deviation of $S_i(t^*-1)$ over M Monte Carlo replicates. (Due to our temporal normalization, all features have mean zero and variance one (approximately) when "recent past" consists of stationarity, which is the assumption when testing for change at time t.) A detailed algorithm of this approach is shown in Algorithm III.1.

Notice that the adaptive weights are a function of the graph $G(t)$ being tested (line 6 of the algorithm). This implies that the features with larger deviations from the norm get higher weights and contribute more to the inference.

B. Examples

A graphical example is illustrated in Figures 6 and 7. In Figure 6, each point represents a Monte Carlo replicate of time series of graph in two-dimensional Euclidean space using the first two features (size and
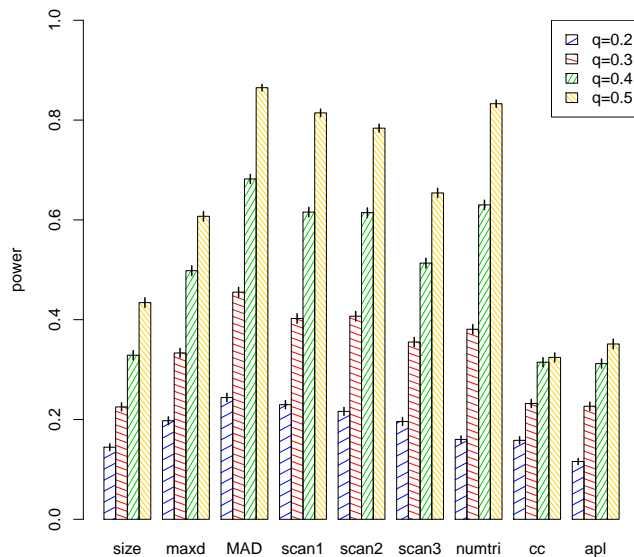
Fig. 5. Statistical power for our nine graph features in the first approximation model. $G(t) = ER(n = 50, p = 0.01)$ for $t = 1, \cdots, t^* - 1$ and $G(t^*) = \kappa(n = 50, p = 0.01, m = 6, q)$, for $q \in \{0.2, 0.3, 0.4, 0.5\}$ and allowable Type I error rate $\alpha = 0.05$, based on $M = 10,000$ Monte Carlo replicates. The error bars represent $1.96 \times$ standard error for the sample means.

maximum degree). The black points (circles) are $H_0 : S_i(t^* - 1)$, and the color points are $H_A : S_i(t^*)$; the points above the detection boundaries (critical values in Algorithm III.1, line 8) are colored in green ("+" symbols) and represent the power of the test. Notice that this boundary is linear for the equal weighting while it is not for the adaptive weighting. The former is because the boundary is calculated based on equal weighting for all $S_i(t^* - 1)$ points; the slope of the line is always -1 and the intercept can be calculated with a given significance level of the test (*i.e.,* $ax + by > c$, $a = b = 1/d$, $\therefore y > -x + dc$, where $c = cv$). For the adaptive weighting case, meanwhile, the color of the $S_i(t^*)$ points are determined by the distance from each point to μ_0, the mean vector of $S_i(t^* - 1)$; the points whose fused values are bigger than the critical value will get the green colors. This means that every $S_i(t^*)$ point gets a different weight and therefore the detection boundary is not linear. Figure 7 shows the adaptive weighting case for various values of q. As q increases, there are more green points, which implies higher power as expected.

IV. FUSION EXPERIMENTS

A. Simulations

Algorithm III.1 Hypothesis Test using Adaptive Weighting Fusion

---

Require: $S_i(t) : M \times t_{max} \times d$ normalized feature matrix, t^*

1: $S_i(t^* - 1) \leftarrow M \times d$ matrix for null at time $t^* - 1$ from $S_i(t)$, and

$S_i(t^*) \leftarrow M \times d$ matrix for alternative at time t^* from $S_i(t)$

2: $\mu_0(t) \leftarrow 1 \times d$ mean vector of $S_i(t^* - 1)$, and

$\sigma_0(t) \leftarrow 1 \times d$ standard deviation vector of $S_i(t^* - 1)$ over M Monte Carlo replicated

