Scan Statistics on Enron Hypergraphs

Youngser Park [*†] Carey E. Priebe [*] David J. Marchette [‡] Abdou Youssef [†]

Abstract

We present a theory of scan statistics on hypergraphs and apply the methodology to a time series of email data. This approach is of interest because a hypergraph is better suited to email data than a graph. This is due to the fact that a hypergraph can contain all the recipients of a message in a single hyperedge rather than treating each recipient separately in a graph. We will discuss our methodology in detail and provide an example of anomaly detection using this technique on a time series of Enron email data.

1 Introduction

In Priebe et al. [14], we introduced a theory of scan statistics on graphs and applied the idea to the problem of anomaly detection in a time series graphs. One example to keep in mind is a communications graph, in which the vertices of the graph represent people or computers and the edges correspond to communications between the entities. In [14], we tested the null hypothesis of "homogeneity" against alternatives suggesting "local subregions of excessive activity."

However, there is a limitation in the graph model, where each edge can connect only two vertices. This is clearly a problem; for example, an email message may have more than one recipient with multiple addresses in the "To:" field or may have non-empty "Cc:" or "Bcc:" fields. One way to represent this kind of message in a graph is to use individual edge for each (from, recipient) pair, but then an extra bit of information is necessary to identify edges associated with the same message. A better solution for this is a *hypergraph*, a generalized graph where edges can connect any number of vertices.

In this paper, we extend the methodology of [14] to hypergraphs and compare this extension with the original theory. Section 2 of this paper presents a background of scan statistics and its application to graphs and hypergraphs, section 3 introduces time series graphs and hypergraphs, section 4 explains how we generate hypergraphs from the Enron email dataset and a couple of experiments are shown. We conclude the paper with discussion in section 5.

[*]Johns Hopkins University, Baltimore, MD
[†]George Washington University, Washington, DC
[‡]Naval Surface Warfare Center, Dahlgren, VA

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 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 "quadrant counts" [8]; see [7]. 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 [10], [4], [5], and [9]. 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 [10], [1], [9], and [3]. As noted by [6], 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. [11] present an alternative approach, using importance sampling, to this problem of $p-$value approximation.

2.1 Scan Statistics on Graphs

Consider a directed graph (digraph) D with vertex set $V(D)$ and arc set $A(D)$ of directed edges. 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

$$\Omega(N_k[v; D]),$$

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,

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

This idea is introduced in [15].

Under a null model for the random digraph D (for instance, the Erdös-Rényi 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.

2.2 Scan Statistics on Hypergraphs Hypergraphs are a generalization of graphs, in which generalized edges (called *hyperedges*) may connect more than two vertices. A hypergraph $H = (V, \mathcal{E})$ consists of a set of vertices (or nodes) $V = \{v_1, v_2, \cdots, v_n\}$ and a set of hyperedges $\mathcal{E} = \{e_1, e_2, \cdots, e_m\}$, with $e_i \neq \emptyset$ and $e_i \subseteq V$ for $i = 1, \cdots, m$ [2]. An example with $n = 4$ and $m = 6$ is depicted in Figure 1.

We may represent any hypergraph with a $|V| \times |\mathcal{E}|$ incidence matrix $A = [a_{ij}]$ such that $a_{ij} \in \{0, 1\}$, where each row i is associated with a vertex v_i and each column j with a hyperedge e_j. The incidence matrix for the hypergraph depicted in Figure 1 is given by

$$A = \begin{array}{c|cccccc} \vdots & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ \hline v_1 \vdots & 1 & 1 & 0 & 0 & 0 & 1 \\ v_2 \vdots & 0 & 0 & 1 & 1 & 0 & 0 \\ v_3 \vdots & 1 & 0 & 1 & 0 & 1 & 1 \\ v_4 \vdots & 0 & 1 & 0 & 1 & 1 & 1 \end{array}$$

We are using following hypergraph definitions throughout the paper:

