Computational & Mathematical Organization Theory, 11, 229–247, 2005

Scan Statistics on Enron Graphs

CAREY E. PRIEBE
Johns Hopkins University, Baltimore, MD
email: cep@jhu.edu

JOHN M. CONROY
IDA Center for Computing Sciences, Bowie, MD

DAVID J. MARCHETTE
NSWC B10, Dahlgren, VA

YOUNGSER PARK
Johns Hopkins University, Baltimore, MD

Abstract

We introduce a theory of scan statistics on graphs and apply the ideas to the problem of anomaly detection in a time series of Enron email graphs.

Keywords: Enron email data, time series of graphs, scan statistics, statistical inference, anomaly detection

1. Introduction

Consider a directed graph (digraph) D with vertex set $V(D)$ and arc set $A(D)$ of directed edges. For instance, we may think of D as a communications or social network, where the $n = |V(D)|$ vertices represent people or computers or more general entities and an arc $(v, w) \in A(D)$ from vertex v to vertex w is to be interpreted as meaning "the entity represented by vertex v is in directed communication with or has a directed relationship with the entity represented by vertex w." We are interested in testing the null hypothesis of "homogeneity" against alternatives suggesting "local subregions of excessive activity." Toward this end, we develop and apply a theory of scan statistics on random graphs.

2. Scan Statistics

Scan statistics are commonly used to investigate an instantiation of a random field X (a spatial point pattern, perhaps, or an image of pixel values) for the possible presence of a

Previous presentation: Workshop on Link Analysis, Counterterrorism and Security at the SIAM International Conference on Data Mining, Newport Beach, CA, April 23, 2005.

local signal. Known in the engineering literature as "moving window analysis", the idea is to scan a small window over the data, calculating some local statistic (number of events for a point pattern, perhaps, or average pixel value for an image) for each window. The supremum or maximum of these locality statistics is known as the scan statistic, denoted $M(X)$. Under some specified "homogeneity" null hypothesis H_0 on X (Poisson point process, perhaps, or Gaussian random field) the approach entails specification of a critical value c_α such that $P_{H_0}[M(X) \geq c_\alpha] = \alpha$. If the maximum observed locality statistic is larger than or equal to c_α, then the inference can be made that there exists a nonhomogeneity—a local region with statistically significant signal.

An intuitive approach to testing these hypotheses involves the partitioning of the region X into disjoint subregions. For cluster detection in spatial point processes this dates to Fisher's 1922 "quadrat counts" (Fisher and Mackenzie, 1922); see Diggle (1983). Absent prior knowledge of the location and geometry of potential nonhomogeneities, this approach can have poor power characteristics.

Analysis of the univariate scan process ($d = 1$) has been considered by many authors, including Naus (1965), Cressie (1977, 1980), and Loader (1991). For a few simple random field models exact p-values are available; many applications require approximations to the p-value. The generalization to spatial scan statistics is considered in Naus (1965), Adler (1984), Loader (1991), and Chen and Glaz (1996). As noted by Cressie (1993), exact results for $d = 2$ have proved elusive; approximations to the p-value based on extreme value theory are in general all that is available. Naiman and Priebe (2001) present an alternative approach, using importance sampling, to this problem of p-value approximation.

3. Scan Statistics on Graphs

The order of the digraph, $n = |V(D)|$, is the number of vertices. The size of the digraph, $|A(D)|$, is the number of arcs. For $v, w \in V(D)$ the digraph distance $d(v, w)$ is defined to be the minimum directed path length from v to w in D.

For non-negative integer k (the *scale*) and vertex $v \in V(D)$ (the *location*), consider the closed kth-order neighborhood of v in D, denoted $N_k[v; D] = \{w \in V(D) : d(v, w) \leq k\}$. We define the *scan region* to be the induced subdigraph thereof, denoted

$$\Omega(N_k[v; D]) \tag{1}$$

with vertices $V(\Omega(N_k[v; D])) = N_k[v; D]$ and arcs $A(\Omega(N_k[v; D])) = \{(v, w) \in A(D) : v, w \in N_k[v; D]\}$. A *locality statistic* at location v and scale k is any specified digraph invariant $\Psi_k(v)$ of the scan region $\Omega(N_k[v; D])$. For concreteness consider for instance the *size* invariant, $\Psi_k(v) = |A(\Omega(N_k[v; D]))|$. Notice, however, that any digraph invariant (e.g. density, domination number, etc.) may be employed as the locality statistic, as dictated by application. The "scale-specific" *scan statistic* $M_k(D)$ is given by some function of the collection of locality statistics $\{\Psi_k(v)\}_{v \in V(D)}$; consider for instance the maximum locality statistic over all vertices,

$$M_k(D) = \max_{v \in V(D)} \Psi_k(v). \tag{2}$$

This idea is introduced in Priebe (2004).

Under a null model for the random digraph D (for instance, the Erdos-Renyi random digraph model) the variation of $\Psi_k(v)$ can be characterized and $M_k(D)$ large indicates the existence of an induced subdigraph (scan region) $\Omega(N_k[v; D])$ with excessive activity. A test can be constructed for a specific alternative of interest concerning the structure of the excessive activity anticipated. However, if the anticipated alternative is, more generally, some form of "chatter" in which one (small) subset of vertices communicate amongst themselves (in either a structured or an unstructured manner) then our scan statistic approach promises more power than other approaches.

Finally, we wish to consider the scan statistic which accounts for variable scale. Let $K \subset \{1, \ldots, n-1\}$ be a collection of scales, and let Ψ'_k be a scale-standardized version of the locality statistic Ψ_k. For instance, for given $\alpha \in (0, 1)$, find $g_{k,\alpha}(\cdot)$ such that $\Psi'_k(v) = g_{k,\alpha}(\Psi_k(v))$ satisfies $P[\Psi'_k(v) \geq c_\alpha] \approx \alpha$ for all $v \in V(D)$ and for all $k \in K$. This standardization imposes upon each locality statistic the same probability of exceedance. Then the *scan statistic* $M_K(D)$ is given by

$$M_K(D) = \max_{k \in K} \max_{v \in V(D)} \Psi'_k(v) \tag{3}$$

and we reject for large values of $M_K(D)$.

For the Enron data considered in this paper, as for much social network data, no appropriate simple null random graph model is obvious. The dataset, as we process it, consist of a time series of digraphs $D_1, D_2, \ldots, D_{T=189}$. We will proceed conditionally: we will assume that the data (or the statistics derived from the data) have some short-time stationarity properties under the null, so that a moving window approach is appropriate. We will be concerned with discovering anomalies that appear as digraphs which differ substantially from those seen in the recent past. In particular, we wish to detect subdigraphs with an unusually high connectivity, as measured by our statistic. This conditional approach alleviates the requirement to posit an appropriate and simple null graph model—but does require some (approximate) stationarity.

