Computational Statistics and Data Analysis 54 (2010) 1766–1776

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Statistical inference on attributed random graphs: Fusion of graph features and content: An experiment on time series of Enron graphs

Carey E. Priebe [a,*], Youngser Park [a], David J. Marchette [b], John M. Conroy [c], John Grothendieck [d], Allen L. Gorin [e]

[a] Johns Hopkins University, United States
[b] Naval Surface Warfare Center, United States
[c] IDA Center for Computing Sciences, United States
[d] BBN Technologies, United States
[e] U.S. Department of Defense, United States

ARTICLE INFO

Article history:
Received 8 June 2009
Received in revised form 10 January 2010
Accepted 10 January 2010
Available online 2 February 2010

Keywords:
Time series analysis
Clustering
Metadata
Feature representation
Statistical methods
Graph theory

ABSTRACT

Fusion of information from graph features and content can provide superior inference for an anomaly detection task, compared to the corresponding content-only or graph feature-only statistics. In this paper, we design and execute an experiment on a time series of attributed graphs extracted from the Enron email corpus which demonstrates the benefit of fusion. The experiment is based on injecting a controlled anomaly into the real data and measuring its detectability.

1. Introduction

Let $G_1, G_2, \ldots$ be a time series of attributed graphs, with each graph $G_t = (V, E_t)$; that is, all graphs are on the same vertex set $V = [n]$ and the attributed edges at time t, denoted by E_t, are given by triples $(u, v, \ell) \in V \times V \times \mathcal{L}$, where $\mathcal{L}$ is a finite set of possible categorical edge attributes. We think of $(u, v, \ell) \in E_t$ as a *communication* or *message* from vertex u to vertex v at time t. The ordered pair (u, v) represents the *externals* of the communication which has *content* labeled by *topic* $\ell \in \mathcal{L}$, the *topic set*. The *graph features* are extracted from the overall collection of externals $\{E_t\}_{t=1,2,\ldots}$. (We are considering *directed* graphs, so the (u, v) are *ordered* pairs.)

We present here an experiment using the Enron email corpus (WWWa,b,c) which demonstrates that a statistic which *combines* content and graph feature information can provide superior inference compared to the corresponding content-only or graph feature-only statistics. Furthermore, we characterize empirically the alternative space in terms of comparative power for joint vs. marginal inference—when does a combined statistic offer more power, and when is it in fact *less* powerful.

Our purpose is to begin investigation of fusion of graph features and content for statistical inference on attributed random graphs. Grothendieck et al. (2010) presents theoretical results and simulations. This paper presents an experiment on real data, based on injecting a controlled anomaly into the data and measuring its detectability as a function of its parameters,

* Corresponding address: Johns Hopkins University, Whiting School of Engineering, 21218-2682 Baltimore, MD, United States. Tel.: +1 410 516 7200; fax: +1 410 516 7459.
E-mail address: cep@jhu.edu (C.E. Priebe).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.008

and demonstrates an anomaly detection scenario in which a graph feature-only and a content-only statistic can be fused into a combination statistic with understandable properties in the space of alternatives characterized by a small collection of vertices changing their within-group activity (in terms of graph features, content, or both) while the majority of vertices continue with their null behavior.

2. Hypotheses

The purpose of our inference is to detect a local behavior change in the time series of attributed graphs. In particular, we wish to consider as our alternative hypothesis that a small (unspecified) collection V^A of vertices increase their within-group activity at some (unspecified) time t^* as compared to recent past (an anomaly based on graph feature information) and that the content of these additional communications is also different as compared to recent past. The null hypothesis, then, is some form of temporal homogeneity—no probabilistic behavior changes in terms of either graph features or content. See Fig. 1.

We do not utilize an explicit null hypothesis. Fusion of graph features and content for statistical inference on attributed random graphs is considered in a theoretical setting in Grothendieck et al. (2010); in that work, simple null models are posited and inference for an alternative of the form depicted in Fig. 1 is considered. However, as is the case in many real communication graph applications, the Enron email data set is not the result of a designed experiment, and any reasonably simple model for time series of attributed random graphs is likely to be inappropriate. Instead, we consider an "empirical null", meaning that recent past is used to characterize baseline activity. We assume that there is no anomaly of the type represented by our alternative hypothesis in the time series under consideration – that every vertex is in its default stochastic state (short-time stationarity) during the time up until the current time t – and our goal at time t is to detect deviations of the type represented by our anomaly alternative from this baseline activity. (For example, in the Enron email corpus, Figure 1 of Priebe et al. (2005) shows that the total number of pairs of vertices which communicate during a given week increases dramatically from 1998–2000, and drops off dramatically after 2001; during 2001 this one measure of homogeneity might be considered relatively stable. Similarly, Fig. 2 shows that all three of the graph feature, content, and fusion statistics considered in this paper are also relatively stable during 2001—no anomaly is detected.)

This insistence on considering an empirical null, and avoiding an explicit null model, allows us to sidestep the inevitable model-mismatch criticisms. It does, however, lead to fundamental problems when attempting to assess and compare inferential procedures. Our approach will be to assume that the real data is "operating under the null" – and none of our statistics suggest otherwise – and to estimate rejection thresholds (critical values) for our tests by using a fixed-length sliding window, as is often done for short-time stationary signals (Rabiner and Schafer, 1978). We then seed an anomaly of interest into the data. The strength of the anomalous signal required for rejection for each of the statistics under consideration provides a useful comparison of detection power.

3. Statistics

Recall that for a graph $G = (V, E)$ and a vertex $v \in V$, the closed one-neighborhood of v in G is the collection of vertices given by $N_1[v; G] = \{w \in V : d_G(v, w) \leq 1\} \subset V$, where d_G denotes the graph distance—the length of the shortest path between two vertices. Furthermore, given $V' \subset V$, the subgraph in G induced by V' is denoted by $\Omega(V'; G)$ and is the graph (V', E') where $E' = \{(u, v) \in E : u, v \in V'\}$; that is, the induced subgraph $\Omega(V'; G)$ has vertices V' and all edges from G connecting any pair of those vertices. Therefore, $\Omega(N_1[v; G]; G)$ denotes the subgraph in G induced by the closed one-neighborhood of v.

3.1. Graph features

For each vertex v and time t, let

$$G_t(v) = \Omega(N_1[v; G_t]; G_t).$$

This *locality region* $G_t(v)$ is the subgraph of G_t induced by the closed one-neighborhood in G_t of vertex v.