1. The order of a hypergraph H, denoted $order(H) = |V| = n$, is the number of vertices.
2. The size of a hypergraph H, denoted $size(H) = |\mathcal{E}| = m$, is the number of hyperedges,
3. The closed 1st-order neighborhood of v in H, denoted $N_1(v, H) = \bigcup_{v \in e_i, e_i \in \mathcal{E}} e_i$,
4. The closed kth-order neighborhood of v in H, denoted $N_k(v, H) = \bigcup_{v \in N_{k-1}(v,H)} N_1(v, H)$ for $k \geq 2$,
5. The induced subgraph $H(N_k, \mathcal{E}_k)$, denoted $\Omega(N_k(v, H))$, where $\mathcal{E}_k = \{e_i \in \mathcal{E} : e_i \subset N_k\}$,

The *locality statistic* at location v and scale k for a hypergraph H is denoted by $\Psi_k(v, H) = size(\Omega(N_k(v, H)))$, for $k \geq 1$. When $k = 0$, we will let it be a degree of vertex v, and $\Psi_0(v, H) = |\{e_i \in \mathcal{E} : v \in e_i\}|$. That is,

1. The degree of vertex v for a hypergraph H, denoted $\Psi_0(v, H) = |\{e_i \in \mathcal{E} : v \in e_i\}|$, that is, it is the number of edges of the induced subgraph formed by the edges containing v.
2. The locality statistic at vertex v and scale k for a hypergraph H, denoted $\Psi_k(v, H) = size(\Omega(N_k(v, H)))$ for $k \geq 1$,

The "scale-specific" *scan statistic* $M_k(H)$ is given by some function of the collection of locality statistics $\{\Psi_k(v)\}_{v \in V(H)}$; consider for instance the maximum locality statistic over all vertices,

$$M_k(H) = \max_{v \in V(H)} \Psi_k(v, H).$$

Let's consider a following graph, which we will introduce as an *authorship* graph for illustration; a

graph $G_1 = (V, E)$ consists of vertices (or authors) $V = \{v_1, v_2, v_3, v_4\}$ and edges (or papers) $E = \{e_1, e_2, e_3, e_4, e_5\}$, where $e_i = (v_j, v_k)$ means the paper e_i is written by authors v_j and v_k. The graph G_1 is shown in the top panel of Figure 1. When we calculate scan statistic on G_1, $\Psi_0(G_1) = \{2, 2, 3, 3\}$, and $\Psi_1(G_1) = \{3, 3, 5, 5\}$. Note that $\Psi_k(v_1, G_1) = \Psi_k(v_2, G_1)$, and $\Psi_k(v_3, G_1) = \Psi_k(v_4, G_1)$ so it is not clear whether the author v_1 is the same person as v_2 or not. (Note that we know $v_3 \neq v_4$ because there is an edge e_5.) Meanwhile, let us add another paper written by coauthors v_1, v_3, and v_4 and call the resulting graph G_2. If we use an unweighted graph as used in [14], G_2 becomes identical to G_1, and therefore it is still not possible to distinguish v_1 and v_2. Considering G_2 as a hypergraph H as shown in the bottom panel of Figure 1, however, the statistics are $\Psi_0(H) = \{4, 2, 5, 5\}$[1], and $\Psi_1(H) = \{5, 3, 7, 7\}$, and it is now clear that $v_1 \neq v_2$.

3 Statistics and Time Series

Our time-dependent scale-k locality statistic on time series hypergraph H_t is given by

$$\Psi_{k,t}(v, H_t) = size(\Omega(N_k(v, H_t)))$$

for $k \in \{1, 2, \cdots, k\}$. (We will let $\Psi_{0,t}(v) = degree(v, H_t)$.) And, their corresponding scan statistics are

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

.

