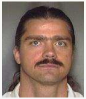
INTERFACE 2008

Scan Statistics on Enron Hypergraphs

Youngser Park*, Carey E. Priebe*, David J. Marchette†

*Johns Hopkins University, Baltimore, MD
†Naval Surface Warfare Center, Dahlgren, VA



Carey E. Priebe

David J. Marchette

May 24, 2008
Durham, NC

Citation

- C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, "Scan Statistics on Enron Graphs," *SIAM International Conference on Data Mining, Workshop on Link Analysis, Counterterrorism and Security*, Newport Beach, California, April 23, 2005
- C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, "Scan Statistics on Enron Graphs," *Computational and Mathematical Organization Theory*, Vol. 11, No. 3, pp. 229-247, 2005.
- http://www.ams.jhu.edu/~priebe/sseg.html

Introduction

Problem: Time series of graphs are becoming more and more common, *e.g.*, communication graphs, social networks, *etc.*, and methods for *statistical inferences* are required.

Objective: *To extend a theory of scan statistics on hypergraphs to perform change point / anomaly detection in graphs and in time series thereof.*

Hypotheses: H_0: homogeneity
H_A: local subregion of excessive activity

Scan Statistics

"moving window analysis":

to scan a small "window" (*scan region*) over data, calculating some *locality statistic* for each window; *e.g.*,

- number of events for a point pattern,
- average pixel value for an image,
- ...

scan statistic ≡ maximum of locality statistic:

If maximum of observed locality statistics is large, then the inference can be made that *there exists a subregion of excessive activity!*

Scan Statistics on Graphs

directed graph (digraph): $D = (V, A)$

order: $|V(D)|$

size: $|A(D)|$

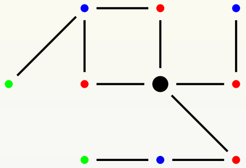
neighborhood: k^{th} order neighborhood of v:
$N_k[v; D] = \{w \in V(D) : d(v, w) \leqslant k\}$

scan region: (example: induced subdigraph): $\Omega(N_k[v; D])$

locality statistic: (example: size): $\Psi_k(v) = |A(\Omega(N_k[v; D]))|$

scan statistic: ("scale specific") $M_k(D) = \max_{v \in V(D)} \Psi_k(v)$

Example of Scan Statistics



scan	color	locality
0	●	4
1	● + ●	4
2	● + ● + ●	8
3	● + ● + ● + ●	10

Scan Statistics and Time Series

- Let $\{D_t\}$ $t = 1, \ldots, t_{max}$ be a time series of directed graphs.
- Scan region: induced subgraph of k-neighborhood: $\Omega(N_k(v; D_t))$.
- Locality statistic: $\Psi_{k,t}(v) = \text{size}(\Omega(N_k(v; D_t)))$.
- Scan statistic: $M_{k,t} = \max_v(\text{size}(\Omega(N_k(v; D_t))))$.
- Let τ be an integer (temporal window).

Scan Statistics and Time Series

Vertex Standardization

- We want to standardize the vertices ("loud" vertices don't drown out "quiet" ones).
- Vertex-dependent standardized locality statistic:

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

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

Scan Statistics and Time Series

Normalizing the Scan Statistic

- If we want to detect anomalies, we need to *detrend*.
- temporally-normalized scan statistics:

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

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

Scan Statistics and Time Series

Some Examples

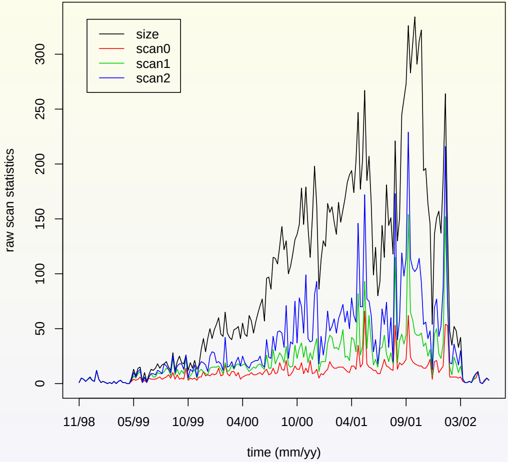


Figure: Time series scan statistics for weekly Enron email graphs.