Let

$$\Psi_t(v) = \text{size}(G_t(v)),$$

where the graph invariant *size* denotes the number of edges; thus $\Psi_t(v)$ represents a *local activity estimate* for vertex v at time t—the number of ordered pairs of vertices in v's neighborhood which communicate with each other at time t.

Due to potentially different overall activity levels amongst vertices, we let

$$\widetilde{\Psi}_t(v) = (\Psi_t(v) - \widehat{\mu}_t(v))/\widehat{\sigma}_t(v),$$

where $\widehat{\mu}_t(v)$ and $\widehat{\sigma}_t(v)$ are vertex-dependent normalizing parameter estimates – sample mean and variance – based on recent past. (Here, for "recent past", we use times $t - \Delta_1, \ldots, t - 1$.) For each vertex v, then, $\widetilde{\Psi}_t(v)$ represents the *normalized activity estimate* for message activity local to v (see Priebe et al. (2005) for details).

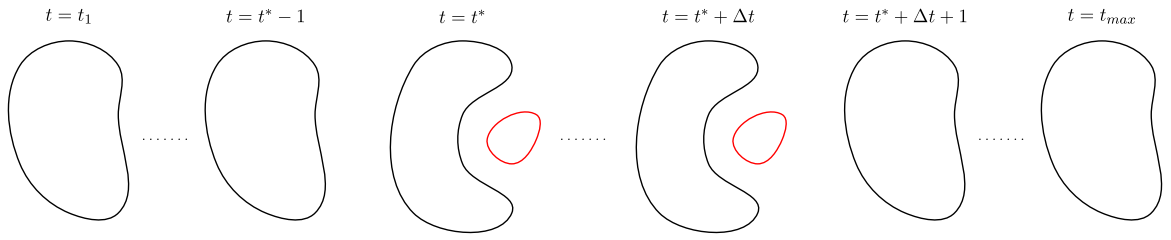

Fig. 1. Conceptual depiction of our "anomaly" alternative hypothesis for time series of random graphs. The purpose of our inference is to detect local behavior changes in the time series. The null hypothesis is "homogeneity in time"—no probabilistic changes in terms of the graph features or content behavior in time. The first collection of blobs, from time $t = t_1$ through time $t = t^* - 1$, represent the collection of vertices V behaving the same through time. At time t^*, a small collection of vertices (the small "egg", V^A) change their within-group activity (in terms of graph features, content, or both), while the majority of vertices (the large "kidney") continue with their null behavior. This local alternative behavior endures for some amount of time, and then the anomalous vertices revert to null behavior.

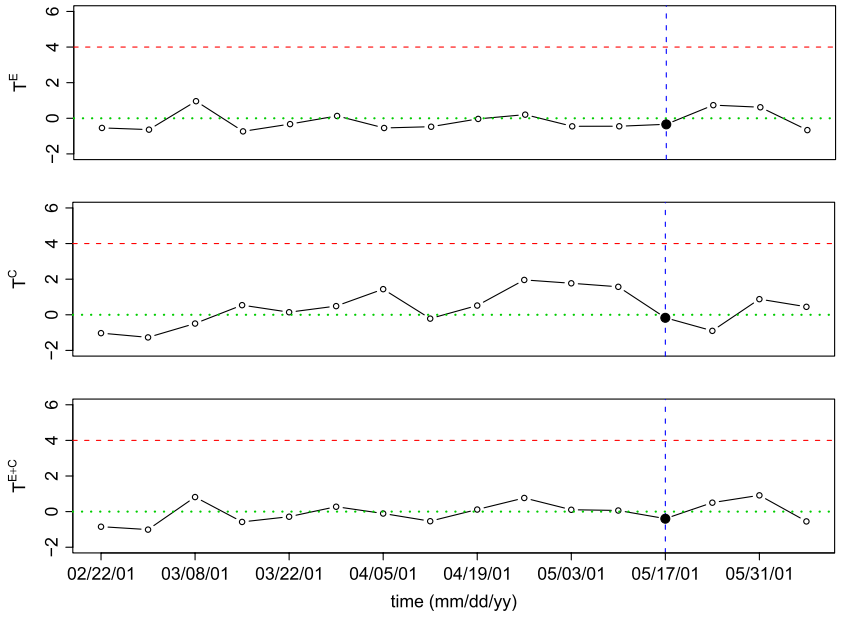

Fig. 2. The three statistics T_t^E, T_t^C, and $T_t^{E\&C}$, plotted as functions of time for the collection of weeks under consideration. (These are normalized versions of the three statistics, so that their mean is zero (dotted horizontal line) and standard deviation is one. The dashed horizontal line represents the critical value of $c_{t^*} = \widehat{\mu} + a\widehat{\sigma}$ with $a = 4$.) This figure indicates that no anomaly is detected during this time period, for any of the three statistics. Furthermore, note that all three statistics take their observed values very near their mean at time t^* (the third week of May 2001, depicted by the large dots and vertical line).

Finally, consider the *graph feature statistic* T_t^E given by

$$T_t^E = \max_v \widetilde{\Psi}_t(v);$$

a large value of T_t^E suggests that there is *excessive communication activity* in some neighborhood at time t.

(NB: T_t^E is precisely the vertex-normalized scan statistic from Priebe et al. (2005).)

3.2. Content

For each vertex v and time t, again let $G_t(v) = \Omega(N_1[v; G_t]; G_t)$.

Given the topic set $\mathcal{L}$ of cardinality K for the messages under consideration, let the vector $\widehat{\theta}_t(v)$ be the local *topic estimate* – the K-nomial parameter vector estimate, an element of the simplex $\{z \in \mathbb{R}_+^K : \|z\|_1 = 1\}$ – for the collection of messages associated with $G_t(v)$; $\widehat{\theta}_t(v)$ represents the proportions of messages in the locality region $G_t(v)$ which are deemed to be about each of the various topics. (NB: Clearly, this "aboutness" assessment requires non-trivial text processing and classification; these issues are, however, beyond the scope of present concerns.)

Then we consider the statistic

$$T_t^C = \sum_v I\left\{\arg\max_k \widehat{\theta}_t(v) \neq \arg\max_k \widehat{\theta}_{t-1}(v)\right\}.$$

T_t^C counts the number of vertices which see a change in their main topics from time $t-1$ to time t, and thus a large value of T_t^C suggests that *an excessive number of actors experienced a change in their dominant topic* for the messages associated with their neighborhood at time t.

Under the null, we assume that a certain amount of topic changing is normal—there is some tendency to stay on a topic for a period of time, and there is some (unknown) distribution on the number of individuals who change topics from one time to the next. One might posit some "change process" and proceed to fit the model to the data. Instead, we are interested in detecting these changes without regard to a particular null model.