4. The Enron Data

The Enron email dataset is available online (http://www.cs.queensu.ca/home/skill/siamworkshop.html). This dataset consists of a collection of 150 folders corresponding to the email to and from senior management and others at Enron, collected over a period from about 1998 to 2002. The emails have been minimally processed to correct integrity problems. Some emails have been deleted, as have all attachments. Thus, while imperfect, this dataset represents a rich environment in which to perform text analysis and link analysis. More information on this dataset can be found online (http://www-2.cs.cmu.edu/∼enron).

One consequence of the processing of these data is that some of the original email addresses have been changed. Invalid addresses were converted to *no_address@enron.com*. In several cases, individuals have multiple addresses, which are clearly a result of some post-processing: for example, John Q. Public has email addresses *john.public@enron.com* and *q..public@enron.com*. In this study we will treat such cases as distinct; one potential goal might be to recognize this "aliasing" from the link analysis alone, without reference to the content of the messages. This will be discussed further in Section 7.1.

5. Whence our Enron Graphs?

The data are collected from "about 150 users"—mostly Enron executives, but also some energy traders, executive assistants, etc. However, our graphs are based on 184 users, which is the number of unique addresses we obtain from the 'From' line of emails in the 'Sent' boxes after manually removing some addresses which are clearly not associated with the 150 users. (NB: Neither of the two extreme options—keeping all addresses, or merging to the point of one-to-one correspondence between addresses and known users—seems practical; the former yields too many obvious aliases and extraneous addresses, and no simple unassailable version of the latter presents itself to us. Thus, we proceed with an admittedly imperfect collection of vertices.) In addition, some of the time stamps in the original data are clearly invalid, occurring before Enron existed, so we restrict our attention to a period of 189 weeks, from 1998 through 2002.

For each week $t = 1, \ldots, 189$, there is a digraph $D_t = (V, A_t)$ with $|V| = 184$ vertices and directed edges (arcs) A_t, where $(v, w) \in A_t \Leftrightarrow$ vertex v sends at least one e-mail to vertex w during the t-th week. We make no distinction between emails sent "To", "CC" or "BCC".

6. Statistics and Time Series

Our time-dependent scale-k locality statistic is given by

$$\Psi_{k,t}(v) = |A(\Omega(N_k[v; D_t]))| \tag{4}$$

for $k \in \{1, 2, \ldots, K\}$. In an abuse of notation, we will let $\Psi_{0,t}(v) = \text{outdegree}(v; D_t)$.

Figure 1 shows the three statistics

$$M_{k,t} = \max_v \Psi_{k,t}(v); \quad k = 0, 1, 2 \tag{5}$$

as well as $size(D_t)$, as functions of time (weeks) $t = 1, \ldots, 189$ for the 189 weeks under consideration. (Figures 8–11 show these four curves separately.)

The raw locality statistics $\Psi_{k,t}(v)$ are inadequate for our purposes. Consider, for instance, the situation in which one vertex, v, has a lot of activity throughout time, and another vertex, w, has but one tenth this amount of activity until one week in which w triples its activity. Without some form of vertex-dependent standardization, the increase in activity for w will

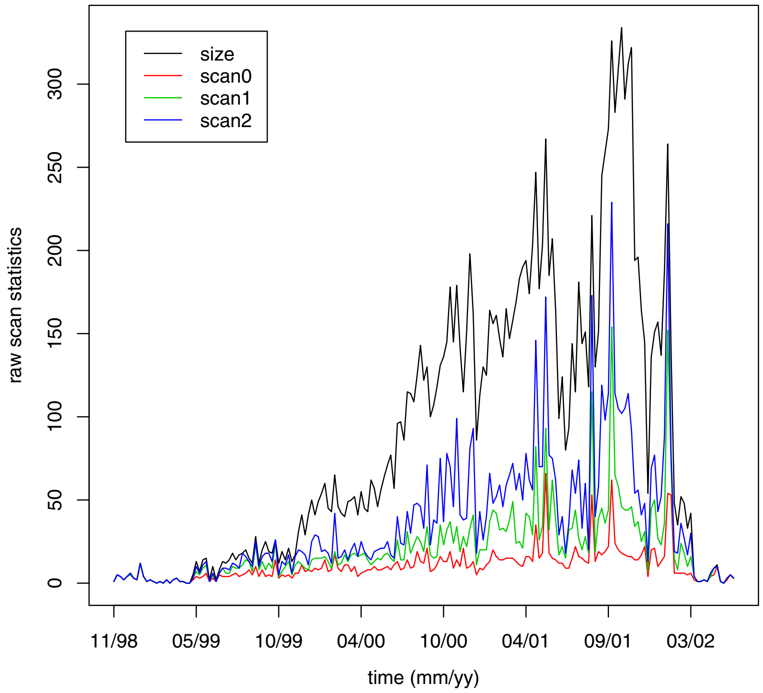

Figure 1. Time series of scan statistics and max degree ($M_{k,t}$ for $k = 0, 1, 2$), as well as digraph size, for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figures 8–11.)

go unnoticed, as $v = \arg\max \Psi_{k,t}(v)$ regardless of w's increased activity. Thus the locality statistics $\Psi_{k,t}(v)$ must be standardized using vertex-dependent recent history.

Our vertex-standardized locality statistic, for $k = 0, 1, 2$, is given by

$$\tilde{\Psi}_{k,t}(v) = (\Psi_{k,t}(v) - \hat{\mu}_{k,t,\tau}(v)) / \max(\hat{\sigma}_{k,t,\tau}(v), 1) \tag{6}$$

where

$$\hat{\mu}_{k,t,\tau}(v) = \frac{1}{\tau} \sum_{t'=t-\tau}^{t-1} \Psi_{k,t'}(v) \tag{7}$$

and

$$\hat{\sigma}^2_{k,t,\tau}(v) = \frac{1}{\tau - 1} \sum_{t'=t-\tau}^{t-1} (\Psi_{k,t'}(v) - \hat{\mu}_{k,t,\tau}(v))^2. \tag{8}$$

That is, we standardize the locality statistic $\Psi_{k,t}(v)$ by a vertex-dependent mean and standard deviation based on recent history. (The denominator in $\tilde{\Psi}_{k,t}(v)$ is forced to be greater than or equal to one to eliminate fragility due to vertices with little or no variation in activity.)