As mentioned in [14], these raw locality statistics $\Psi_{k,t}$ are standardized using vertex-dependent recent history:

$$\widetilde{\Psi}_{k,t} = (\Psi_{k,t}(v) - \widehat{\mu}_{k,t,\tau}(v))/\max(\widehat{\sigma}_{k,t,\tau}(v), 1)$$

where

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

and

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

That is, we standardize the local statistic $\Psi_{k,t}(v)$ by a vertex-dependent mean and standard deviation based on recent history. The corresponding standardized scan statistics are

$$\widetilde{M}_{1,t} = \max_v \widetilde{\Psi}_{1,t}(v)$$

[1]For example, the edges for the subgraph containing v_1 are (v_1, v_3), (v_1, v_4), and another edges of (v_1, v_3) and (v_1, v_4) from the hyperedge e_6, so $\Psi_0(v_1, H) = 4$, and so on.

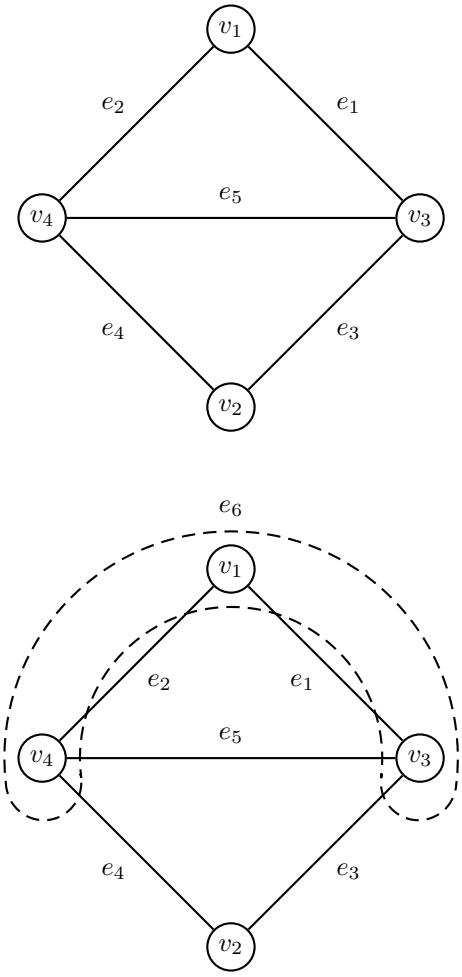

Figure 1: *Top*: A simple graph G_1. $\Psi_0(v, G_1) = \{2, 2, 3, 3\}, \Psi_1(v, G_1) = \{3, 3, 5, 5\}$. Note that $\Psi_k(v_1, G_1) = \Psi_k(v_2, G_1)$. *Bottom*: A corresponding hypergraph H with an extra hyperedge $e_6 = \{v_1, v_3, v_4\}$. $\Psi_0(v, H) = \{4, 2, 5, 5\}, \Psi_1(v, H) = \{5, 3, 7, 7\}$. Note that $\Psi_k(v_1, H) \neq \Psi_k(v_2, H)$.

.

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

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

where $\widetilde{\mu}_{k,t,\ell}$ and $\widetilde{\sigma}_{k,t,\ell}$ are the running mean and standard deviation estimates of $\widetilde{M}_t$ based on the most recent ℓ time steps.

4 Experiments

The Enron email dataset [12, 13] is processed exactly the same way as previously (also available at `http://www.cis.jhu.edu/~parky/Enron`) except that edges are now hyperedges considering all recipients, that is, in [14] emails sent "To", "Cc", and "Bcc" were undistinguished, they are now all in one hyperedge.

For each week $t = 1, \cdots, 189$, there is a hypergraph $H_t = (V, \mathcal{E}_t)$ with $|V| = 184$ vertices and hyperedges $\mathcal{E}_t$, where $\{v_1, \cdots, v_k\} \in \mathcal{E}_t \Leftrightarrow$ vertices $v_1, \cdots, v_k$ consists of the list of recipients and sender of at least one email during the t-th week.

4.1 Experiment 1 Figure 2 shows the three scan statistics ($\Psi_{k,t}(v, D_t)$ and $\Psi_{k,t}(v, H_t)$ for $k = \{0, 1, 2\}$) as well as the size for each of time series Enron graphs (dashed lines) and hypergraphs (solid lines), as functions of time (weeks) $t = 1, \cdots, 189$ for the 189 weeks under consideration.