3.3. Graph features and content combined

We have identified statistics which are appropriate for testing our null hypothesis of homogeneity against our alternative – T_t^E which uses graph features only, and T_t^C which uses local content. We do not claim that these are necessarily the best such statistics – the situation is far too complicated to derive uniformly best tests even if they did exist – but these statistics suffice to illustrate our point regarding fusion of graph features and content. To that end, we consider the combined "fusion" statistic

$$T_t^{E\&C} = g(T_t^E, T_t^C)$$

for some g.

The fusion statistic $T_t^{E\&C}$ combines the information concerning excessive communication activity in some neighborhood at time t compared to recent past and excessive number of actors changing their dominant topic from time $t-1$ to time t. An anomaly at time t of the type described by our alternative hypothesis should, under some circumstances, be detectable using $T_t^{E\&C}$ even though neither T_t^E nor T_t^C detect. This we shall demonstrate, in the sequel.

4. Experimental design

The time series of Enron email graphs $G_1, G_2, \ldots$ is assumed to be observed "under the null"; that is, we assume that no anomaly is present during this time period, for any of the three statistics developed in the previous section. We choose a time t^* and consider *seeding* the graph G_{t^*} with anomalous behavior so that the seeded graph $\widetilde{G}_{t^*}$ behaves as "under the alternative".

4.1. Critical values

For each time t in $\{t^* - \Delta_2, \ldots, t^* - 1\}$ we evaluate the three statistics T_t^E, T_t^C, and $T_t^{E\&C}$. Since this time period is assumed to be "under the null", we use these values to naively calculate the sample mean and variance and an estimated critical value for each statistic; since all three statistics reject for large observed values, we use $c_{t^*} = \widehat{\mu} + a\widehat{\sigma}$ for each of T_t^E, T_t^C, and $T_t^{E\&C}$ in turn.

This approach to setting the critical values is ad hoc, but principled. Chebyshev's inequality suggests that, for $a = 4$ for example, all three tests should have a significance level in the range $[0, 1/16]$. This range is only approximate, because we are estimating the mean and standard deviation, and assuming (short-time) stationarity. Nevertheless, without specifying an explicit null model (which we are loath to do for this Enron email data set) this sliding window analysis provides a useful first approximation for comparative power analysis (Rabiner and Schafer, 1978).

4.2. Anomaly seeding

Our seeding procedure for time t^* proceeds as follows:

To model *additional communication activity*, we choose a value $q \in (0, 1]$ which represents the probability that a new edge will be added to G_{t^*} as described below.

To model *altered topic behavior*, we choose a K-nomial parameter vector θ' which represents the topic distribution for a newly added edge.

An anomaly graph $\widetilde{G}_{t^*}$ based on q, θ', and $m > 1$ selected vertices $V^A \subset V$ is constructed from G_{t^*} such that for each of the $m(m-1)$ ordered pairs from V^A, if there is not already an edge in G_{t^*}, an edge is added with probability q and this added edge is given a topic value drawn according to θ'.

The effect of this anomaly seeding depends on the attributes and graph features of the selected vertices. A "coherent signal" (constrained to the small set of vertices V^A) with parameters q, θ', and m is added based on a particular choice of V^A. Which V^A is used strongly effects the detectability of the signal. Hence, integration over the choice of V^A is necessary. Monte Carlo is employed.

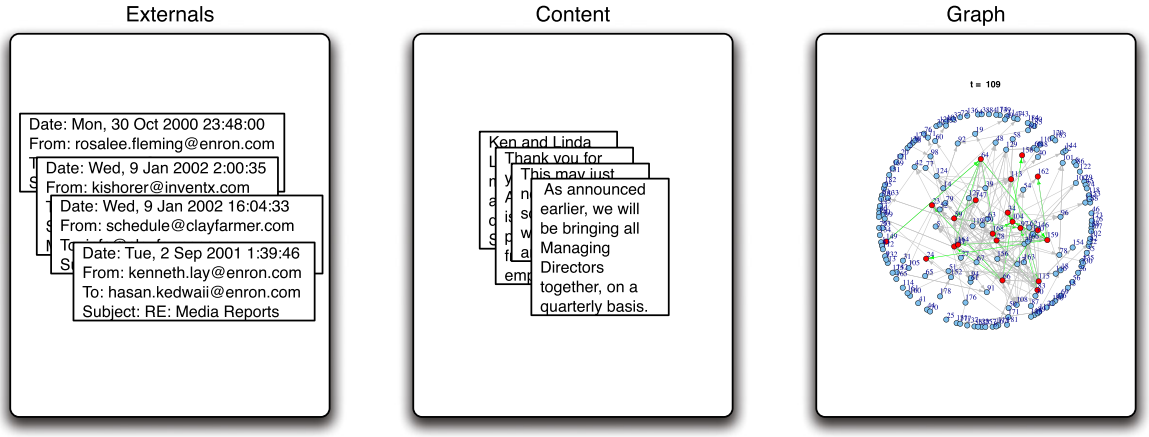

Fig. 3. Conceptual depiction of the Enron email corpus data. Each message includes externals (to, from, date) and content (the text of the email). In addition to information derivable solely from a given message's externals, the overall time series of communications graphs gives rise to graph features relevant to the exploitation task.

4.3. Monte Carlo

Given the time series of graphs $G_1, G_2, \ldots$ and t^*, q, θ', m, our Monte Carlo experiment proceeds as follows:

For each Monte Carlo replicate r, we *seed* G_{t^*} with an anomaly, randomly selecting m vertices and adding "messages" – new edges with topic attributes – as described above. We calculate $T^E_{t^*,r}$, $T^C_{t^*,r}$, and $T^{E\&C}_{t^*,r}$, and for each statistic we check whether its seeded value exceeds its associated critical value $c^E_{t^*}$, $c^C_{t^*}$, and $c^{E\&C}_{t^*}$; that is, we evaluate the rejection indicator $I\{T_{t^*,r} > c_{t^*}\}$ for each of the three statistics for each Monte Carlo replicate $r = 1, \ldots, R$.

The result of the Monte Carlo experiment – estimated probabilities of detection $\widehat{\beta}^E$, $\widehat{\beta}^C$, and $\widehat{\beta}^{E\&C}$, where $\widehat{\beta} = (1/R)\sum_{r=1}^{R} I\{T_{t^*,r} > c_{t^*}\}$ for each of the three statistics – provides information concerning the sensitivity of the statistics to the signal parameterized by q, θ', and m.