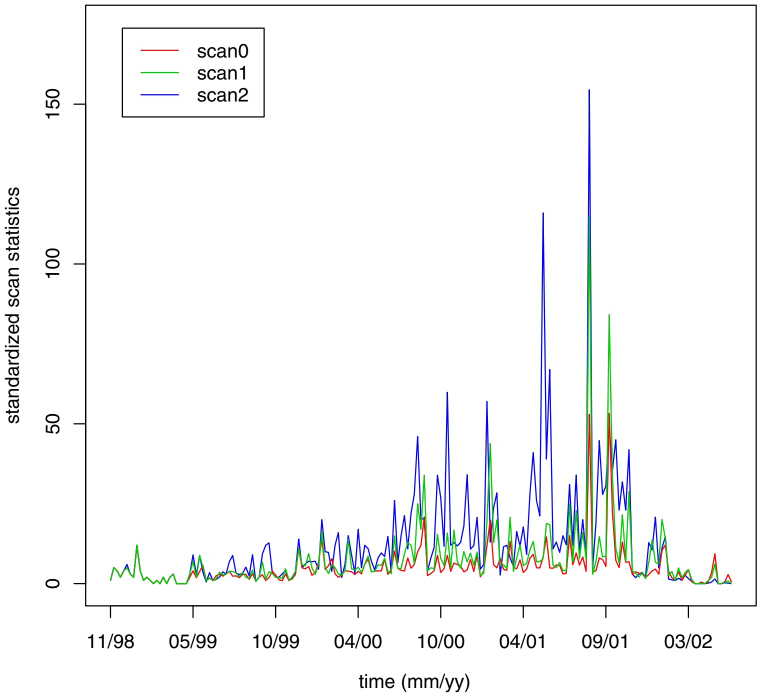

Figure 2. Time series of standardized scan statistics and max degree ($\tilde{M}_{k,t}$ for $k = 0, 1, 2$) for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figures 12–14.)

In figure 2 we plot the standardized scan statistics

$$\tilde{M}_{k,t} = \max_{v} \tilde{\Psi}_{k,t}(v) \tag{9}$$

against t over the 189 weeks. (Figures 12–14 show these three curves separately.)

This approach requires a vertex-dependent local stationarity assumption. The validity of a stationarity assumption is obviously suspect over the entire 189 weeks, but short-time near-stationarity (we use $\tau = 20$) may be reasonable as a null model.

7. Anomaly Detection

Given the standardized scan statistic time series $\tilde{M}_{k,t}$ presented in figure 2, we now consider anomaly detection.

For simplicity, we consider a temporally-normalized version of $\tilde{M}_{k,t}$,

$$S_{k,t} = (\tilde{M}_{k,t} - \tilde{\mu}_{k,t,\ell})/\max(\tilde{\sigma}_{k,t,\ell}, 1), \tag{10}$$

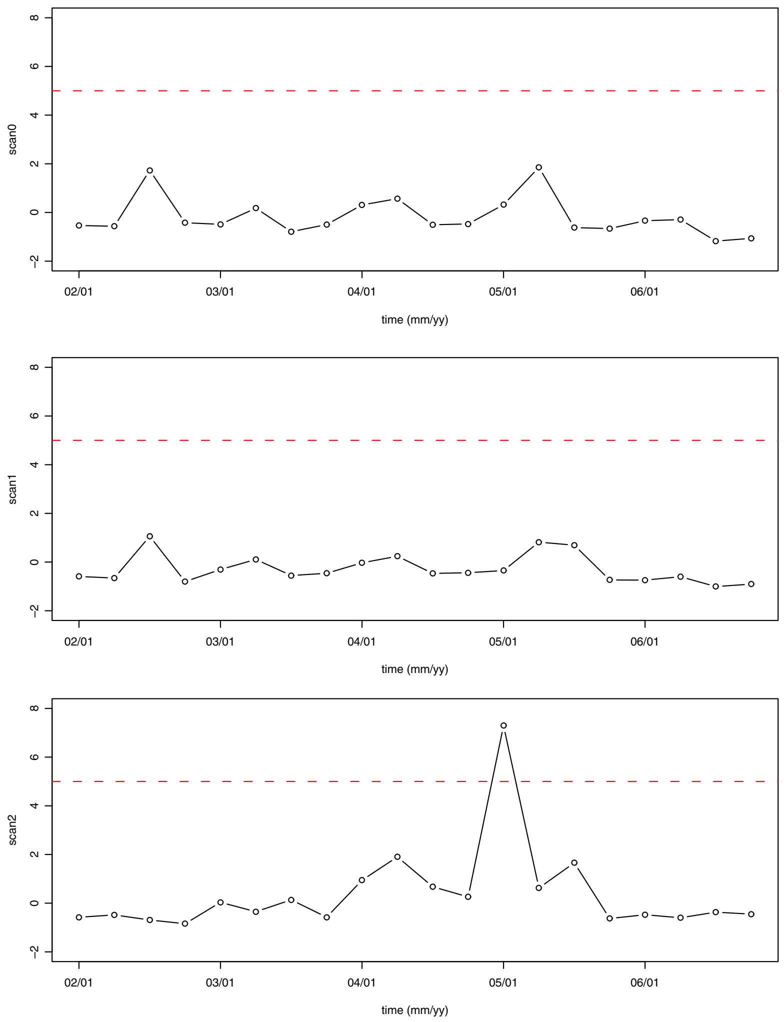

Figure 3. $S_{k,t}$, the temporally-normalized standardized scan statistics, on zoomed-in time series of Enron e-mail graphs during a period of 20 weeks in 2001. Top: $k = 0$; Middle: $k = 1$; Bottom: $k = 2$. This figure shows a detection (a standardized statistic $\tilde{M}_{k,t}$ which achieves a value greater than 5 standard deviations above its running mean, or a temporally-normalized standardized statistic $S_{k,t}$ in this plot taking a value greater than 5) at week $t^* = 132$ in May 2001 for scale $k = 2$, but not for $k = 1$ or $k = 0$.

where $\tilde{\mu}_{k,t,\ell}$ and $\tilde{\sigma}_{k,t,\ell}$ are the running mean and standard deviation estimates of $\tilde{M}_{k,t}$ based on the most recent ℓ time steps. (Here we use $\ell = 20$.) Detections are defined here as weeks for which $\tilde{M}_{k,t}$ achieves a value greater than five standard deviations above its mean; i.e., times t such that $S_{k,t} > 5$.

Figure 3 depicts $S_{2,t}$ for a 20 week period from February 2001 through June 2001. We observe that the second order scan statistic indicates a clear anomaly at $t^* = 132$

(max$_v$ $\tilde{\Psi}_{2,132}(v)$ is a seven sigma event) in May 2001. This anomaly is apparent, in hindsight, in figure 2.

Inference performed using simple sigmages is inadequate in this case, of course, because there is no reason to believe that the distribution of $S_{k,t}$ is normal or that $S_{k,t}$ and $S_{k,t'}$ are independent. Computational methods such as the bootstrap would be appropriate. We consider exceedance probabilities of an extreme value distribution, the Gumbel, fit via the method of moments. $S_{2,132} = 7.3$; 7.3 standard deviations yields a p-value $<10^{-10}$, assuming normality. While the significance for the detection at $t^* = 132$ is not so drastic under the more reasonable Gumbel model, we nevertheless obtain an exceedance probability $<10^{-6}$, which remains convincing. Bonferonni analysis suggests that if the $\tilde{\Psi}_{k,t}$ are approximately distributed as a t_{19} then the detection is significant; however, if the distribution of the $\tilde{\Psi}_{k,t}$ has extraordinarily heavy tails (e.g., Cauchy) then the $\alpha = 0.05$ level critical value may be greater than 7.3. Thus, under a reasonable range of null distributions, the detection at $t^* = 132$ is statistically significant.

