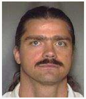
IASC 2008

Statistical Modeling for Computer Security

Anomaly Detection using Scan Statistics on Enron Graphs and Hypergraphs

Youngser Park

Johns Hopkins University, Baltimore, MD



Carey E. Priebe

David J. Marchette

December 1–3, 2008

Seoul, Korea

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.

- Y. Park, C.E. Priebe, D.J. Marchette, A. Youssef,
  "Scan Statistics on Enron Hypergraphs", submitted to
  SIAM International Conference on Data Mining,
  Sparks, Nevada, April 30, 2009.

- http://www.cis.jhu.edu/~parky/Enron

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 develope and apply a theory of scan statistics on graphs and 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" [1992 R.A. Fisher, 1965 J. Naus]:

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,
- number of email messages, ...

scan statistic $\equiv$ 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 $\rightarrow$ *detection!*

Scan Statistics on Graphs

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

order: $|V(D)|$

size: $|A(D)|$

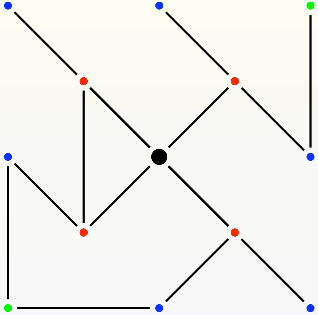
neighborhood: k^{th} order neighborhood of v:
$N_k[v; D] = \{w \in V(D) : d(v, w) \leqslant k\}$

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	● + ●	5
2	● + ● + ●	11
3	● + ● + ● + ●	14

Erdős-Rényi Random Graphs

The simplest and most common type of random graph is the Erdos-Renyi random graph [Bollobas02]:

- Given a probability p
- Place an edge $v_i v_j$ between vertices v_i and v_j with probability p.

Bollobas, B.,
Random Graphs, 2nd ed.,
Cambridge University Press, 2001.

Social Network Motivation

- A social network is a set of vertices corresponding to “actors” (individual entities) and edges representing relationships.
- Out intuition is that actors with similar interests should be related: In some sense, the probability of an edge should be proportional to the amount of overlap of interests. (e.g., religion, education, sports, ...)

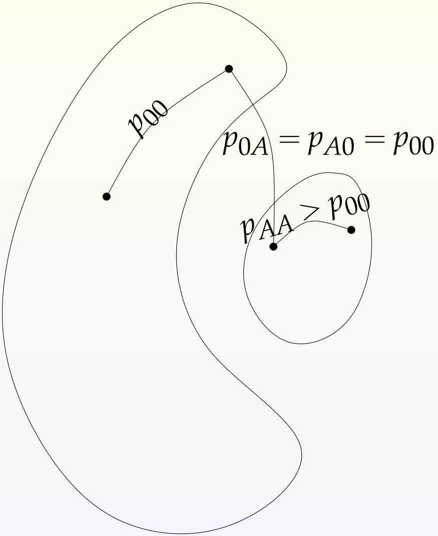
Random Dot Product Graphs

[Scheinerman07]:

- Each vertex v_i has associated with it a vector x_i.
- Place an edge $v_i v_j$ between vertices v_i and v_j with probability proportional to $x_i x_j$, the dot product of x_i and x_j.
- Thus $p_{ij} = f(x_i, x_j)$. e.g., identity function.
- The edges in the random graph are no longer independent.
- Further, this can be interpreted in the manner of our social network motivation.

Ed Scheinerman and Kimberly Tucker,
Modeling graphs using dot product representations,
Computational Statistics, 2007.