5. Experiment

For the experiment, we use the Enron email corpus. Fig. 3 presents a conceptual depiction of the data, and Fig. 4 presents a conceptual depiction of the experiment.

The times considered are a collection of weeks in 2001, for which short-time stationarity is at least plausible. There are $n = |V| = 184$ vertices and a total of 28,266 messages during this period. Fig. 2 shows that no anomaly is detected during this time period, for any of the three statistics T^E_t, T^C_t, or $T^{E\&C}_t$. We choose t^* to be the third week of May 2001. All three statistics take their observed value very near their running mean at t^*, so that none of the statistics enjoys an obvious advantage when it comes to detecting the seeded anomaly.

For vertex-dependent normalization of T^E, and for critical value estimation, we choose lags $\Delta_1 = \Delta_2 = 20$ to identify "recent past". These time-window constants should be as large as possible, so long as short-time stationarity can be safely assumed.

To identify the topic set $\mathcal{L}$ of cardinality K, we first choose $\mathcal{M} =$ {50 messages randomly selected from each of the weeks under consideration}. We cluster the collection of messages $\mathcal{M}$ into K clusters (see Appendix) and identify each resulting cluster with a topic, thus providing the topic set $\mathcal{L} = \{1, \ldots, K\}$. For all of the experiments presented herein, we use $K = 3$. (The choice of $K = 3$ topics is for illustrative purposes; we make no claim that the email message content is best modeled as three topics. For instance, (Berry et al., 2007) manually index a subset of the Enron email messages into 32 topics. Our choice of $K > 2$ avoids degeneracy from multinomial to binomial, and K small avoids unnecessary complexity, for our demonstration.)

Note that the clustering of $\mathcal{M}$ described above gives rise to a *classification* of each message in $\mathcal{M}$ – each message is labeled with the topic estimate given by the cluster into which that message falls. Next, for the messages not in $\mathcal{M}$, we employ out-of-sample embedding and nearest neighbor classification using the now-labeled set $\mathcal{M}$ (see Appendix). For T^C, to identify the *topic estimate* $\widehat{\theta}_t(v)$ – the K-nomial parameter vector estimate for the collection of messages associated with $G_t(v)$ – we use the sample proportions from the now-classified messages.

To model *altered topic behavior*, we first set $\theta = [\theta_1, \ldots, \theta_K]^T$ to be the K-nomial parameter vector estimate obtained from the classification described above; we obtain $\theta = [0.391, 0.184, 0.425]^T$. We identify the altered topic vector θ' based on θ and a single parameter δ, so that $\theta' = h_\theta(\delta)$, by setting θ' to be the K-nomial parameter vector obtained by increasing the smallest element of θ by some $\delta \leq (K-1)/K$ and decreasing the other $K-1$ elements of θ by $\delta/(K-1)$, so $\theta'_k = \theta_k + \delta$

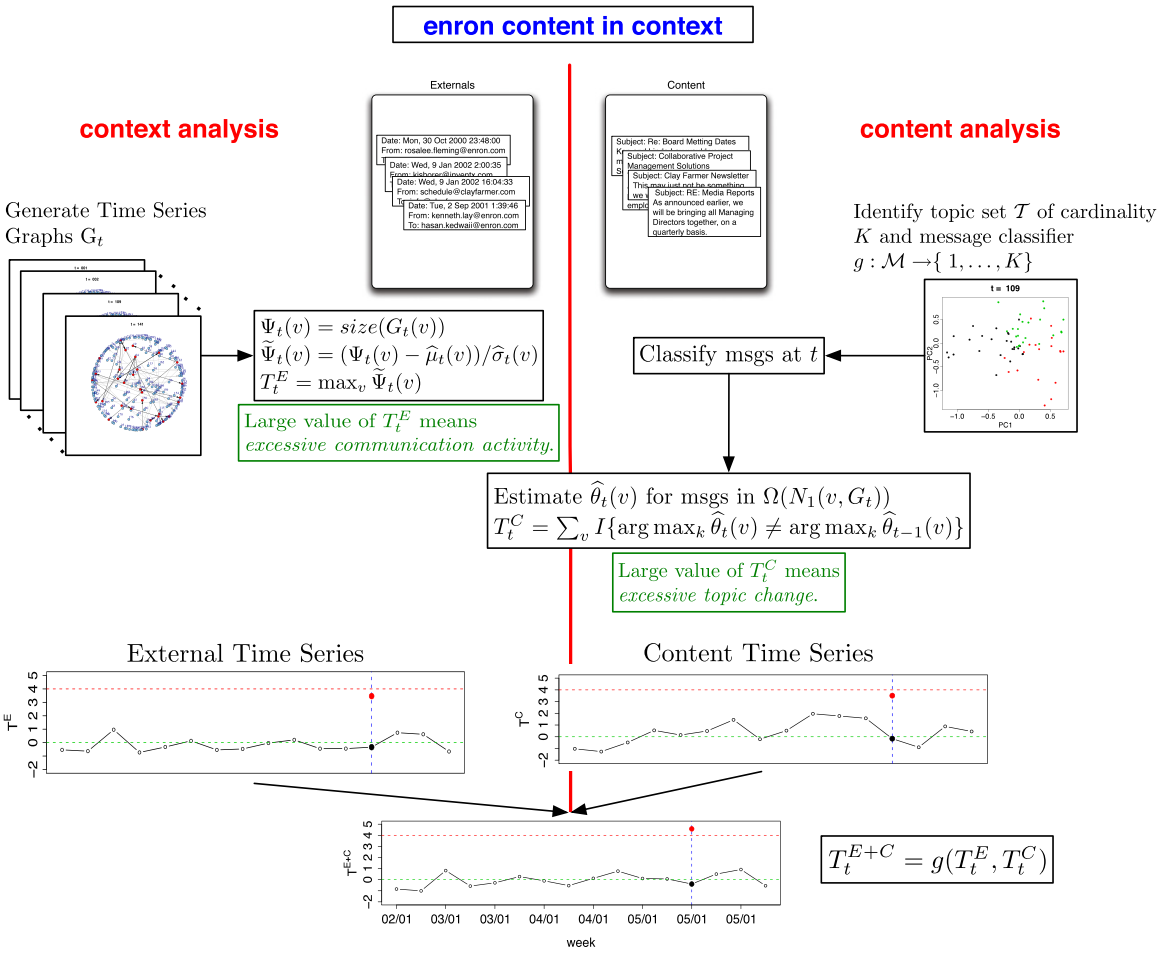

Fig. 4. Conceptual depiction of the experiment described in Section 5.

for $k = \arg\min_k \theta_k$ and $\theta'_k = \theta_k - \delta/(K-1)$ for $k \neq \arg\min_k \theta_k$. Thus our altered topic vector θ' is obtained from θ via a single parameter $\delta \in (0, (K-1)/K]$.