Figure 4 shows the graph topology, sans isolates, for our 'detection' graph D_{132}. Our vertex of interest, $v^* = \arg\max_v \tilde{\Psi}_{2,132}(v)$, is identified with email address *email*90. Of note is the fact that $\arg\max_v \Psi_{0,132}(v) =$ *email*83. That is, the vertex of maximum outdegree for $t^* = 132$ is *not* the cause of our detection. Furthermore, $\arg\max_v \Psi_{1,132}(v) =$ *email*83, $\arg\max_v \Psi_{2,132}(v) =$ *email*147, $\arg\max_v \tilde{\Psi}_{0,132}(v) =$ *email*147, and $\arg\max_v \tilde{\Psi}_{1,132}(v) =$ *email*75. Thus the detection based on $v^* =$ *email*90 is apparent only when using the standardized second order scan statistic.

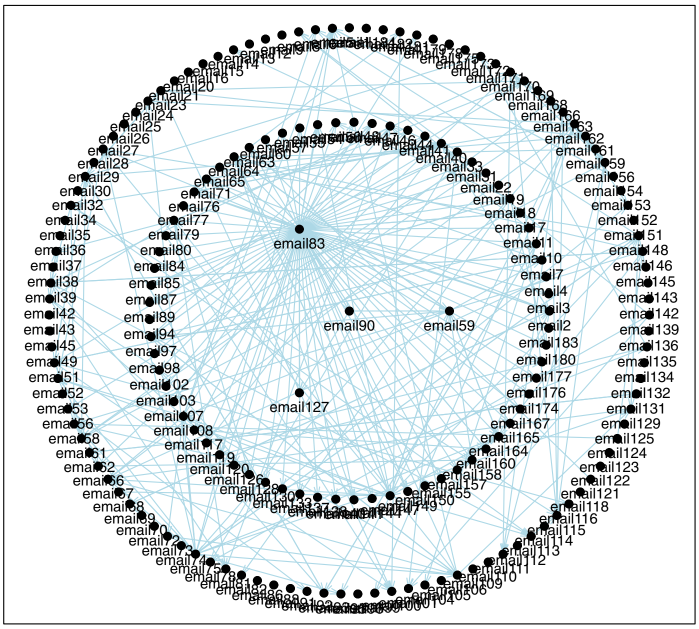

Figure 4. Plot of the 'detection' Enron email graph D_{132} (sans isolates) for which our scan statistic methodology detects an anomaly. The center vertex, *email*90, is $v^* = \arg\max_v \tilde{\Psi}_{2,132}$.

Table 1. Details for the 'detection' graph D_{132}.

Time t^*	132 (week of May 17, 2001)		
size(D_{132})	267		
Scale k	$M_{k,132}$	$\tilde{M}_{k,132}$	$S_{k,132}$
0	66	8.3	0.32
1	93	7.8	−0.35
2	172	116.0	7.30
3	219	174.0	5.20
Number of isolates		50	

Table 1 gives some relevant numerical values for the 'detection' graph D_{132}.

There is excessive activity among the elements of the closed 2-neighborhood of our vertex of interest v^* which is not accounted for by its outdegree (or its closed 1-neighborhood). In fact, v^* communicates, in particular, with other vertices each of which has high outdegree. This type of excessive local activity is precisely the raison d'etre for our scan statistics; our approach exhibits the ability to detect this anomaly.

Is this detection an event of interest? It is statistically significant, but the objective of our scan statistic methodology is to sift through massive communications data to find potentially informative events for the purpose of directing additional, more time consuming investigations. The ultimate determination of the practical significance of this or any detection must be made on the basis of subsequent analysis. There is a coinciding insider trading event on the Enron time line ... but there are many insider trading events on the Enron time line! Ideally, one would hope to find a link between the detected excess activity and that insider trading. Such a forensic analysis will require delving into the content of the email messages and associated meta-data.

Time $t^* = 132$ is the only week among the 189 under consideration for which $S_{2,t} \geq 5$. Detections for the other scan statistics—orders 0, 1, and 3—that may be worth pursuing are summarized here. For maximum standardized outdegree, there are three weeks with $S_{0,t} \geq 5$: 58 (week of December 16, 1999), 96 (week of September 7, 2000), 146 (week of August 23, 2001) for the standardized first order scan statistic, we obtain (almost) the same three detections: 58, 94, 146. The standardized third order scan statistic produces detections at $t^* = 132$ and at week 87.

7.1. Aliasing

In the case of the detection at $t^* = 132$,

$$v^* = \arg\max_{v} \tilde{\Psi}_{2,132}(v) = \mathit{email}90, \tag{11}$$

perusal of the emails shows that *email*90 and *email*141 are really the same person. User *email*90 had no activity before $t^* = 132$, at which time *email*141 switched to the *email*90 identifier. Thus we have detected an instance of aliasing, which could perhaps have been addressed during the manual merging stage wherein we settled on the collection of 184 vertices to consider. Of course, this identification does in fact require perusal of the emails, which perusal was suggested by the detection ... precisely the point of the exercise!

However, it may be possible to automatically identify such aliasing events. Given the detection (v^*, t^*) we can immediately identify *email*90 as having had no activity prior to $t^* = 132$. From this point, we may employ a "matched filter" scheme to determine candidates for aliasing by matching the pattern of *email*90's activity at or after $t^* = 132$ against the pattern of other vertices' activity prior to $t^* = 132$. Vertices with a high score for some matching function will be deemed likely candidates for further investigation.

For instance, we may compute, for each vertex $v \in V \setminus \{v^*\}$, the simple score

$$s_{t^*,\kappa}(v; v^*) = \sum_{t'=t^*-\kappa}^{t^*-1} |N_1(v; D_{t'}) \cap N_1(v^*; D_{t^*})|. \tag{12}$$

In this case we obtain $email141 = \arg\max_v s_{t^*,\kappa}(v; v^*)$ with $\kappa \geq 5$. That is, for this simple case, the aliasing can be automatically identified and resolved.

This idea of employing matched filters to time series of graphs, introduced here in a very simplistic fashion, will be pursued in more detail elsewhere.

7.2. Another Detection

The detection of $v^* = \arg\max_v \tilde{\Psi}_{2,132}(v) = email90$ at $t^* = 132$, while real and interesting, is due to the fact that *email*90 had not been active prior to $t^* = 132$. We may be interested, instead, in detections for which activity increases from a non-zero baseline. That is, we consider the statistic

$$\tilde{\Psi}_{k,t}(v) \cdot I\{\hat{\mu}_{0,t,\tau}(v) > c\}, \tag{13}$$

where $I\{E\}$ is the indicator function taking value one if event E occurs and taking value zero otherwise, which requires there to have been some recent activity.