3: $pwr \leftarrow 0$

4: **for all** replicate j such that $1 \leq j \leq M$ **do**

5:    $x \leftarrow S_i(t^*)[j,]$ {single replicate of $S_i(t^*)$}

6:    $w \leftarrow |x - \mu_0(t)|/\sigma_0(t)$ {$1 \times d$ weight vector}

7:    $S^w(t^* - 1) \leftarrow \sum_i^d w_i S_i(t^* - 1)$ {$1 \times M$ fused null vector}

8:    $cv \leftarrow \text{quantile}(S^w(t^* - 1), 0.95)$ {critical value: 95% quantile}

9:    $S^w(t^*) \leftarrow \sum_i^d w_i x_i$ {fused scalar of x}

10:    **if** $S^w(t^*) > cv$ **then**

11:      $pwr \leftarrow pwr + 1$

12:    **end if**

13: **end for**

14: **return** pwr/M, power of the test

The simulation setup of this experiment is the same as the one in Section II-C except that fusion of graph features is applied. The performance of fusion with all nine features is depicted as horizontal lines in Figure 8. In all cases, the fusion lines are above the corresponding individual bars, and the adaptive weighting fusion lines are above the equal weighting fusion lines.

Figure 9 depicts power as a function of fusion dimension for the different weighting schemes for the three models in [1]. Given a fusion dimension d', all $\binom{9}{d'}$ possible combinations of features are considered for the fusion and the best performance is plotted. The difference in performance among the three models in [1] is minimal ("qualitatively similar"), while the superiority of the adaptive weighting scheme (with $\triangle$ symbol) is apparent. Table I depicts the actual weightings obtained via the adaptive weighting scheme for $d' = 4$. We see that, for the most part, the same features are selected for all three models in [1].

In Figure 10 we present a statistical power plot of fusion using all nine features ($d' = d = 9$) with
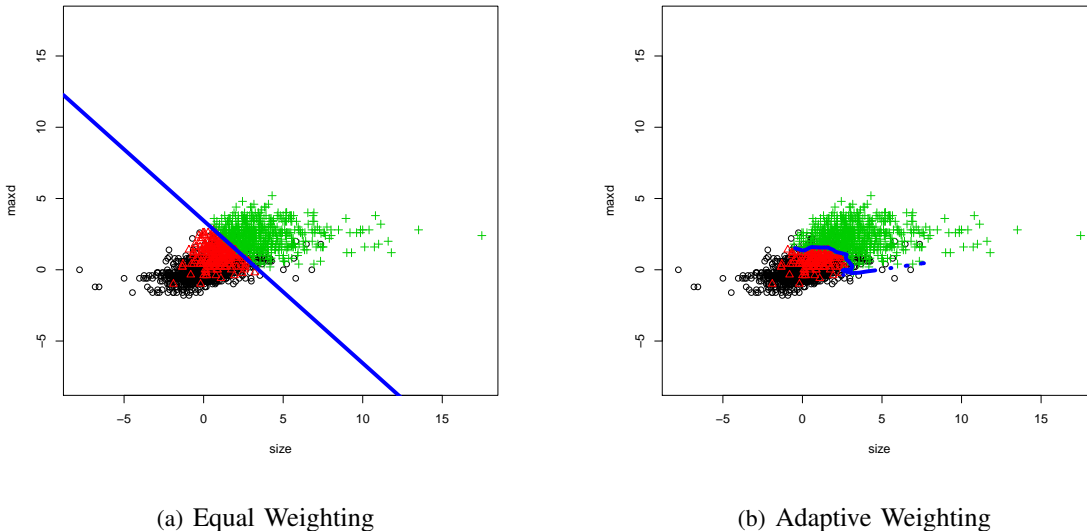
(a) Equal Weighting

(b) Adaptive Weighting

Fig. 6. Scatter plots for size versus maximum degree for each fusion technique. Each point represents a Monte Carlo replicate. The black points (circles) are $S_i(t^* - 1)$, and the color points are $S_i(t^*)$; the points above the detection boundaries (critical values) are colored in green ("+" symbols). The ratio of the number of green points over the total of green and red points represents the power of the test: power = 0.457 for the equal weighting and power = 0.564 for the adaptive weighting. Blue lines represent detection boundaries, which provide *quantitative* rejection regions.