To model *additional communication activity*, we choose $q \in (0, 1]$.

The alternative space for our seeded messages, then, is $(\delta, q) \in \Theta_A = (0, (K-1)/K] \times (0, 1]$.

For the combined statistic, we use a linear combination with parameter γ;

$$T_t^{E\&C} = g(T_t^E, T_t^C) = \gamma T_t^E + (1-\gamma) T_t^C.$$

This simple form for the fusion statistic is a first consideration only; optimal fusion need not take this form (see Grothendieck et al. (2010)).

6. Experimental results

Fig. 2 shows that all three statistics, normalized to have mean zero and standard deviation one, take their observed values at time t^* very near their mean. We conclude that the three tests, using critical values $c_{t^*} = \widehat{\mu} + a\widehat{\sigma}$ with $a = 4$, have approximately the same α-level, and so power comparisons under our anomaly seeding scheme are at least illustrative.

Fig. 5 depicts illustrative results from one Monte Carlo replicate, with $q = 0.5$ (additional communication activity), $\delta = 0.6$ (altered topic behavior), $\gamma = 0.5$ (combination coefficient), and $m = 13$ (number of anomalous vertices). The large squares indicate that this particular random seeding replicate adds signal for all three statistics. While the added signal is not sufficiently strong to cause a null hypothesis rejection for either T^E or T^C alone, the fusion statistic $T^{E\&C}$ does reject for the same added signal.

Figs. 6 and 7 present results from the full Monte Carlo experiment. We use $R = 1000$ Monte Carlo replicates throughout.

Fig. 6 depicts power curves, probability of detection as a function of m, with $q = 0.5$, $\delta = 0.6$, $\gamma = 0.5$. We see that when m is small (less than 9) all three statistics have power approximately zero; when m is greater than 18 all three statistics have power approximately one. For moderate m (between five and ten percent of the total 184 vertices, or $m \approx \sqrt{n}$ as suggested

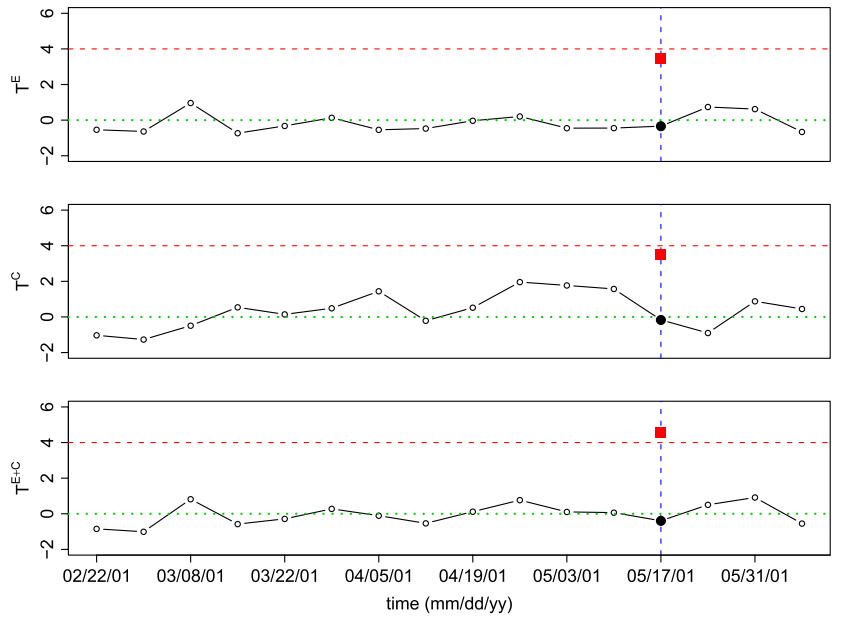

Fig. 5. Illustrative results from one Monte Carlo replicate, with $q = 0.5$, $\delta = 0.6$, $\gamma = 0.5$ and $m = 13$. Normalized versions of the three statistics T_t^E, T_t^C, and $T_t^{E\&C}$, plotted as functions of time, with their running mean at zero (dotted horizontal line) and critical value of $c_{t^*} = \widehat{\mu} + a\widehat{\sigma}$ with $a = 4$ (dashed horizontal line), as in Fig. 2. The large dots indicate that all three statistics take their observed values very near their mean at time t^*. The large squares indicate that this particular random seeding replicate adds signal for all three statistics, but that while the added signal is not sufficiently strong to cause a null hypothesis rejection for either T^E or T^C alone, the fusion statistic $T^{E\&C}$ does reject for this same instantiation of added signal. (Figs. 6 and 7 present results from the full Monte Carlo experiment.).

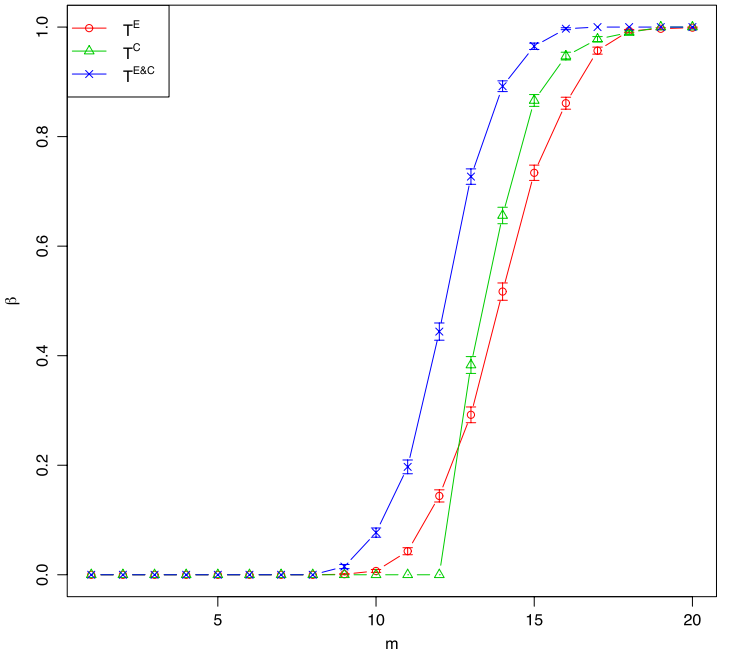

Fig. 6. Results from Monte Carlo experiment. Depicted are power curves as a function of m, with $q = 0.5$, $\delta = 0.6$ and $\gamma = 0.5$. When m is small (less than 9) all three statistics have power approximately zero; when m is greater than 18 all three statistics have power approximately one. For moderate m (between five and ten percent of the total 184 vertices, or $m \approx \sqrt{n}$) the fusion statistic $T^{E\&C}$ yields a more powerful test than either T^E or T^C alone. In particular, for $m = 13$, $q = 0.5$, $\delta = 0.6$ and $\gamma = 0.5$ we obtain $\widehat{\beta}^{E\&C} \approx 0.75 > \max\{\widehat{\beta}^E \approx 0.29, \widehat{\beta}^C \approx 0.39\}$.