"Kidney-Egg Model"



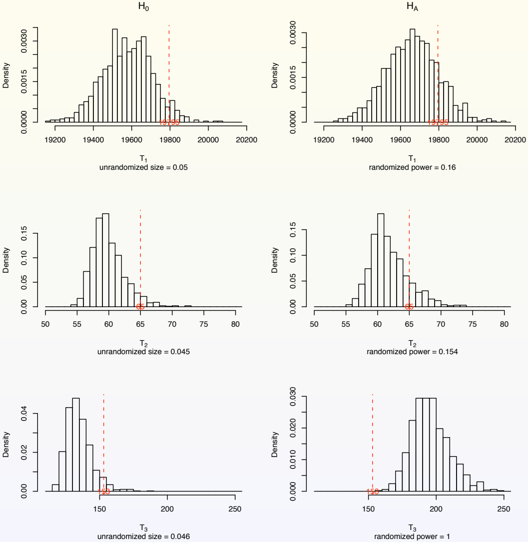
Monte Carlo Simulation

$n = 1000, \;\; n' = 13, \;\; \alpha = 0.05, \;\; MC = 1000$

$T = size$ $\qquad$ $\beta = 0.160$

$T = scan0$ $\qquad$ $\beta = 0.154$

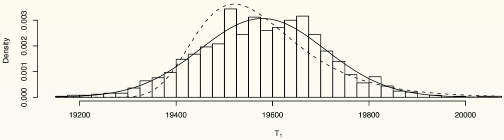
$T = scan1$ $\qquad$ $\beta = 1.000$



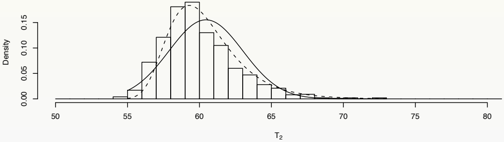
Gumbel Conjecture

$n = 1000,\ MC = 1000$

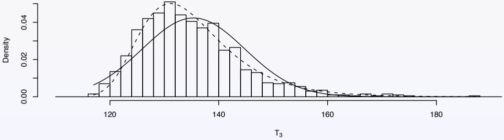
$T = size$



$T = scan0$

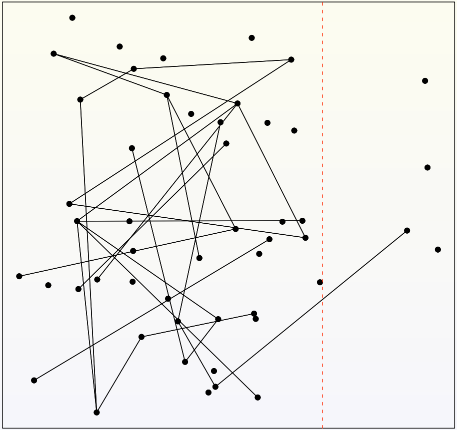


$T = scan1$



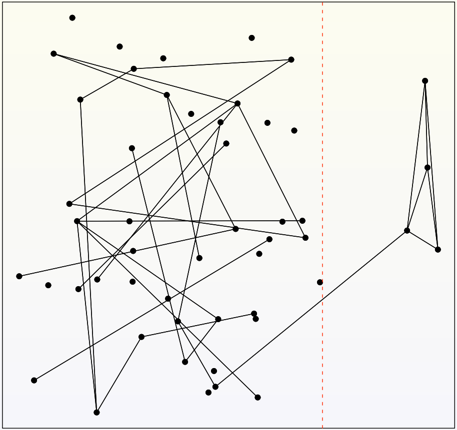
Example: H_0

$size = 26, \; scan0 = 5, \; scan1 = 5, \; scan2 = 11$



Example: H_{A_1}

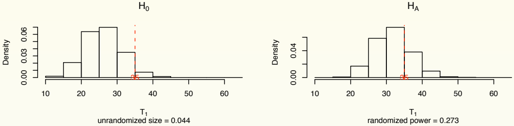
$size = 32, \; scan0 = 5, \; scan1 = 7, \; scan2 = 11.$



Example: Monte Carlo Simulation H_0 vs H_{A_1}

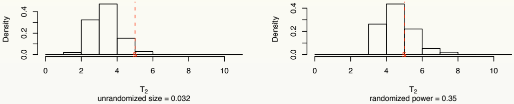
$n = 50, \;\; n' = 4, \;\; \alpha = 0.05, \;\; MC = 1000$