For $c = 1$, one such detection of this type, for which the order $k = 2$ scan statistic detects but the order $k = 0$ and $k = 1$ scan statistics do not detect, is $v^* = email152$ at $t^* = 152$ (the week of October 4, 2001).

Table 2 gives the scan statistics for this detection for the weeks up to and including t^*. Here we see clearly the increase in activity, and we see that it is not due to order 0 or order 1 locality statistics. (N.B. It does appear that a detection at $t^* - 2$ may be appropriate.)

However, further investigation indicates that this detection is due to the fact that *email*152 communicates with *email*154, and *email*154 is an order 0 locality statistic detection at $t^* = 152$ due to a massive increase in outdegree (see Table 3).

Thus, in some sense, neither the *email*90/*email*141 detection at $t^* = 132$ nor the *email*152/*email*154 detection at $t^* = 152$ is really due to the type of excessive "chatter" in which we are most interested.

Table 2. Locality statistics $\Psi_{k,t}(v^* = email152)$ for the time range $\{t^* - 5, \ldots, t^*\}$ leading up to the $v^* = email152$ detection at $t^* = 152$.

Scale k	$\Psi_{k,t^*-5:t^*}(v^*)$
0	[1, 2, 1, 3, 1, 2]
1	[1, 2, 2, 9, 2, 4]
2	[1, 2, 2, 19, 4, 175]
3	[1, 2, 2, 58, 6, 268]

Table 3. Locality statistics $\Psi_{k,t}$ $(v = email154)$ for the time range $\{t^* - 5, \ldots, t^*\}$ leading up to the $v^* = email152$ detection at $t^* = 152$.

Scale k	$\Psi_{k,t^*-5:t^*}(v)$
0	[3, 2, 0, 2, 3, 62]
1	[3, 3, 0, 3, 6, 154]
2	[4, 3, 0, 37, 11, 229]
3	[4, 3, 0, 98, 16, 267]

7.3. Detecting Chatter

For each time t and vertex v, consider the order 2 statistic

$$\tilde{\Psi}'_t(v) = (\tilde{\Psi}_{2,t}(v) \cdot \mathcal{I}_{t,\tau}(v)) / \max(\gamma_t(v), 1). \tag{14}$$

Here the term $\mathcal{I}_{t,\tau}(v)$ is the product of three indicator functions,

$$I\{\hat{\mu}_{0,t,\tau} > c_1\}, \tag{15}$$

$$I\{\Psi_0(v) < \hat{\sigma}_{0,t,\tau}(v)c_2 + \hat{\mu}_{0,t,\tau}(v)\}, \tag{16}$$

$$I\{\Psi_1(v) < \hat{\sigma}_{1,t,\tau}(v)c_3 + \hat{\mu}_{1,t,\tau}(v)\}. \tag{17}$$

That is, we gate the second order scan statistic so that some minimal level of recent activity is required, and we insist that the order 0 and order 1 scan statistics do not yield detections. In this way we narrow the class of alternatives under consideration—the types of anomalous activities that will be deemed detections; we seek a detection in which the excess activity is due to chatter amongst the 2-neighbors. We include an "inhomogeneity penalty" $\gamma_t(v)$, the standard deviation of the outdegrees of the neighbors $N_1(v^*; D_{t^*})$, in the denominator of $\tilde{\Psi}'_t(v)$ to further narrow our search to the case of "balanced chatter" (and to rule out events such as the $email152/email154$ detection at $t^* = 152$).

The $\arg\max_{(v,t)} \tilde{\Psi}'_t(v)$ is given by $(v^*, t^*) = (email164, 109)$. (The value of $t^* = 109$ corresponds to the week of December 7, 2000.) Figure 5 displays $\tilde{M}'_t = \max_v \tilde{\Psi}'_t(v)$ as well as the temporally-normalized version S'_t.

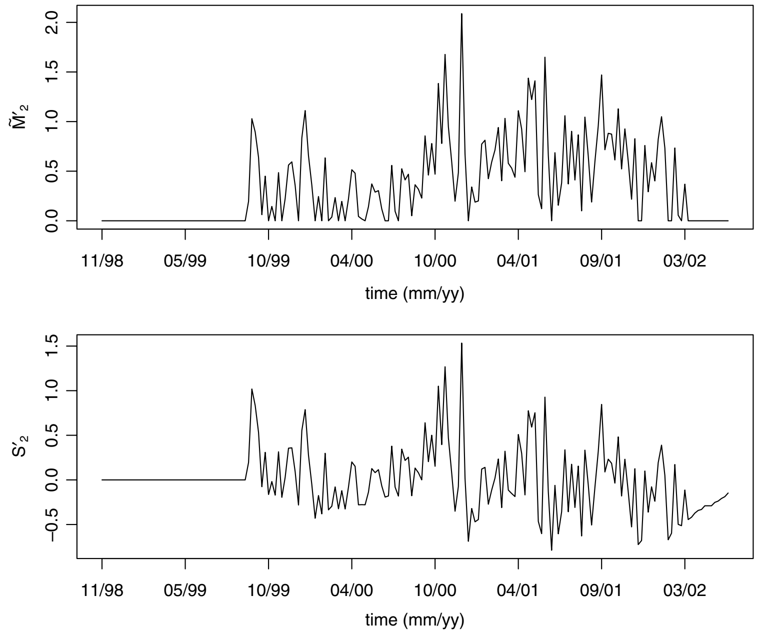

Figure 5. Plot of order 2 statistics $\tilde{M}'_t$ and S'_t showing the maximum at $t^* = 109$ in December 2000. This is the $email164$ "excessive chatter" detection.

The raw locality statistics $\Psi_{k,t}(v^*)$ for the time range $\{t^* - 5, \ldots, t^*\}$ leading up to this detection are given in Table 4. As can be seen from Table 4, the raw locality statistics for $k = 0$ and $k = 1$ do not have a substantial signal at $t^* = 109$, while for $k = 2$ the presence of an anomaly is clear.

The inhomogeneity penalty for this detection is $\gamma_{t^*}(v^*) \approx 1.7$; the outdegrees of the five neighbors of $v^* = email164$ are 6,6,6,7,10.

The induced subdigraph at $t^* = 109$, $\Omega(N_2[v^*; D_{t^*}])$, is depicted in figure 6. We see that $v^* = email164$ has five neighbors, each of which has outdegree between six and ten. That is, this detection is due to v^* communicating with a moderate subset of vertices, each of whom communicates with another moderate subset. Comparing this graph with $email164$'s

Table 4. Locality statistics $\Psi_{k,t}(v^*)$ for the time range $\{t^* - 5, \ldots, t^*\}$ leading up to the $email164$ detection at $t^* = 109$.