First, we use raw locality scan statistics $\Psi_{k,t}(v, H_t)$ on time series Enron hypergraph H_t. Let $\|V\| = n$. For time t, let $M_{(j)}$ be order statistics, where $j = 1, \cdots, n$. Then, $M_{(n)} = \max_v \Psi(v, H_t)$ and $v^* = \{v \ s.t. \ \Psi(v, H_t) = M_{(n)}\} = \arg\max_v \Psi(v, H_t)$. Note that v is not necessarily unique. Let $k = k_t$ be such that $M_{(k)} = \max_{v \in v^*} \Psi(v, H_t)$. The value of k_t will be the maximum possible value $k_{\max}$ if $\arg\max_v \Psi(v, D_t) = \arg\max_v \Psi(v, H_t)$.

The Figure 3 depicts k_t versus t. Detections are defined here as weeks for which k_t is less than $k_{\max}$, that is, the index of $\arg\max_v \Psi(v, D_t)$ is much smaller than $\arg\max_v \Psi(v, H_t)$, k_{max} in this case. It shows that there are three potential noticible detections at $t^* = \{116, 121, 169\}$, which are in January 2001, February 2001, and January 2002 respectively.

4.2 Experiment 2 In Figure 4 we plot the standardized scan statistics $\widetilde{M}_{1,t}$ against t over the 189 weeks. We use $\tau = 20$ in this experiment.

Figure 5 depicts a temporally-normalized version of $\widetilde{M}_{1,t}$, $S_{1,t}$, based on the most recent ℓ time steps ($\ell = 20$ in this example). Detections are defined here as time for which $\widetilde{M}_{1,t}$ achieves a value greater than five

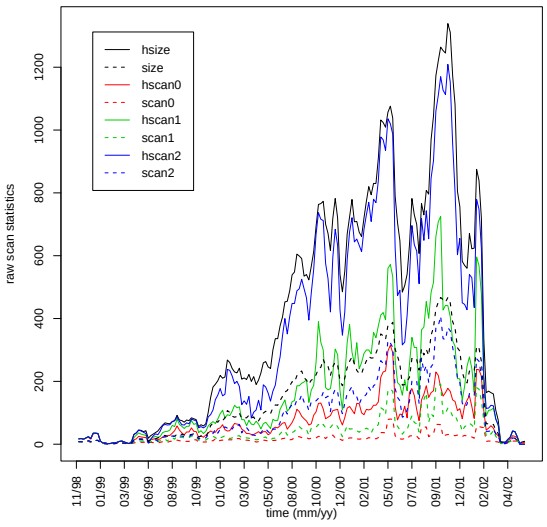

Figure 2: Time series of scan statistics ($\Psi_{k,t}(v, D_t)$ and $\Psi_{k,t}(v, H_t)$ for $k = \{0, 1, 2\}$) as well as graph size for weekly Enron email graphs and hypergraphs (please see the online colored version for better view).

standard deviations above its mean, *i.e.,* times t such that $S_{1,t} > 5$.

To compare with the original detection found in [14], we look into the same period of time (a 20 week period from February 2001 through June 2001), and the zoomed-in plots are depicted in Figure 6. The figure shows a detection (a standardized statistic $\widetilde{M}_{1,t}$ which achieves a value greater than 5 standard deviations above its running mean, or a temporally-normalizsed standardized statistic $S_{1,t}$ in this plot taking a value greater than 5) at week $t^* = 130$ in May 2001 for a hypergraph, but not for a graph. Note that the detection from $S_{2,t}(D_t)$ was $t^* = 132$, while our new detection using $S_{1,t}(H_t)$ is now $t^* = 130$, both are in May 2001. Recall that the former uses directed graphs while the latter uses undirected hypergraphs. That is, an email to r people in the directed graph causes a different effect from an email among r people in the undirected hypergraph; there is no degree change for r recipients in the former (because those messages are all incoming edges from the recipients point of view) while all $r+1$ people will get their degrees increased by r from a single hyperedge in the latter.