in Grothendieck et al. (2010)) the fusion statistic $T^{E\&C}$ yields a more powerful test than either T^E or T^C alone. In particular, for $m = 13$, $q = 0.5$, $\delta = 0.6$, $\gamma = 0.5$ we obtain $\widehat{\beta}^{E\&C} \approx 0.75 > \max\{\widehat{\beta}^E \approx 0.29, \widehat{\beta}^C \approx 0.39\}$.

Fig. 7 presents Monte Carlo power estimates as a function of q and δ with $m = 13$, $\gamma = 0.5$. We see that for small q and small δ, none of the three statistics have good power characteristics, while for large q the statistic T^E is appropriate.

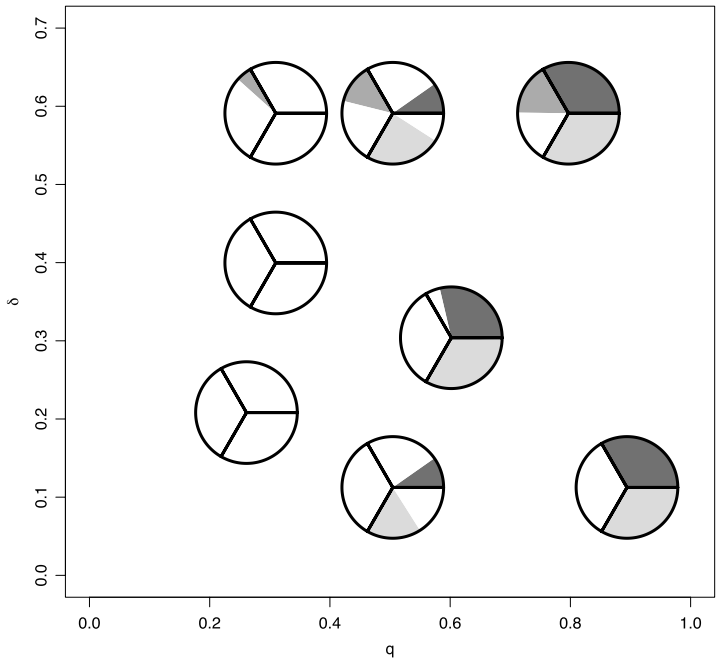

Fig. 7. Results from Monte Carlo experiment. Depicted are power values as a function of q and δ with $m = 13$ and $\gamma = 0.5$. Each pie has three thirds, with one third representing each of the three statistics and the amount of that third which is filled in representing the power – lightest gray for $\widehat{\beta}^{E\&C}$, darkest gray for $\widehat{\beta}^{E}$, and medium gray for $\widehat{\beta}^{C}$ – for the point (q, δ) in the alternative space denoted by the center of the pie. (The top center pie corresponds to Fig. 6 with $m = 13$.) We see that for small q and small δ, none of the three statistics have good power characteristics, while for large q the statistic T^E is appropriate. For moderate q, the fusion statistic $T^{E\&C}$ yields a more powerful test than either T^E or T^C alone.

For moderate q, the fusion statistic $T^{E\&C}$ can yield a more powerful test than either T^E or T^C alone. Note, however, that for $q = 0.3, \delta = 0.6$ the content-only statistic is most powerful–$\widehat{\beta}^{E\&C} \approx 0.01, \widehat{\beta}^{E} \approx 0.00, \widehat{\beta}^{C} \approx 0.15$ (Nowhere in this plot do we see T^E being most powerful. Such alternatives do exist, however; for example, $q = 0.5, \delta = 0.1, m = 13, \gamma = 0.1$ yields $\widehat{\beta}^{E\&C} \approx 0.12, \widehat{\beta}^{E} \approx 0.29, \widehat{\beta}^{C} \approx 0.00$.)

Figs. 8 and 9 depict an investigation of the best fusion statistic in the class $\{\gamma T_t^E + (1-\gamma)T_t^C : \gamma \in (0,1)\}$ for various signal parameters q, δ with $m = 13$. We see that the specific anomaly parameters q, δ strongly influence the optimal γ, as expected.

These results, taken as a whole, indicate clearly that fusion of graph features and content in this setting can indeed provide superior power than is attainable using either graph features or content alone. We also see, however, that (as the bias-variance tradeoff suggests) for some anomalies fusion can have a negative effect, due to additional estimation variance for too little signal gain. This phenomenon indicates that, for general alternatives, more elaborate inferential procedures are necessary, rather than simply using both graph features and content just because both are available—a version of the classical "curse of dimensionality".

6.1. Incoherent signal

We have presented our experimental results in terms of comparative power, as if all three tests are constrained to have the same α-level. Since we do not posit a specific null model, this claim is at best approximately true. Nevertheless, the changing relative powers for various values of the signal parameters indicate that fusion has a variable effect and can be sometimes wise and sometimes unwise.

Another approach to the analysis considers, rather than comparative power in a null vs. alternative sense, comparative probability of detection for coherent vs. incoherent signals. In our anomaly seeding experiments, we choose signal parameters q, δ, and m, and then choose a specific collection V^A of m vertices and consider adding attributed edges for each of the $m(m-1)$ directed pairs. That is, the signal is *coherent* in that the additional communication activity and the altered topic behavior are concentrated amongst the m vertices V^A. An *incoherent* version of this anomaly seeding utilizes the same signal parameters, but considers adding attributed edges for $m(m-1)$ directed pairs randomly selected from the entire collection of $n(n-1)$ possible directed pairs. Thus we have the same amount of added signal, but it is not concentrated in one coherent subgraph.