Scale k	$\Psi_{k,t^*-5:t^*}(v^*)$
0	[3, 5, 4, 5, 4, 5]
1	[11, 13, 10, 10, 11, 18]
2	[14, 35, 21, 38, 13, 65]

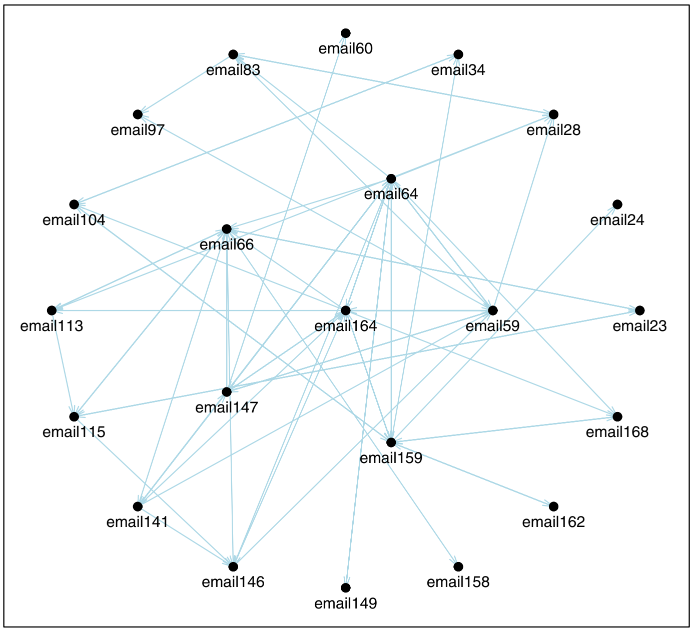

Figure 6. Plot of the 'detection' Enron email graph $\Omega(N_2[v^* = email164; D_{t^*=109}])$.

induced subdigraph $\Omega(N_2[v^*; D_{t^*-1}])$ at $t^* - 1 = 108$ (black arcs and associated vertices in figure 7) gives a clear, albeit simplistic, indication that change has occurred. Figure 7 gives additional information regarding this change, depicting the subdigraph induced at $t^* - 1 = 108$ by the union of *email*164's 2-neighborhood at $t^* - 1 = 108$ and *email*164's 2-neighborhood at $t^* = 109$. The arcs corresponding to communications between members of *email*164's closed 2-neighborhood at $t^* - 1 = 108$ are depicted in black; gray arcs represent other communications in D_{108} between vertices in *email*164's 2-neighborhood at $t^* = 109$. Figure 7 shows that this detection is not the result of a simple increase in the size of v^*'s neighborhood, but that the vertices in the neighborhood at t^*, while active at $t^* - 1$, have also increased their activity. Thus, the detection is not due solely to v^* joining a larger group; in addition, the group itself is more active as well. We interpret this figure as suggesting that this detection is robust—insensitive to small changes in the graph.

8. Discussion

A theory of scan statistics on graphs offers promise for detecting anomalies in time series of graphs.

We have employed perhaps overly-simplistic time series and inference methods, for purposes of illustration; more elaborate methods such as exponential smoothing, detrending, and variance stabilization may be appropriate. In addition, multivariate time series (one time

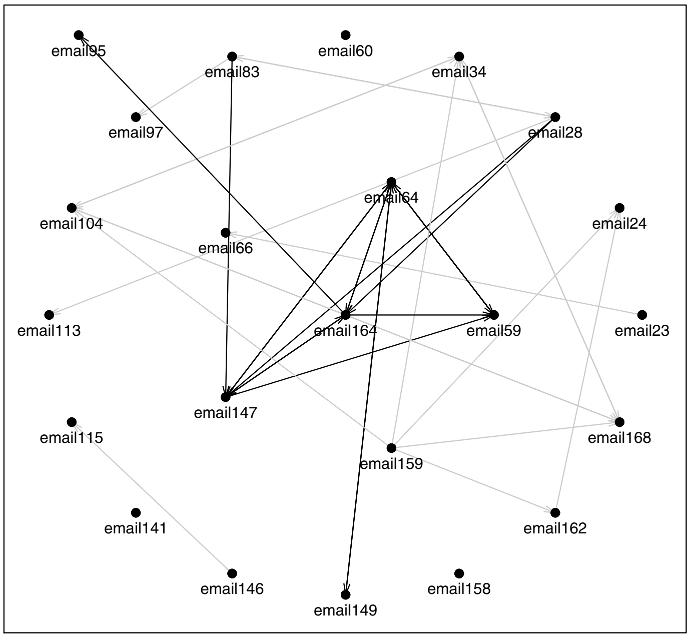

Figure 7. An induced subgraph of D_{108}. Black arcs and associated vertices represent *email*164's induced subdigraph $\Omega(N_2[v^*; D_{t^*-1}])$ at $t^* - 1 = 108$. Gray arcs represent other communications in D_{108} between vertices in *email*164's 2-neighborhood at $t^* = 109$. Comparing this figure with figure 6 provides information regarding the change from $t^* - 1 = 108$ to $t^* = 109$ for the ($v^* = email164$, $t^* = 109$) detection.

series for each vertex v, in this case) have a theory all their own—e.g., vector autoregressive models—which we have ignored here. And, of course, for data such as this Enron corpus, robust versions of moment estimates we have employed are called for.

Nevertheless, despite our simplistic approach to these various issues, we have demonstrated the potential utility of the scan statistic approach to the problem of anomaly detection in a time series of Enron email graphs. Much remains to be done—mathematically, computationally, and with respect to data and meta-data analysis. Of particular interest is the extension of these scan statistics to weighted graphs (and hypergraphs), allowing for the detection of anomalies related to the number (and possibly type) of messages sent, as opposed to the simpler case considered herein.

Noteworthy as a closing fact is that the procedures introduced herein can all be performed in a real-time, streaming data environment. That is, a sliding one-week window, rather than disjoint one-week windows, can be utilized and nothing presented herein causes a common laptop computer difficulty in keeping up. Thus, these procedures can be applied in scenarios of on-line analysis, in addition to the forensic scenario offered by this Enron corpus.

Appeindix

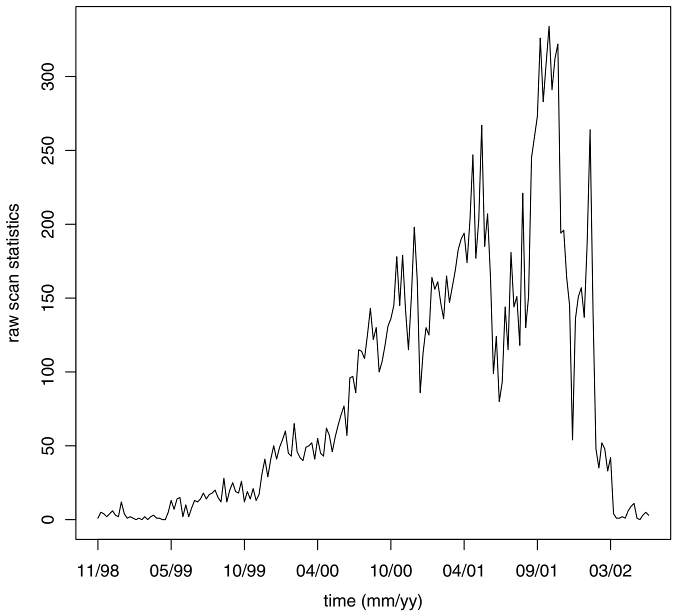

Figure 8. Time series of digraph size for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 1.)