TABLE I

THE ESTIMATED WEIGHTINGS OBTAINED VIA THE ADAPTIVE WEIGHTING SCHEME FOR $d' = 4$ FROM FIGURE 9. WE SEE THAT, FOR THE MOST PART, THE SAME FEATURES ARE SELECTED FOR ALL THREE MODELS IN [1].

model	$\arg \max_i$	w_1	w_2	w_3	w_4
1st approx	(1,2,6,7)	2.66	0.86	1.30	0.10
2nd approx	(1,2,6,7)	2.24	3.88	4.62	0.11
exact	(1,2,6,7)	1.25	5.14	6.01	13.9

$q = 0.3$ and $\alpha = 0.05$ as a function of the rate parameter r for the vertex processes[3]. These results demonstrate that (1) adaptive weighting is superior to equal weighting, (2) the second approximation is more faithful to the exact model than is the first approximation, and (3) both approximations are accurate

[3]The parameter r controls the variability of the latent stochastic processes $\{X_v(t)\}$ for the vertices. In particular, a large value of r corresponds to small variability in $\{X_v(t)\}$ (the second-order approximation), and as $r \to \infty$ the processes $\{X_v(t)\}$ converge to the stationary probability vectors π_0 and π_A (the first-order approximation). See [1] for detail. We have used $r = 1024$ for all other results presented herein.
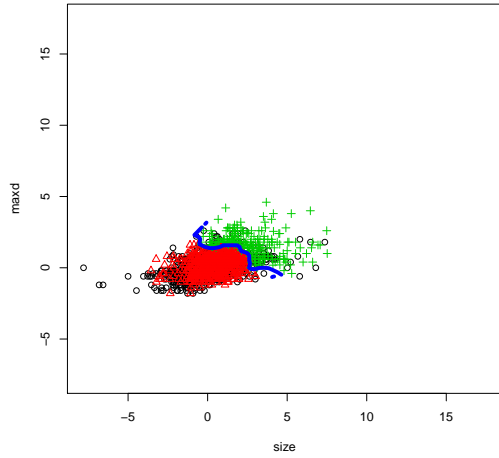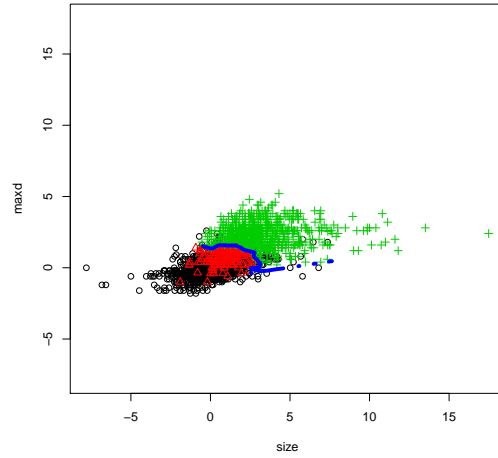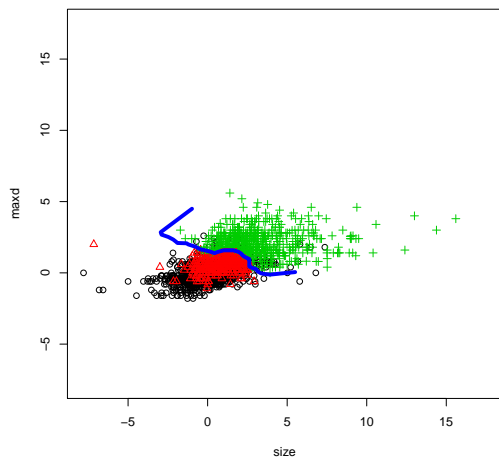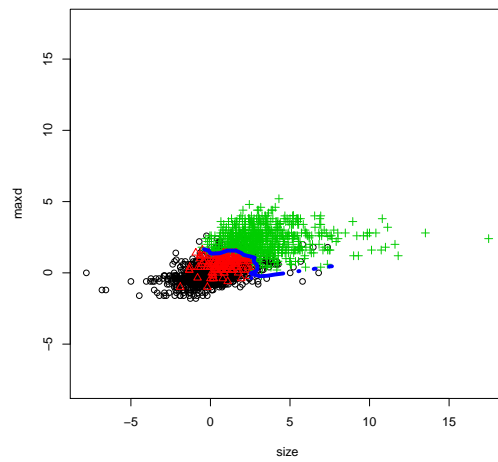
(a) $q = 0.2$