Figure 7 shows the locality statistics $\Psi_1(v, H_t)$ on hypergraph as a function of $\Psi_1(v, D_t)$ on a graph at week 130. Each data point represents an employee id. The points further away from the dotted line are the

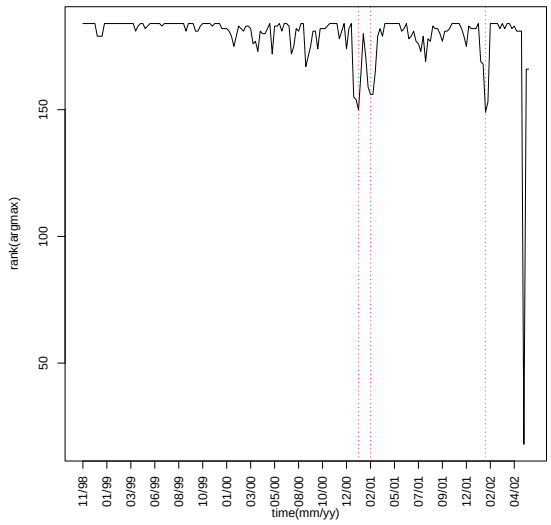

Figure 3: This plot depicts k_t versus t for $\Psi_{1,t}(v, D_t)$ and $\Psi_{1,t}(v, H_t)$. We consider t such that $k_t < k_{max}$ as detections. There are three potential noticible detections at $t^* = \{116, 121, 169\}$, which are in January 2001, February 2001, and January 2002, shown in red dotted lines.

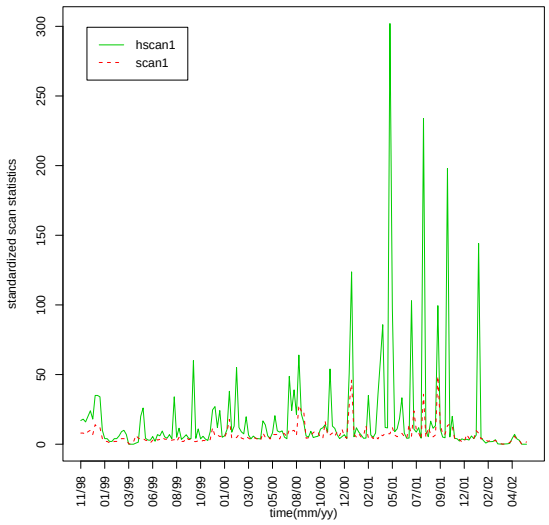

Figure 4: Time series of standardized scan statistics $\widetilde{M}_{1,t}(D_t)$ and $\widetilde{M}_{1,t}(H_t)$ for weekly Enron email graphs and hypergraphs.

ones we are interested in, *e.g.,* employees 79 and 97.

As we can see from Figures 8 and 9, locality statistics $\Psi_1(\{79, 97\}, H_t)$ show anomalous behavior at $t^* = 130$.

Figure 10 shows the similar plot as Figure 7 except that it's plotting $\widetilde{\Psi}_1(v, H_t)$ versus $\widetilde{\Psi}_1(v, D_t)$. It shows that employees 17 and 76 are of interest, and their behaviors are shown in Figures 11 and 12.

5 Discussion

We demonstrated the extended work of scan statistics on hypergraph in this paper.

Much remains to be done; 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 of messages sent, as opposed to the simpler case considered in this paper. A directed version of hypergraph is also essential.

Another important extension will be a *content analysis* along with a context analysis. After extracting features from corpus, clustering and document summarization methods can be applied to expose the *topics* within the region of time-series communication network, then these topic information can be used for detection of anomalies as time changes.

References

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

[2] C. Berge. *Hypergraphs: Combinatorics of Finite Sets*, North-Holland, 1989.