$T = size$



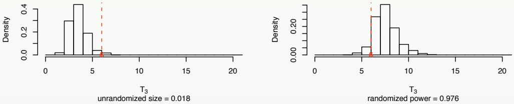
$\beta = 0.273$

$T = scan0$



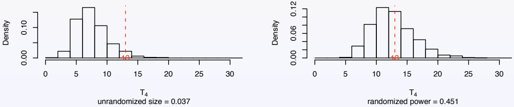
$\beta = 0.350$

$T = scan1$



$\beta = 0.976$

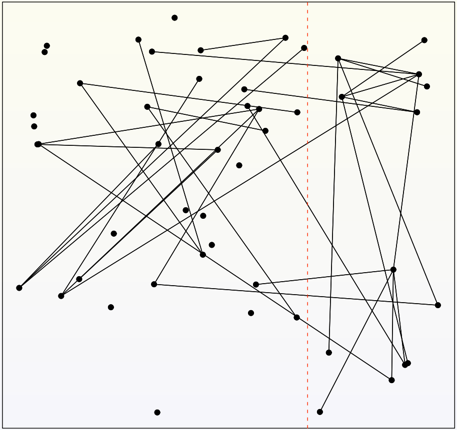
$T = scan2$



$\beta = 0.451$

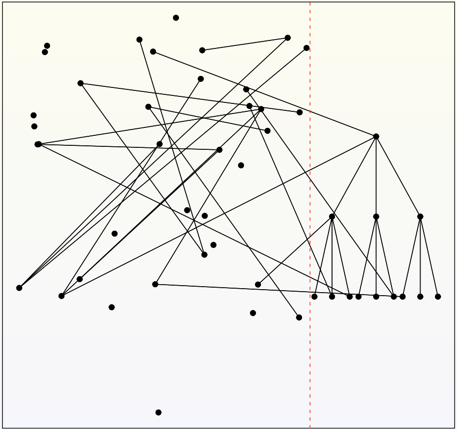
Example: H_{A_2}

$size = 34,\ scan0 = 5,\ scan1 = 5,\ scan2 = 17$



Example: H_{A_2}

$size = 34, \; scan0 = 5, \; scan1 = 5, \; scan2 = 17$



Example: Monte Carlo Simulation H_0 vs H_{A_2}

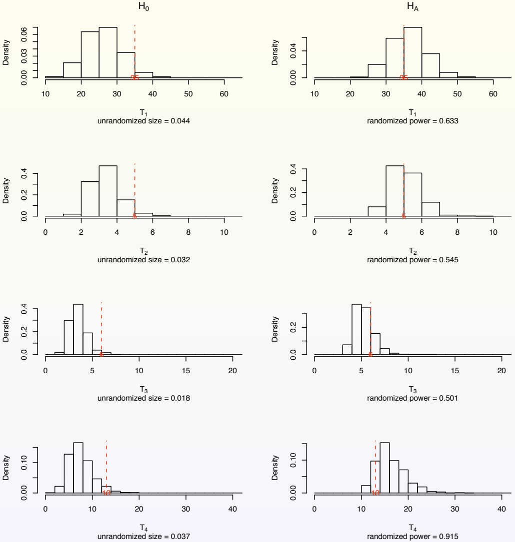
$n = 50,\ n' = 13,\ \alpha = 0.05,\ MC = 1000$

$T = size$ $\quad \beta = 0.633$

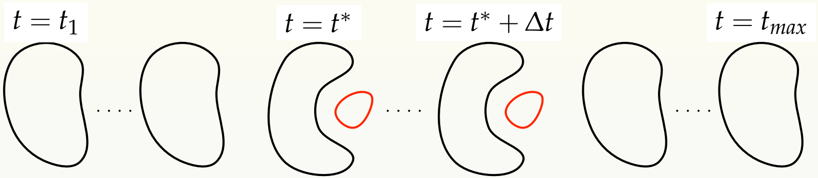
$T = scan0$ $\quad \beta = 0.545$

$T = scan1$ $\quad \beta = 0.501$

$T = scan2$ $\quad \beta = 0.915$



Time Series "Kidney-Egg Model"



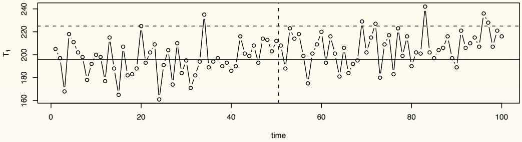
Scan Statistics and Time Series

- 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: $M_{k,t} = \max_v(\Psi_{k,t}(v))$.