(b) $q = 0.3$

(c) $q = 0.4$

(d) $q = 0.5$

Fig. 7. Scatter plots for size vs. maximum degree for adaptive weighting for $q = \{0.2, 0.3, 0.4, 0.5\}$. Each point represents a Monte Carlo replicate. The black points (circles) are $S_i(t^* - 1)$, and the color points are $S_i(t^*)$; the points above the detection boundaries (critical values) are colored in green ("+" symbols). The actual powers of the test are 0.332, 0.564, 0.775, and 0.917, respectively. As q increases, there are more green points ("+" symbols), which implies higher power. Blue lines represent detection boundaries, which provide *quantitative* rejection regions.

for large r.

Fig. 8. Statistical power for our nine graph features and two fusion schemes in the first approximation model. $G(t) = ER(n = 50, p = 0.01)$ for $t = 1, \cdots, t^* - 1$ and $G(t^*) = \kappa(n = 50, p = 0.01, m = 6, q)$, for $q \in \{0.2, 0.3, 0.4, 0.5\}$ and allowable Type I error rate $\alpha = 0.05$, based on $M = 10,000$ Monte Carlo replicates. The horizontal lines indicate the power using fusion statistics $S^w(t)$ with $d' = 9$. The error bars represent $1.96 \times$ standard error for the sample means. The superiority of adaptive weighting (solid lines) over equal weighting (dashed lines) is apparent.

B. Enron Email Data

We use the Enron email data used in [7] for this experiment. The nine features, $S_i(t)$ for $1 \leq t \leq 189$, are calculated for graphs derived from email messages among $n = 184$ executives during one week periods. Figure 11 depicts histograms of $S_i(t)$ for each i.

Our interest is the "alias" detection identified at week 132 in [7], when an employee changes his/her email address. Therefore, we choose $t^* = 132$, the third week of May 2001. Figure 12 depicts scatter plots of $S_i(t)$ for $t = \{1, \ldots, 132\}$ for various pairs of invariants, where $S_i(t^*)$ is shown in red. Unlike the simulation in Figure 7, Monte Carlo replicates of graph are not available for real data; therefore the 131 previous graphs (shown as black points in the figure) are used to determine detection boundaries. This investigation reveals that the combination of size and maximum degree allows detection based on $S_i(t^*)$ for both weighting schemes (the red point is above both critical lines, in panel a), while only the adaptive weighting scheme detects the anomaly for the other three feature pairs depicted (panels b,c,d).
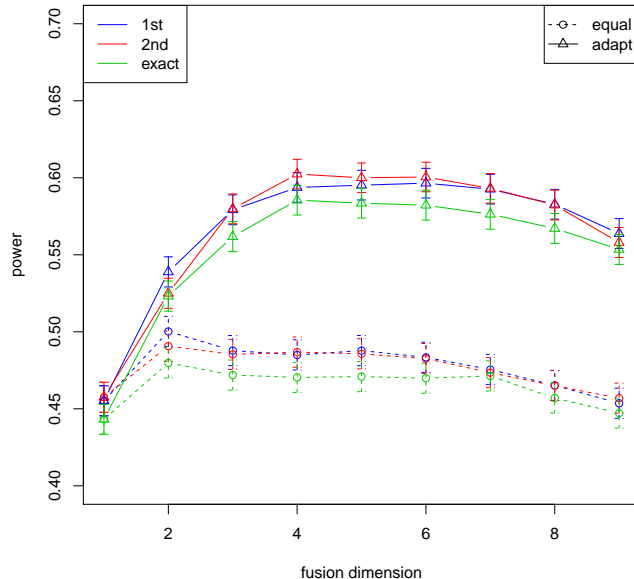
Fig. 9. Statistical power plots for fusion statistics for the three models in [1] as a function of fusion dimension when $q = 0.3$, $M = 10,000$, and $\alpha = 0.05$. The error bars represent $1.96 \times$ standard error for the sample means. The fusion dimensions (d') are chosen from the best possible combinations. The difference in performance among the three models is minimal. The adaptive weighting scheme (with $\triangle$ symbol) is superior to equal weighting.