Scan Statistics and Time Series

Some Examples

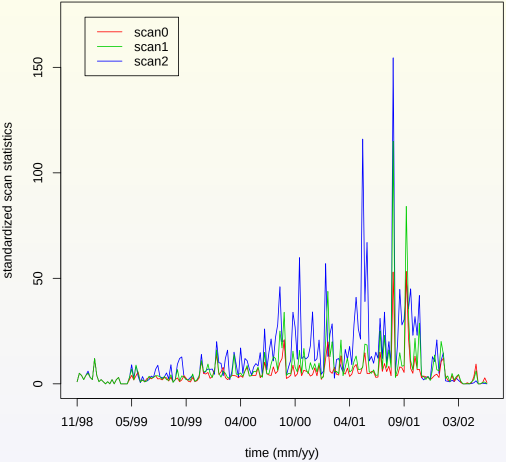


Figure: Time series of standardized scan statistics $\widetilde{M}_{k,t}(G)$ for $k = 0, 1, 2$.

Scan Statistics and Time Series

Some Examples

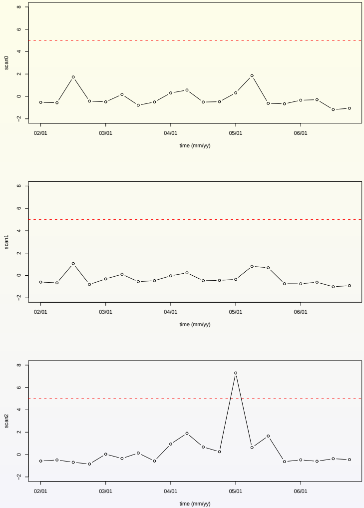


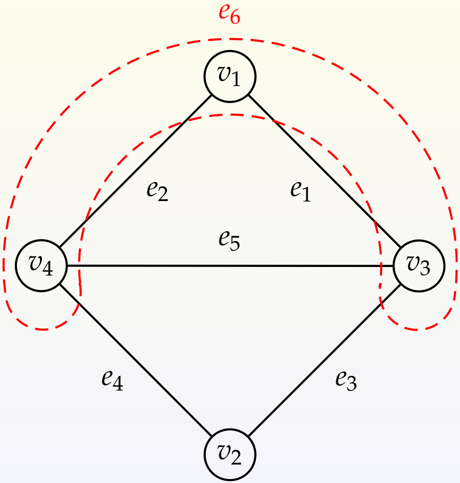
Figure: $S_{k,t}$, temporally-normalized scan statistics, on zoomed in time series of Enron email graphs.

Definition of Hypergraph

- A graph in which generalized edges (called *hyperedges*) may connect more than two vertices.
- A hypergraph $H = (V, \mathcal{E})$ consists of a set of vertices $V = \{v_1, \cdots, v_n\}$ and a set of hyperedges $\mathcal{E} = \{e_1, \cdots, e_m\}$, with $e_i \neq \emptyset$ and $e_i \subseteq V$ for $i = 1, \cdots, m$ [Berge89].

C. Berge,
Hypergraphs: Combinatorics of Finite Sets,
North-Holland, 1989.

Example of Hypergraph



Incidence Matrix

	e_1	e_2	e_3	e_4	e_5	e_6
v_1	1	1	0	0	0	1
v_2	0	0	1	1	0	0
v_3	1	0	1	0	1	1
v_4	0	1	0	1	1	1

Scan Statistics on Hypergraphs

hypergraph: $H = (V, \mathcal{E})$

order: $order(H) = |V| = n$,

size: $size(H) = |\mathcal{E}| = m$,

neighborhood: (1^{st}-order) $N_1(v, H) = \bigcup_{v \in e_i, e_i \in \mathcal{E}} e_i$,

neighborhood: (k^{th}-order) $N_k(v, H) = \bigcup_{v \in N_{k-1}(v,H)} N_1(v, H)$ for $k \geqslant 2$,

induced subgraph: $\Omega(N_k(v, H))$, where $\mathcal{E}_k = \{e_i \in \mathcal{E} : e_i \subset N_k\}$,

Scan Statistics on Hypergraphs

hypergraph: $H = (V, \mathcal{E})$

locality statistic: $\Psi_k(v, H) = size(\Omega(N_k(v, H)))$, for $k > 1$,

locality statistic: (vertex-dependent standardized)

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

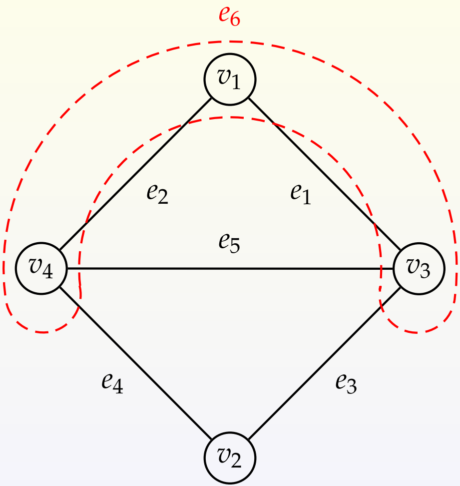
scan statistic: ("scale-specific") $M_k(H) = \max_{v \in V(H)} \Psi_k(v, H)$.