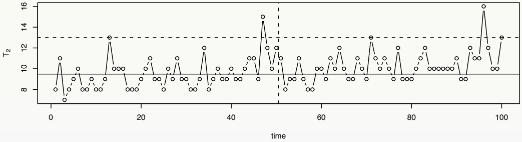
Time Series

$n = 100,\ n' = 4,\ \alpha = 0.05$

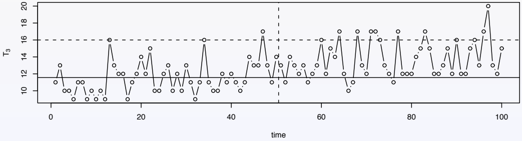
$T = size$



$T = scan0$



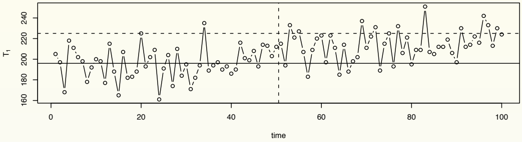
$T = scan1$



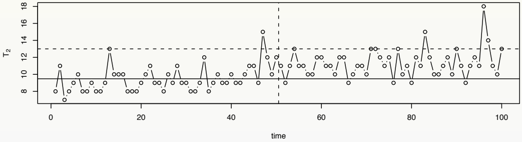
Time Series

$n = 100, n' = 6, \alpha = 0.05$

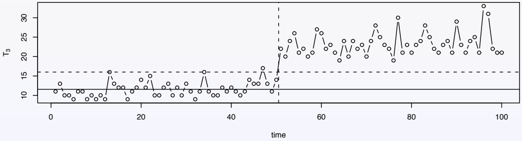
$T = size$



$T = scan0$



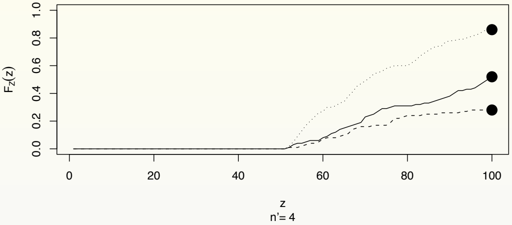
$T = scan1$



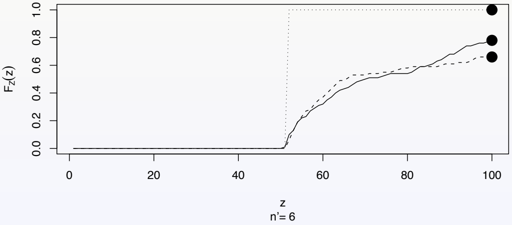
Time Series

$n = 100,\ n' = 4\ \&\ 6,\ \alpha = 0.05$

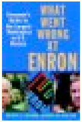
$n = 4$



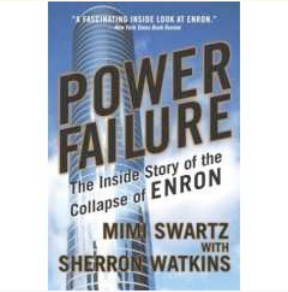
$n = 6$

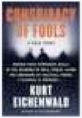


Enron


WHAT WENT WRONG AT ENRON


Anatomy of Greed


"A FASCINATING INSIDE LOOK AT ENRON."
POWER FAILURE
The Inside Story of the Collapse of ENRON
MIMI SWARTZ
WITH
SHERRON WATKINS


CONSPIRACY OF FOOLS
KURT EICHENWALD


ENRON
LOREN FOX

Enron Email 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 4 years (189 weeks).
- 125,409 distinct messages from 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, CC, or BCC field during the week.
  - Duplicates not counted.