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

[4] N.A.C. Cressie, *On Some Properties of the Scan Statistic on the Circle and the Line*, Journal of Applied Probability, 14 (1977), pp. 272–283.

[5] N.A.C. Cressie, *The Asymptotic Distribution of the Scan Statistic under Uniformity*, Annals of Probability, 8 (1980), pp. 828–840.

[6] N.A.C. Cressie, *Statistics for Spatial Data*, John Wiley, New York, 1993.

[7] P.J. Diggle, *Statistical Analysis of Spatial Point Patterns*, Academic Press, New York, 1983.

[8] R.A. Fisher, H.G. Thornton, and W.A. Mackenzie, *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 (1922), pp. 325–359.

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

[10] J. Naus. *Clustering of random points in two dimensions*, In Biometrika, volume 52, pp. 263-267, 1965.

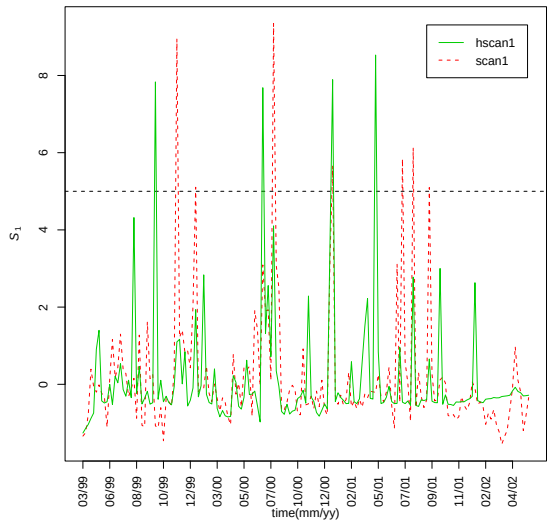

Figure 5: Time series of temporally-normalized standardized scan statistics ($S_{1,t}(D_t)$ and $S_{1,t}(H_t)$) for weekly Enron email graphs. The dotted horizontal line shows five standard deviations above a running mean.

[11] D.Q. Naiman and C.E. Priebe, *Computing Scan Statistic p–Values using Importance Sampling, with Applications to Genetics and Medical Image Analysis*, Journal of Computational and Graphical Statistics, 10 (2001), pp. 296–328.

[12] `www.cs.queensu.ca/home/skill/siamworkshop.html`

[13] `www-2.cs.cmu.edu/~enron`

[14] C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park. *Scan statistics on enron graphs*, In Computational and Mathematical Organization Theory, volume 11, pp. 229-247. Springer Science+Business Media B.V., October 2005.

[15] C.E. Priebe, *Scan Statistics on Graphs*, Technical Report No. 650, Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682, (2004).

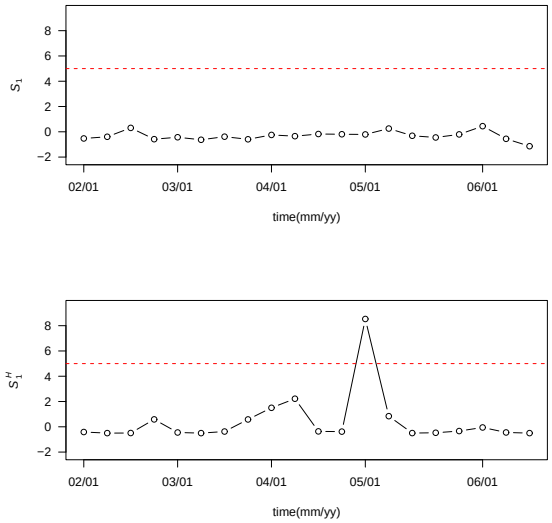

Figure 6: The temporally-normalized standardized scan statistics ($S_{1,t}(D_t)$ on top and $S_{1,t}(H_t)$ in below) on zoomed-in time series of Enron email graphs during a period of 20 weeks in 2001. The figure shows a detection (a standardized statistic $\widetilde{M}_{1,t}$ which achieves a value greater than 5 standard deviations above its running mean, or a temporally-normalizsed standardized statistic $S_{1,t}$ in this plot taking a value greater than 5) at week $t^* = 130$ in May 2001 for the hypergraph but not for the graph.