scan statistic: (standardized) $\widetilde{M}_{k,t}(H) = \max_v \widetilde{\Psi}_{k,t}(v, H)$.

scan statistic: (temporally-normalized)

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

Scan Statistics on Hypergraphs



locality statistic

	Ψ_0	Ψ_1	Ψ_0	Ψ_1
v_1	2	3	4	5
v_2	2	3	2	3
v_3	3	5	6	7
v_4	3	5	6	7

$$\mathbf{v_1} \neq \mathbf{v_2}$$

Enron Graphs

- Energy company famous for "creating accounting" measures to boost stock value.
- Email sent and received between executives at Enron over a period of about 2 years.
- 150 executives (184 email addresses – some duplication).
- From-To pairs extracted from the headers of the email to construct a communications graph:
  - Each graph covers one week (non-overlapping).
  - Vertices correspond to email addresses.
  - An edge between u and v if v sent an email with v in the To or CC field during the week.
  - Duplicates not counted.

Experiments 1

Detection by raw scan statistics

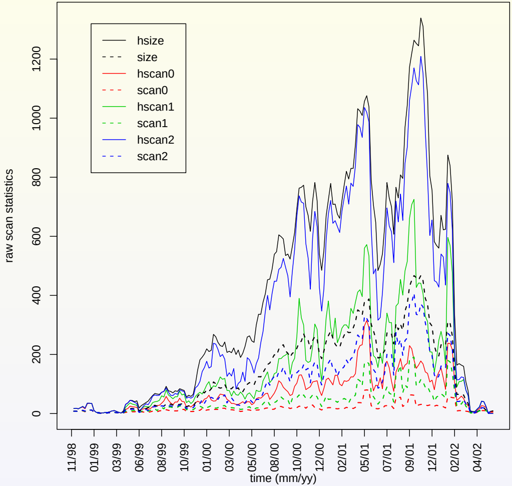


Figure: Time series scan statistics for weekly Enron email graphs.

Experiments 1
Detection by raw scan statistics

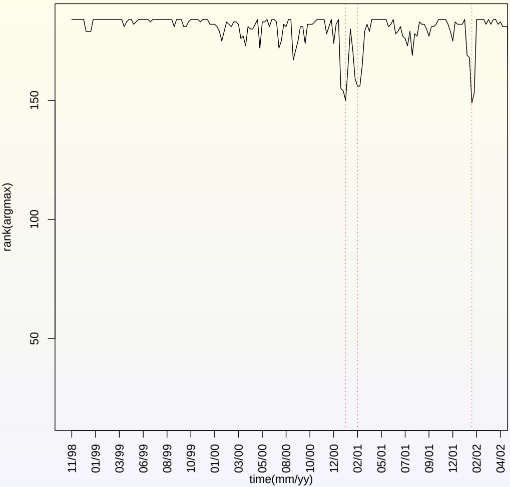


Figure: k_t vs. t for $\Psi_{1,t}(G)$ and $\Psi_{1,t}(H)$.

Experiments 2

Detection by normalized scan statistics

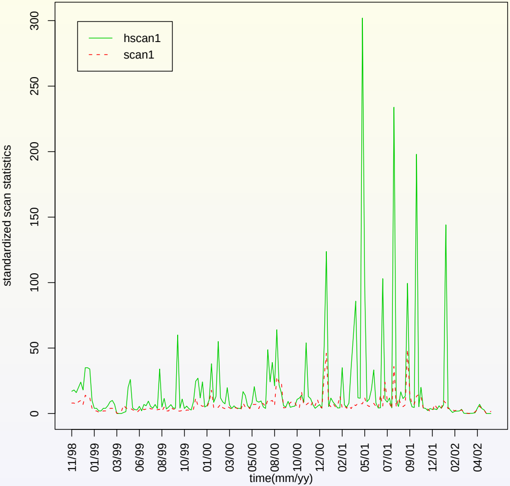


Figure: Time series of standardized scan statistics $\widetilde{M}_{1,t}(G)$ and $\widetilde{M}_{1,t}(H)$.