Enron Email Graphs
Examples

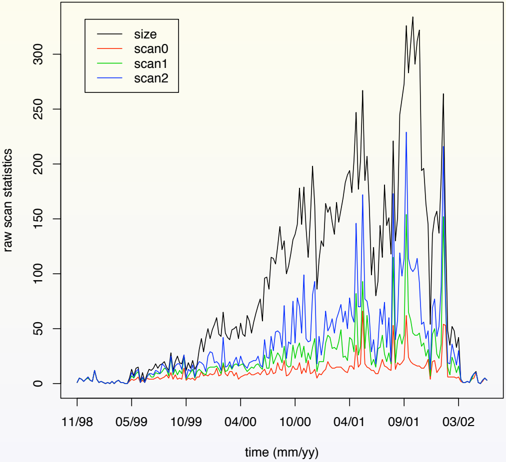


Figure: Time series scan statistics for weekly Enron email graphs.

Scan Statistics and Times Series

Vertex Standardization

- We want to standardize the vertices ("loud" vertices don't drown out "quiet" ones).
- Let τ be an integer (temporal window).
- Vertex-dependent standardized locality statistic:

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

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

Enron Email Graphs
Examples

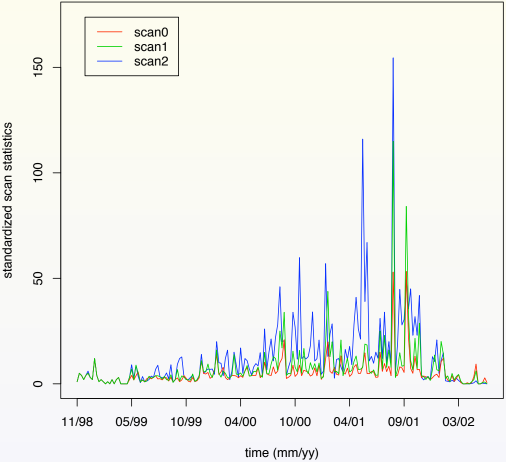


Figure: Time series of standardized scan statistics $\widetilde{M}_{k,t}(G)$ for $k = 0, 1, 2$.

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.
- detection: time t such that $S_{k,t} > 5$

Scan Statistics and Time Series

Examples

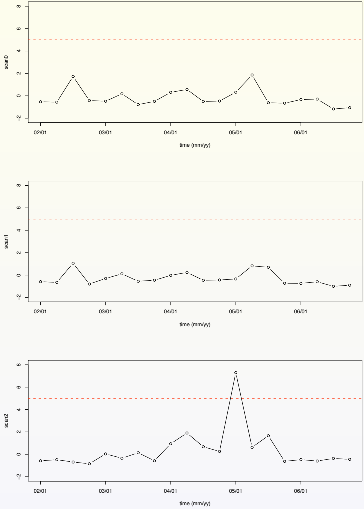


Figure: $S_{k,t}$, temporally-normalized scan statistics, on zoomed in time series of Enron email graphs.

Detection Graph D_{132}

Details for the 'detection' graph D_{132}

time t^*	132 (week of May 17, 2001)		
$size(D_{132})$	267		
scale k	$M_{k,132}$	$\widetilde{M}_{k,132}$	$S_{k,132}$
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		

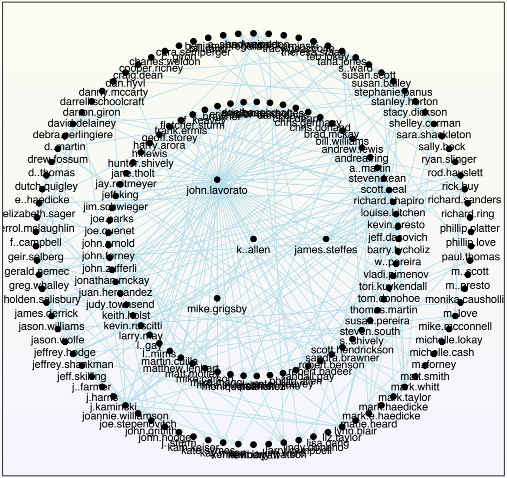
Enron Email Graphs

A detection graph

$\arg\max_v \Psi_{0,132}(v) = \texttt{john.lavorato}$ $\arg\max_v \widetilde{\Psi}_{0,132}(v) = \texttt{richard.shapiro}$

$\arg\max_v \Psi_{1,132}(v) = \texttt{john.lavorato}$ $\arg\max_v \widetilde{\Psi}_{1,132}(v) = \texttt{joannie.williamson}$