The performance of equal and adaptive weighting fusion methods with all possible combinations of features at $t^* = 132$ are summarized in Table II. For example, when the fusion dimension $d' = 2$, the possible number of combination of feature dimensions is 36, and both equal and adaptive weighting methods can detect 24 cases, but only adaptive weighting scheme can detect 5 additional cases. Note that there is no case that only equal weighting scheme can detect while adaptive weighting scheme cannot.

V. DISCUSSION

We have demonstrated, via simulation results using a latent process model for time series of graphs as well as illustrative experimental results for a time series of graphs derived from the Enron email data, that an adaptive weighting methodology for fusing information from graph features provides superior inferential efficacy for a certain class of anomaly detection problems.

One notable implication of this work is that inferential performance in the mathematically tractable approximation models in [1] does indeed provide guidance for methodological choices applicable to the
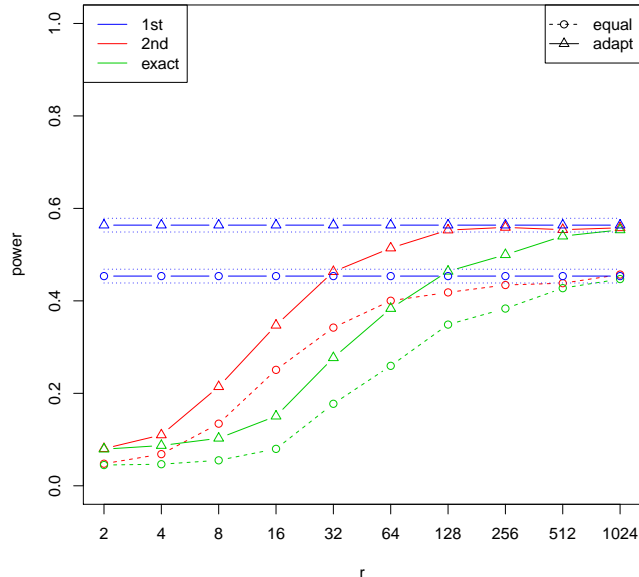
Fig. 10. Statistical power as a function of rate parameter r for models in [1] and both weighting schemes based on $M = 10,000$ Monte Carlo replicates, with $d' = d = 9$, $q = 0.3$, and $\alpha = 0.05$. The horizontal lines represent results for the first approximation $(r \to \infty) \pm$ three standard deviations for adaptive weighting (upper line, at power approximately 0.56) and equal weighting (lower line, at power approximately 0.45).

TABLE II

THE PERFORMANCE OF EQUAL AND ADAPTIVE WEIGHTING FUSION METHODS ON ENRON EMAIL GRAPHS. FOR EXAMPLE, WHEN THE FUSION DIMENSION $d' = 2$, THE POSSIBLE NUMBER OF COMBINATION OF FEATURE DIMENSIONS IS 36, AND BOTH EQUAL AND ADAPTIVE WEIGHTING METHODS CAN DETECT 24 CASES, BUT ONLY ADAPTIVE WEIGHTING CAN DETECT 5 ADDITIONAL CASES.

d'	1	2	3	4	5	6	7	8	9
$\binom{9}{d'}$	9	36	84	126	126	84	36	9	1
both	6	24	65	106	116	81	36	9	1
equal	0	0	0	0	0	0	0	0	0
adapt	0	**5**	**10**	**15**	**9**	**3**	0	0	0

exact (realistic but intractable) model. Furthermore, to the extent possible, we may tentatively conclude that model investigations have some bearing on real data applications.