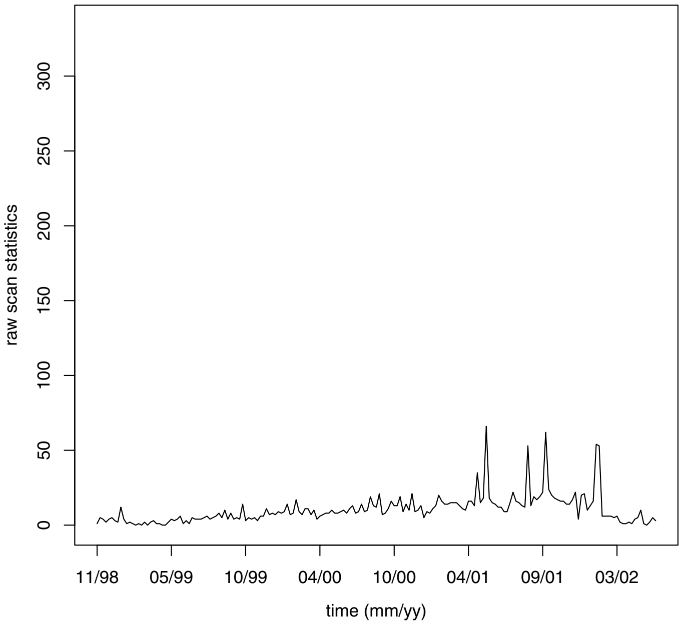

Figure 9. Time series of scan statistic $M_{0,t}$ (max degree) for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 1.)

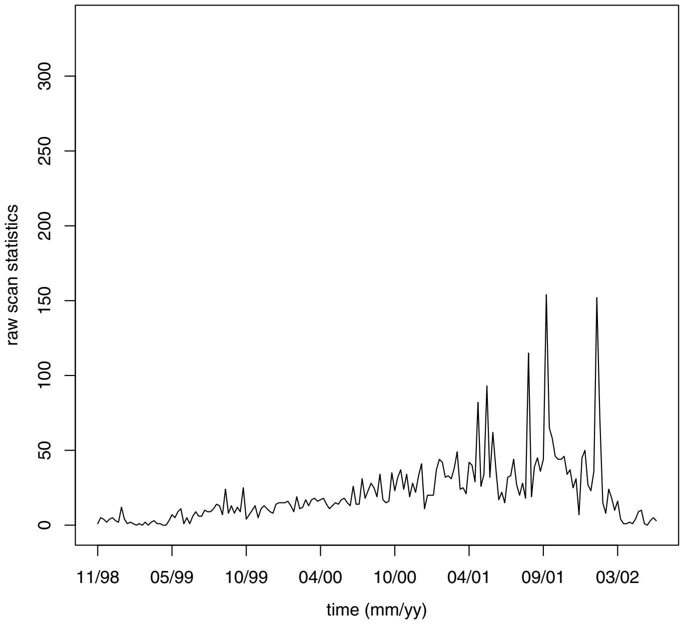

Figure 10. Time series of scan statistic $M_{1,t}$ for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 1.)

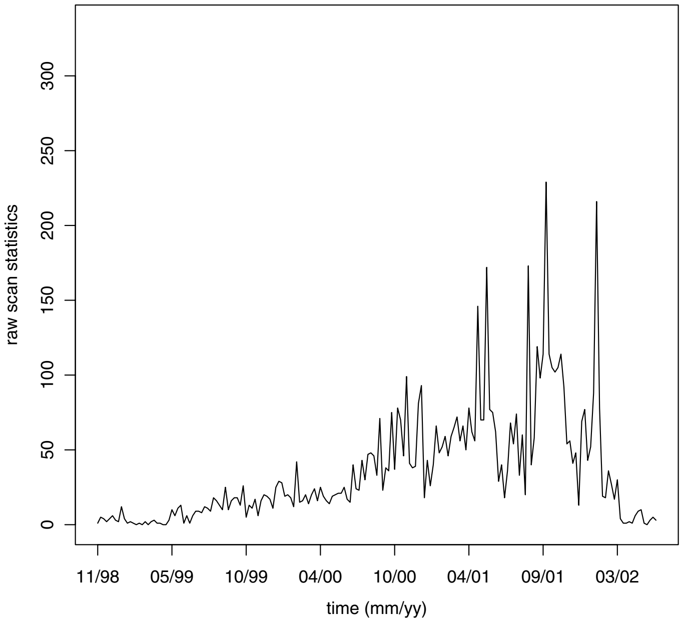

Figure 11. Time series of scan statistic $M_{2,t}$ for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 1.)

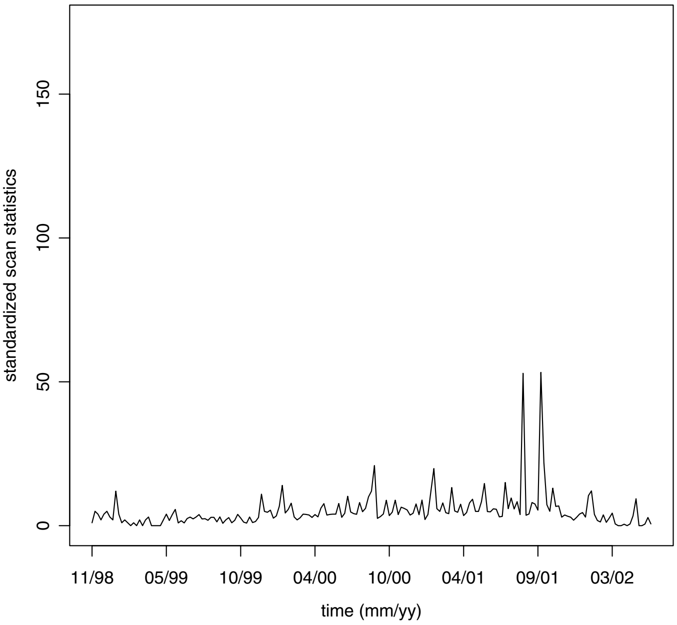

Figure 12. Time series of standardized scan statistic $\tilde{M}_{0,t}$ for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 2.)