Fig. 10 presents the results of such an "incoherent signal" Monte Carlo experiment. Depicted are power values for the combined statistic $T_{t^*}^{E\&C}$ for fixed $\delta = 0.6, q = 0.5, m = 13, \gamma = 0.5$, as a function of a in critical value $c_{t^*} = \widehat{\mu} + a\widehat{\sigma}$. Recall that we obtained $\widehat{\beta}^{E\&C} \approx 0.75 > \max\{\widehat{\beta}^{E} \approx 0.29, \widehat{\beta}^{C} \approx 0.39\}$ for the coherent signal experiment, with $a = 4$. For the

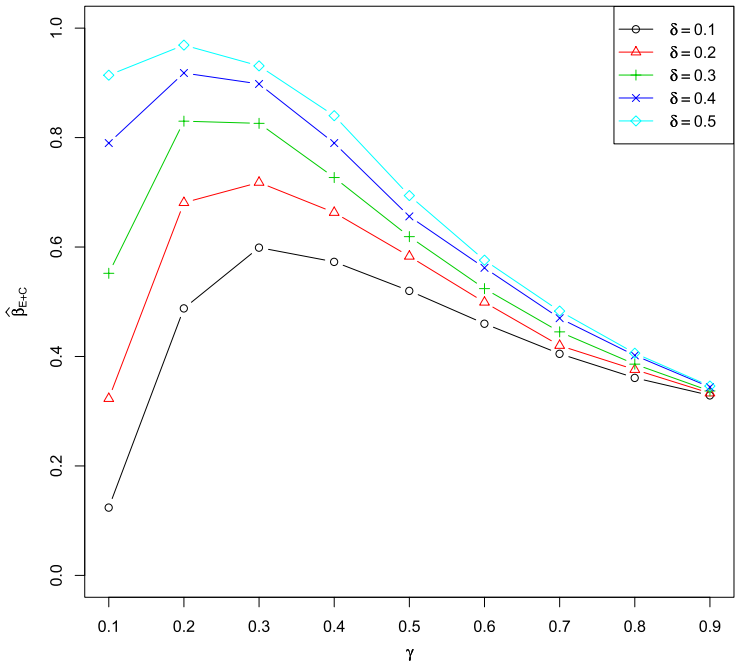

Fig. 8. Results from Monte Carlo experiment. Depicted are power values for the combined statistic $T_{t^*}^{E\&C}$ for fixed $q = 0.5$ and $m = 13$ and selected values of δ, as a function of γ.

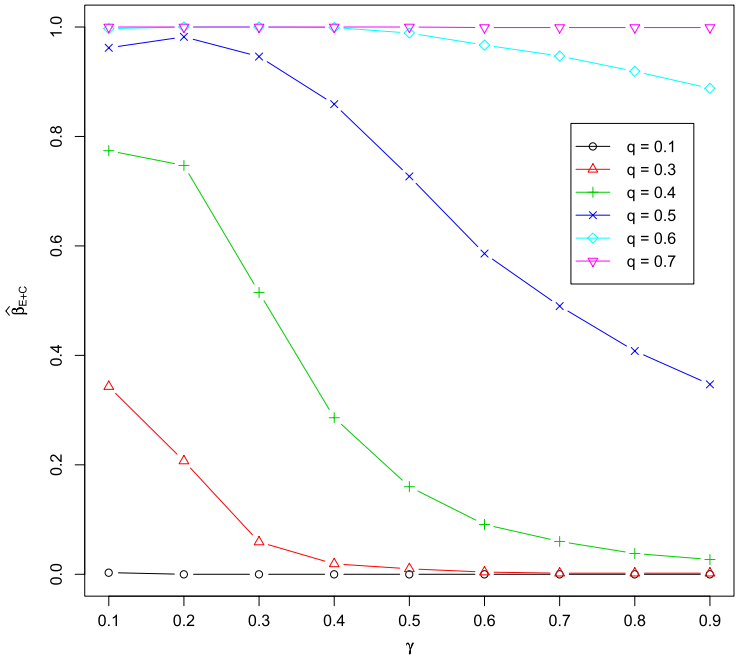

Fig. 9. Results from Monte Carlo experiment. Depicted are power values for the combined statistic $T_{t^*}^{E\&C}$ for fixed $\delta = 0.6$ and $m = 13$ and selected values of q, as a function of γ.

incoherent signal, all three statistics have an estimated power of zero for $a \geq 0.2$. Thus we conclude that it is the *coherence* of the seeded signal that is being detected throughout the foregoing experiments, and that all three statistics (with $a = 4$) have a Type I error rate of approximately zero with respect to the "empirical graph + incoherent signal" null.

7. Discussion and conclusions

The purpose of this experiment is to begin investigation of fusion of graph features and content for statistical inference on attributed random graphs. This experiment on time series of Enron graphs demonstrates explicitly, constructively, and

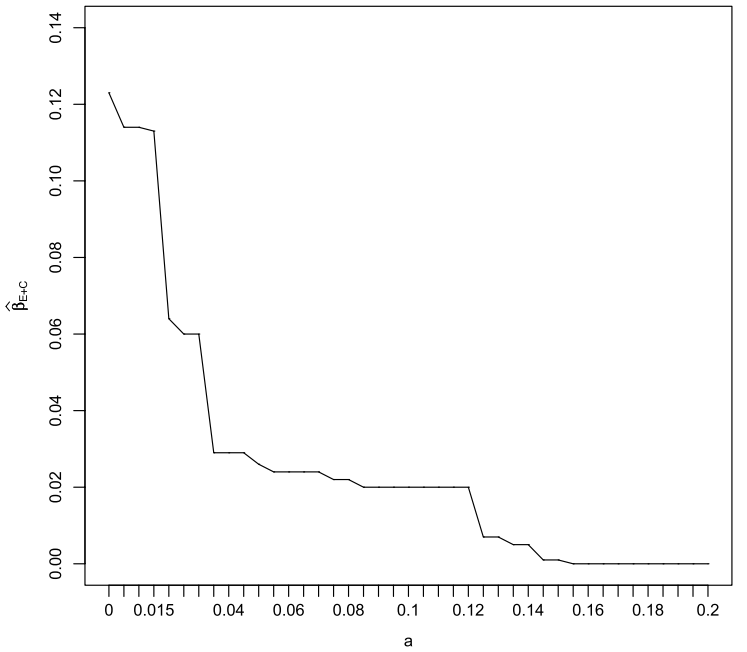

Fig. 10. Results from "incoherent signal" Monte Carlo experiment. Depicted are power values for the combined statistic $T_{t^*}^{E\&C}$ for fixed $\delta = 0.6$, $q = 0.5$, $m = 13$ and $\gamma = 0.5$, as a function of a in critical value.

quantitatively an anomaly detection scenario in which a graph feature-only and a content-only statistic can be fused into a combination statistic with understandable properties in the space of alternatives characterized by a small collection of vertices changing their within-group activity (in terms of graph features, content, or both) while the majority of vertices continue with their null behavior.