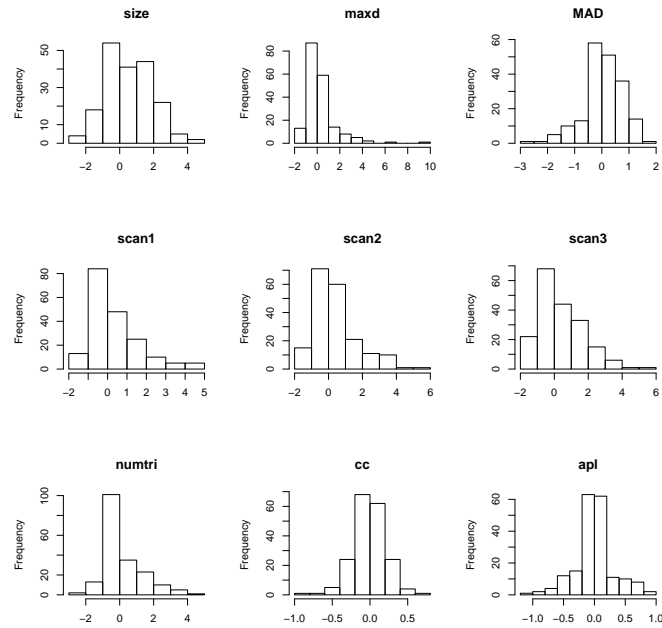An important extension of this work will be to time series of *weighted* and/or *attributed* graphs,

Fig. 11. Enron email data histograms of $S_i(t)$ for 189 weeks.
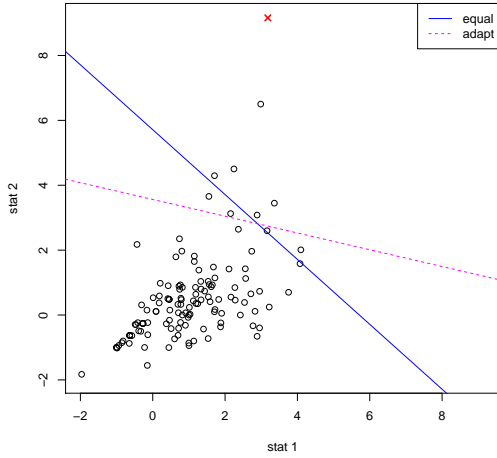
where message count and/or content is used to augment edges with (categorical) "topic" attributes [18]–[20] where authors demonstrated that using content and context together provides superior inferential capability when compared to either alone for a number of inferential tasks. Along with the fusion technique introduced in this paper, changes in communication *content*, in addition to excessive communication probability, can aid detection.
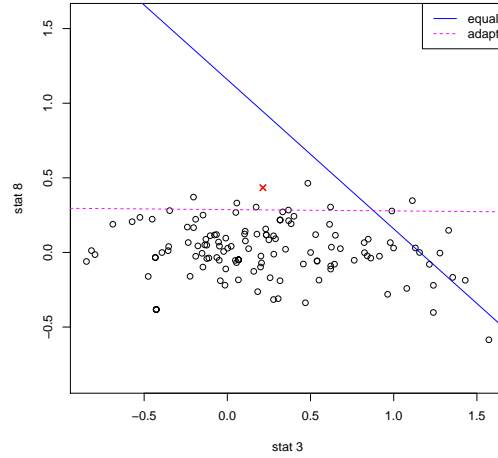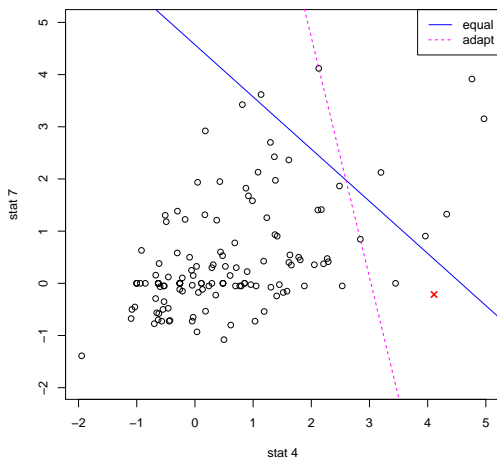
ACKNOWLEDGMENT

REFERENCES

[1] N. H. Lee and C. E. Priebe, "A Latent Process Model for Time Series of Attributed Random Graphs," *Statistical Inference for Stochastic Processes*, vol. 14, no. 3, pp. 231–253, 2011.