Experiments 2

Detection by normalized scan statistics

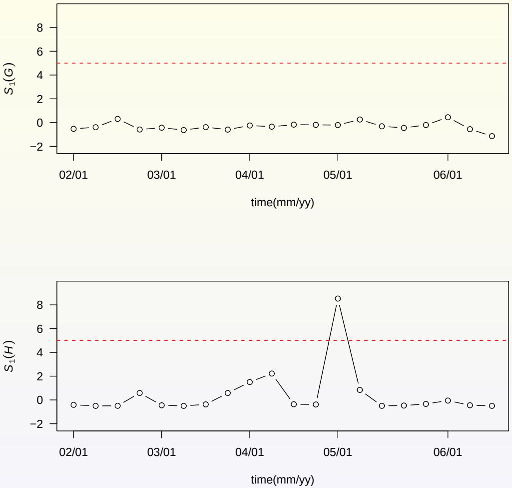


Figure: Time series of temporally-normalized scan statistics $S_{1,t}(G)$ and $S_{1,t}(H)$. It shows that $t^* = 130$ and $v^* = \arg\max_v \widetilde{\Psi}_{1,t^*=130}(H) = 76$.

Experiments 2
Comparison of scan statistics

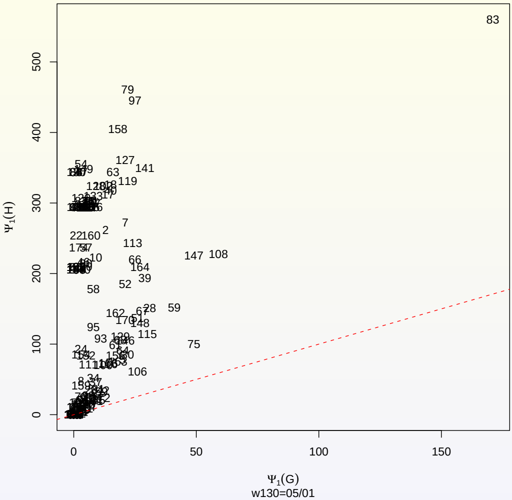
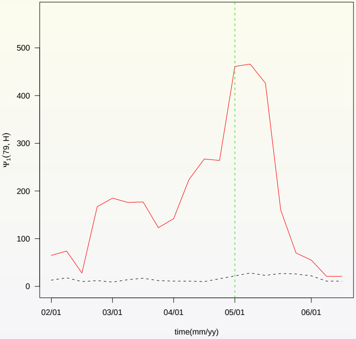


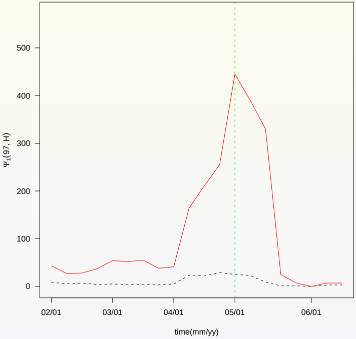
Figure: Locality statistics Ψ_1^H on hypergraph as a function of Ψ_1 on graph at week 130 (= May, 2001).

Experiments 2

Comparison of scan statistics



employee #79



employee #97

Experiments 2

Comparison of scan statistics

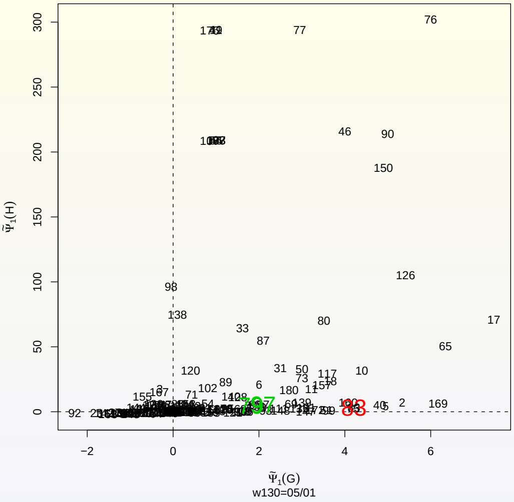
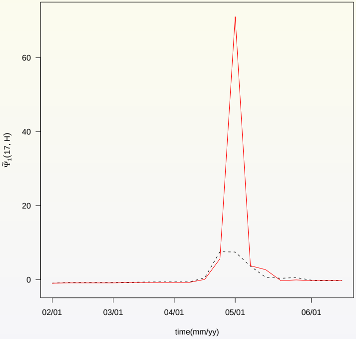


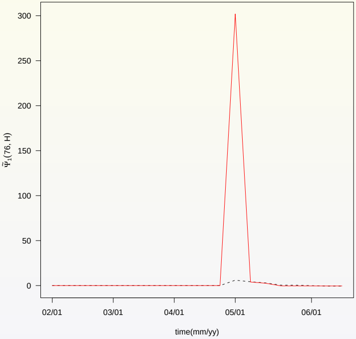
Figure: Standardized scan statistics $\widetilde{\Psi}_1^H$ on hypergraph as a function of $\widetilde{\Psi}_1$ on graph at week 130 (= May, 2001).

Experiments 2

Comparison of scan statistics



employee #17



employee #76

Discussion / Future Works

- Weighted/Directed Hypergraph
- Content Analysis
- Real-time Data