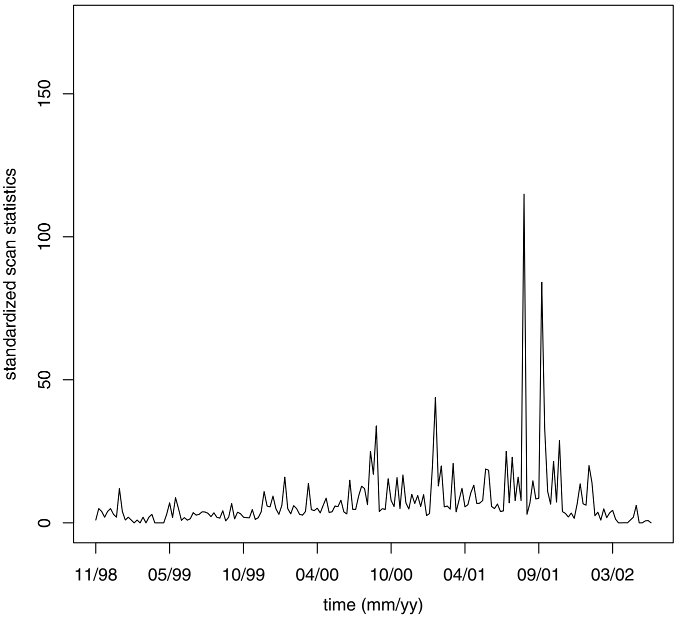

Figure 13. Time series of standardized scan statistic $\tilde{M}_{1,t}$ for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 2.)

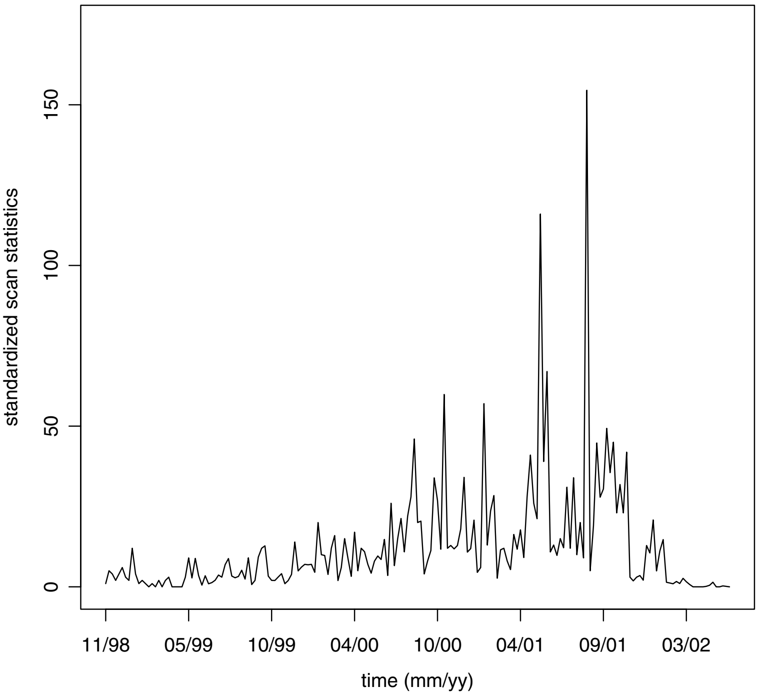

Figure 14. Time series of standardized scan statistic $\tilde{M}_{2,t}$ for weekly Enron email digraphs during a period of 189 weeks from 1998–2002. (See also figure 2.)

References

Adler, R.J. (1984), "The Supremum of a Particular Gaussian Field," *Annals of Probability*, 12, 436–444.

Chen, J. and J. Glaz (1996), "Two-Dimensional Discrete Scan Statistics," *Statistics and Probability Letters*, 31, 59–68.

Cressie, N. (1977), "On Some Properties of the Scan Statistic on the Circle and the Line," *Journal of Applied Probability*, 14, 272–283.

Cressie, N. (1980), "The Asymptotic Distribution of the Scan Statistic Under Uniformity," *Annals of Probability*, 8, 828–840.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley, New York.

Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns*, Academic Press, New York.

Fisher, R.A., T.H., and W. Mackenzie (1922), "The Accuracy of the Plating Method of Estimating the Density of Bacterial Populations, with Particular Reference to the Use of Thornton's Agar Medium with Soil Samples," *Annals of Applied Biology*, 9, 325–359.

Loader, C. (1991), "Large-Deviation Approximations to the Distribution of Scan Statistics," *Advances in Applied Probability*, 23, 751–771.

Naiman, D. and C. Priebe (2001), "Computing Scan Statistic *p*-values Using Importance Sampling, with Applications to Genetics and Medical Image Analysis," *Journal of Computational and Graphical Statistics*, 10, 296–328.

Naus, J. (1965), "Clustering of Random Points in Two Dimensions," *Biometrika*, 52, 263–267.

Priebe, C. (2004), "Scan Statistics on Graphs," Technical Report 650, Johns Hopkins University, Baltimore, MD 21218–2682.

Carey E. Priebe received the B.S. degree in mathematics from Purdue University in 1984, the M.S. degree in computer science from San Diego State University in 1988, and the Ph.D. degree in information technology (computational statistics) from George Mason University in 1993. From 1985 to 1994 he worked as a mathematician and scientist in the US Navy research and development laboratory system. Since 1994 he has been a professor in the Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland. At Johns Hopkins, he holds joint appointments in the Department of Computer Science and the Center for Imaging Science. He is a past President of the Interface Foundation of North America—Computing Science & Statistics, a past Chair of the Section on Statistical Computing of the American Statistical Association, and on the editorial boards of Journal of Computational and Graphical Statistics, Computational Statistics and Data Analysis, and Computational Statistics. His research interests are in computational statistics, kernel and mixture estimates, statistical pattern recognition, statistical image analysis, and statistical inference for high-dimensional and graph data. He was elected Fellow of the American Statistical Association in 2002.

John M. Conroy received a B.S. in Mathematics from Saint Joseph's University in 1980 and a Ph.D. in Applied Mathematics from the University of Maryland in 1986. Since then he has been a research staff member for the IDA Center for Computing Sciences in Bowie, MD. His research interest is applications of numerical linear algebra. He is a member of the Society for Industrial and Applied Mathematics, Institute of Electrical and Electronics Engineers (IEEE), and the Association for Computational Linguistics.

David J. Marchette received a B.A. in 1980, and an M.A. in mathematics in 1982, from the University of California at San Diego. He received a Ph.D. in Computational Sciences and Informatics in 1996 from George Mason University under the direction of Ed Wegman. From 1985–1994 he worked at the Naval Ocean Systems Center in San Diego doing research on pattern recognition and computational statistics. In 1994 he moved to the Naval Surface Warfare Center in Dahlgren Virginia where he does research in computational statistics and pattern recognition, primarily applied to image processing, text processing, automatic target recognition and computer security. Dr. Marchette is a Fellow of the American Statistical Society.

Youngser Park received the B.E. degree in electrical engineering from Inha University in Korea in 1985, the M.S. degree in computer science from The George Washington University in 1991, and had pursued a doctoral degree there. From 1998 to 2000 he worked at the Johns Hopkins Medical Institutes as a senior research engineer. Since 2003 he is working as a research analyst in the Center for Imaging Science at the Johns Hopkins University. His research interests are clustering algorithm, pattern classification, and data mining.