[2] E. R. Scheinerman and K. Tucker, "Modeling Graphs Using Dot Product Representations," *Computational Statistics*, vol. 25, pp. 1–16, January 2010. [Online]. Available: http://dx.doi.org/10.1007/s00180-009-0158-8
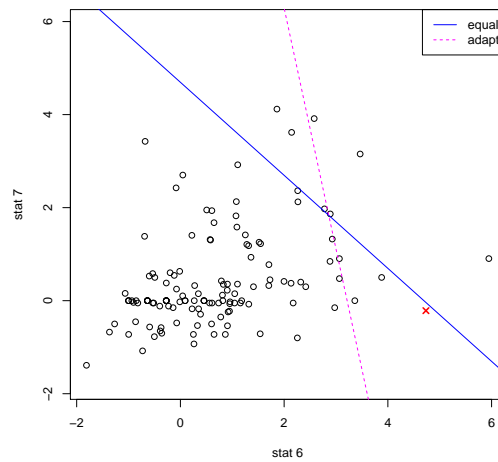
(a) size vs. maxd

(b) MAD vs. cc

(c) scan1 vs. numtri

(d) scan3 vs numtri

Fig. 12. Enron email data scatter plots of $S_i(t)$ for $t = \{1, \ldots, 132\}$ for various pairs of invariants. $S_i(t^*)$ is shown in red. The red point is above both critical lines in panel a, indicating that the combination of size and maximum degree allows detection based on $S_i(t^*)$ for both weighting schemes. In panels b,c,d, it is apparent that only the adaptive weighting scheme detects the anomaly. Unlike Figure 7, the detection boundaries for the adaptive weighting is linear, and it is because there is only one $S_i(t^*)$ graph.

[3] S. J. Young and E. R. Scheinerman, "Random Dot Product Graph Models for Social Networks," *Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph*, pp. 138–149, 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1777879.1777890

[4] B. Bollobás, S. Janson, and O. Riordan, "The Phase Transition in Inhomogeneous Random Graphs," *Random Structures and Algorithm*, vol. 31, pp. 3–122, 2007.

[5] P. Hoff, A. E. Raftery, and M. S. Handcock, "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, vol. 97, pp. 1090–1098, 2002.

[6] B. Bollobás, *Random Graphs*, 2nd ed.   Cambridge University Press, 2001.

[7] 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, October 2005.

[8] H. Pao, G. A. Coppersmith, and C. E. Priebe, "Statistical Inference on Random Graphs: Comparative Power Analyses via Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 20, no. 2, pp. 395–416, 2011.

[9] 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, December 2010.

[10] T. Ide and H. Kashima, "Eigenspace-based anomaly detection in computer systems," in *Proceedings of the Tenth ACM SIGDD International Conference on Knowledge Discovery and Data mining*, 2005, pp. 440–449.

[11] B. A. Miller, M. S. Beard, and N. T. Bliss, "Matched filtering for matched filtering for subgraph detection in dynamic networks," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, 2011.

[12] N. Borges, G. A. Coppersmith, G. G. L. Meyer, and C. E. Priebe, "Anomaly detection for random graphs using distributions of vertex invariants," in *2011 45th Annual Conference on Information Sciences and Systems (CISS)*, March 2011, pp. 1–6.

[13] J. Neil, C. Storlie, C. Hash, A. Brugh, and M. Fisk, "Scan statistics for the online detection of locally anomalous subgraphs," *Technometrics (in review)*, 2012.

[14] C. Horn and R. Willett, "Online anomaly detection with expert system feedback in social networks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 1936–1939.

[15] J. Sharpnack, A. Rinaldo, and A. Singh, "Changepoint detection over graphs with the spectral scan statistic," *arXiv/1206.0773*, 2012.

[16] M. Valko, "Adaptive graph-based algorithms for conditional anomaly detection and semi-supervised learning," Ph.D. dissertation, University of Pittsburgh, 2011.

[17] D. Ullman and E. R. Scheinerman, *Fractional Graph Theory*.   Wiley, 1997.

[18] J. Grothendieck, C. E. Priebe, and A. L. Gorin, "Statistical Inference on Attributed Random Graphs: Fusion of Graph Features and Content," *Computational Statistics and Data Analysis*, vol. 54, pp. 1777–1790, 2010.

[19] C. E. Priebe, Y. Park, D. J. Marchette, J. M. Conroy, J. Grothendieck, and A. 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.

[20] M. Tang, Y. Park, N. H. Lee, and C. E. Priebe, "Attribute fusion in a latent process model for time-series of graphs," 2012, submitted.