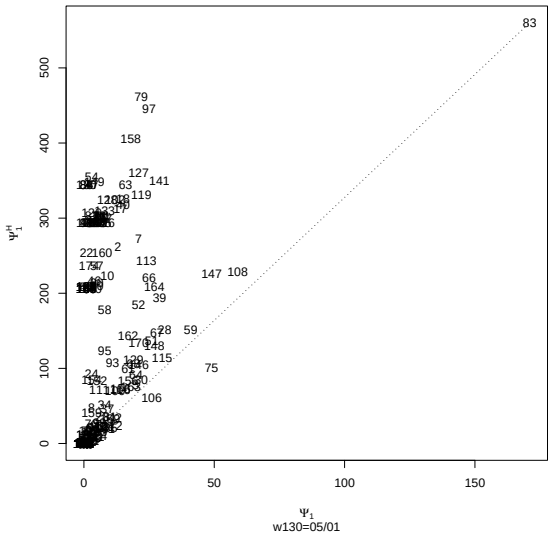

Figure 7: Locality statistics $\Psi_1(v, H_t)$ as a function of $\Psi_1(v, D_t)$ at $t^* = 130$. Each data point represents an employee id. The points further away from the dotted line are the ones we are interested in, *e.g.*, employees 79 and 97.

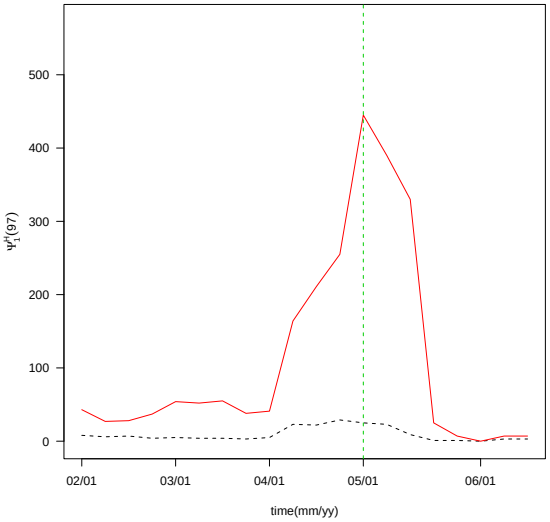

Figure 9: Locality statistics $\Psi_1(97, H_t)$ during a period of 20 weeks in 2001, which is shown in red solid line (the black dotted line is $\Psi_1(97, D_t)$). It shows a sudden value increase at $t^* = 130$ in May 2001.

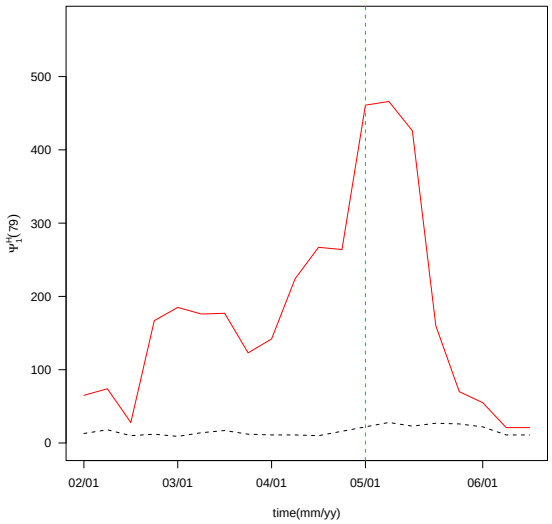

Figure 8: Locality statistics $\Psi_1(79, H_t)$ during a period of 20 weeks in 2001, which is shown in red solid line (the black dotted line is $\Psi_1(79, D_t)$). It shows a sudden value increase at $t^* = 130$ in May 2001.