There are numerous choices that have been made, for expediency, which are in no way claimed to be optimal. Among these issues, we briefly address the most noteworthy here.

The individual statistics T^E and T^C are reasonable, but many other choices are available. As for the combined statistic, another (better?) choice is to combine locally; e.g.,

$$T_t^{E\&C} = \max_v g(\widetilde{\Psi}_t(v), \|\widehat{\theta}_t(v) - \widehat{\theta}_{t-1}(v)\|).$$

It is conjectured here that this more elaborate fusion statistic – assessing a neighborhood's excessive activity and topic change *simultaneously* – will yield dividends.

Methods exist for streaming topic estimation which are superior to the simple approach employed herein. As we are primarily interested in fusion effects, we admittedly give short shrift to this important aspect of the problem.

For our Monte Carlo seeding, we write: "For each Monte Carlo replicate r, we *seed* G_{t^*} with an anomaly, randomly selecting m vertices and adding messages …" Thus, we are integrating over the specific collection of m anomalous vertices in our experiment. Obviously, the amount of signal (graph feature, or content, or both) necessary to cause a detection depends heavily upon the specific vertices involved in the anomaly. It seems that conditional investigations along these lines would be of value. In particular, 14 of the $n = 184$ vertices are inactive during the time period under consideration. Added signal will have a dramatic effect on these isolated vertices. However, the result obtained when re-running the Monte Carlo experiment ($m = 13$, $q = 0.5$, $\delta = 0.6$, $\gamma = 0.5$) with these isolates removed is $\widehat{\beta}^{E\&C} \approx 0.73 > \max\{\widehat{\beta}^E \approx 0.29, \widehat{\beta}^C \approx 0.38\}$ — almost identical to the result with the isolates included, $\widehat{\beta}^{E\&C} \approx 0.75 > \max\{\widehat{\beta}^E \approx 0.29, \widehat{\beta}^C \approx 0.39\}$. From this we conclude that our results are robust to a large percentage of isolates — approximately 8%, in this case.

Since θ and $\theta' = h_\theta(\delta)$ are global, $\delta = 0$ is not equivalent to "no content change" for each local neighborhood. This point helps explain some of the test behavior (e.g. in Fig. 7). This may be an experimental design flaw – local null and alternative content vectors $\theta(v)$ and $\theta'(v)$ could be employed for each neighborhood – but we feel that the advantage (namely, simplicity) of this choice of a univariate content anomaly outweighs the disadvantages.

Finally, we must comment on the level of the three tests being compared. We write: "We conclude that the three tests […] have approximately the same α-level, and so power comparisons […] are at least illustrative". Of course, this assumes properties of the right tail of the null distributions of the statistics that cannot be checked to our satisfaction. Nonetheless, we argue that even if the approximate sameness of level is not accepted, the fusion effect (e.g. Fig. 6) remains, with shifts in the cut-off thresholds. Furthermore, the "incoherent signal" experiment indicates that all three tests have probability approximately zero of detecting the incoherent signal, and thus our investigation demonstrates the comparative detection effect of coherence.

These and other issues notwithstanding, our Monte Carlo experimental results shed light on the behavior of fusion of graph features and content in the specific setting investigated, augmenting theoretical results in Grothendieck et al. (2010). It is hoped that this initial experiment, together with Grothendieck et al. (2010), will provide impetus for additional, more elaborate experimental and theoretical investigations.

Appendix

To cluster the collection of training messages $\mathcal{M}$ into K clusters, we first employ term-document mutual information calculations and low-dimensional Euclidean embedding via multi-dimensional scaling (see Priebe et al. (2004)). K-means clustering is then used to identify the final clusters/topics. For all of the experiments presented herein, we use $K = 3$.

The term-document mutual information calculations can be adapted to allow efficient out-of-sample embedding of test data into the same space as the training data. This approach is used to put all messages not in $\mathcal{M}$ into the same space as $\mathcal{M}$, wherein nearest neighbor classification is possible.

We first calculate the mutual information feature vectors for documents in corpus $\mathcal{M}$. When we calculate the mutual information feature vector for each document x in the remaining test data set $\mathcal{N}$, instead of counting the total frequency of word w in corpus $\mathcal{N}$ only, we also count it in $\mathcal{M}$. That is, instead of using

$$m_{x,w} = \log\left(\frac{f_{x,w}}{\sum_{\mathcal{N}} f_{\mathcal{N},w} \sum_{w} f_{x,w}}\right),$$

we use

$$\begin{aligned}\widetilde{m}_{x,w} &= \log\left(\frac{f_{x,w}}{\sum_{\mathcal{N}\cup\mathcal{M}} f_{\mathcal{N}\cup\mathcal{M},w} \sum_{w} f_{x,w}}\right)\\ &= \log\left(\frac{f_{x,w}}{\left(\sum_{\mathcal{N}} f_{\mathcal{N},w} + \sum_{\mathcal{M}} f_{\mathcal{M},w}\right)\sum_{w} f_{x,w}}\right).\end{aligned}$$

Here $f_{x,w} = c_{x,w}/|\mathcal{N}|$ where $c_{x,w}$ is the number of times word w appears in document x and $|\mathcal{N}|$ is the total number of words in the corpus. Since we already have a list of words and their counts in $\mathcal{M}$, we can reuse that information without recalculating it.

This approach is not as accurate as calculating feature vectors in the whole corpus $\mathcal{M} \cup \mathcal{N}$, but is good enough for an approximation and also can save a significant amount of computing time.

References

www.cs.queensu.ca/home/skill/siamworkshop.html.
www-2.cs.cmu.edu/~enron.
www.cis.jhu.edu/~parky/Enron.
Berry, M.W., Browne, M., Signer, B., 2007. 2001 Topic annotated enron email data set, Linguistic Data Consortium. Philadelphia.
Grothendieck, J., Priebe, C.E., Gorin, A.L, 2010. Statistical inference on attributed random graphs: Fusion of graph features and content. Computational Statistics and Data Analysis 54, 1777–1790.
Priebe, C.E., et al. 2004. Iterative denoising for cross-corpus discovery. In: COMPSTAT 2004, pp. 381–392.
Priebe, C.E., Conroy, J.M., Marchette, D.J., Park, Y., 2005. Scan statistics on enron graphs. Computational & Mathematical Organization Theory 11, 229–247.
Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, NJ.