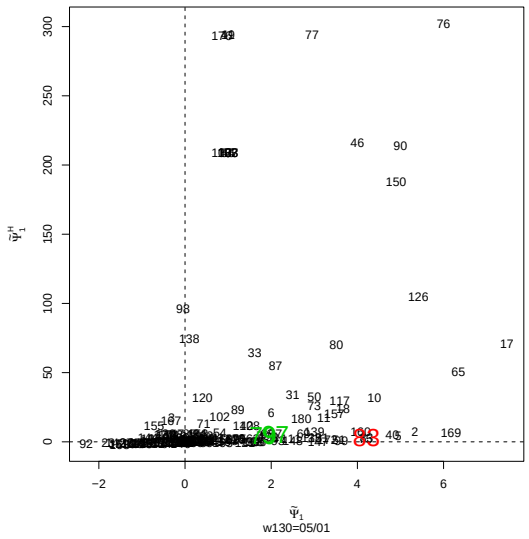

Figure 10: Standardized scan statistics $\widetilde{\Psi}_1(v, H_t)$ as a function of $\widetilde{\Psi}_1(v, D_t)$ at week $t^* = 130$. Each data point represents an employee id. The $\arg\max \Psi_1(v, D_{130}) = \arg\max \Psi_1(v, H_{130}) = 83$ is shown as red and employees 79 and 97 from Figure 7 are shown in green (please see the online colored version).

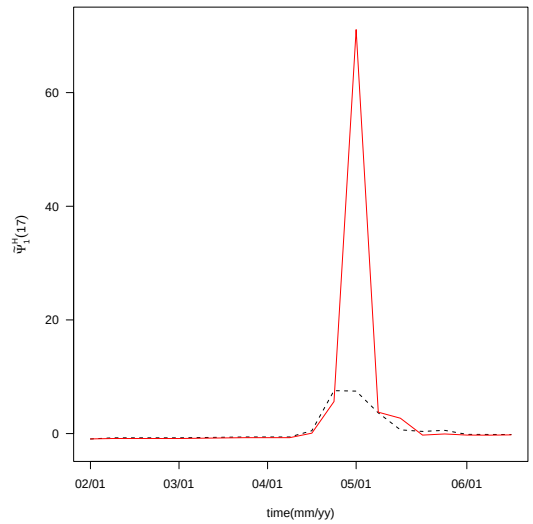

Figure 11: Locality statistics $\widetilde{\Psi}_1(17, H_t)$ during a period of 20 weeks in 2001, which is shown in red solid line (the black dotted line is $\widetilde{\Psi}_1(17, D_t)$). It shows a sudden value increase at $t^* = 130$ in May 2001.

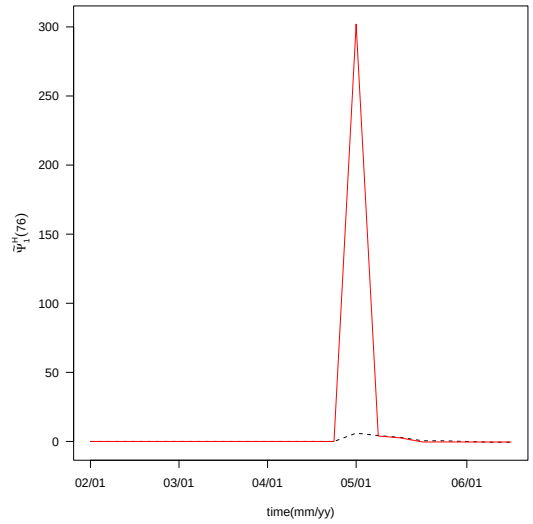

Figure 12: Locality statistics $\widetilde{\Psi}_1(76, H_t)$ during a period of 20 weeks in 2001, which is shown in red solid line (the black dotted line is $\widetilde{\Psi}_1(76, D_t)$). It shows a sudden value increase at $t^* = 130$ in May 2001.