$\arg\max_v \Psi_{2,132}(v) = \texttt{richard.shapiro}$ $\arg\max_v \widetilde{\Psi}_{2,132}(v) = \texttt{k..allen}$



Anomaly Detection (Aliasing)

- $v^* = \arg\max_v \widetilde{\Psi}_{2,132}(v) = \texttt{k..allen}$
- `k..allen == phillip.allen?`
  - `k..allen` had no activity before $t^* = 132$.
  - At $t^* = 132$, `phillip.allen` switched to `k..allen`.
- Matched Filter:
  - For each vertex $v \in V \setminus \{v^*\}$,

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

- Is this a detection we want?

New York Times (May 22, 2005)

The New York Times **Week in Review**

NYTimes.com | Go to a Section

Site Search: NYT Since 1996 | Submit

Enron Offers an Unlikely Boost to E-Mail Surveillance

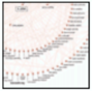
By GINA KOLATA
Published: May 22, 2005

E-Mail This
Printer-Friendly
Reprints

AS an object of modern surveillance, e-mail is both reassuring and troubling. It is a potential treasure trove for investigators monitoring suspected terrorists and other criminals, but it also creates the potential for abuse, by giving businesses and government agencies an efficient means of monitoring the attitudes and activities of employees and citizens.

Multimedia

GRAPHIC



Finding Patterns in Corporate Chatter

Now the science of e-mail tracking and analysis has been given a unlikely boost by a bitter chapter in the history of corporate malfeasance - the Enron scandal.

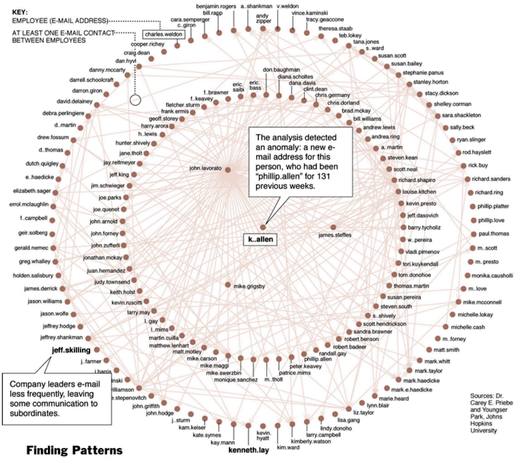
In 2003, the Federal Energy

New York Times (May 22, 2005)

The New York Times > Week in Review > Image > Finding Patterns in Corporate Chatter 05/23/2005 08:32 AM

The New York Times

May 21, 2005



Sources: Dr. Carey E. Priebe and Youngser Park, Johns Hopkins University

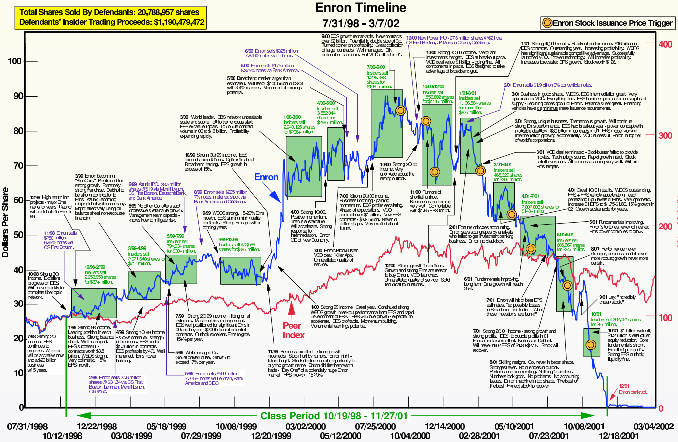
Finding Patterns In Corporate Chatter

Computer scientists are analyzing about a half million Enron e-mails. Here is a map of a week's e-mail patterns in May 2001, when a new name suddenly appeared. Scientists found that this week's pattern differed greatly from others, suggesting different conversations were taking place that might interest investigators. Next step: word analysis of these messages.

Bill Marsh/The New York Times

Close Window

Enron Timeline



Anomaly Detection (another)

- Non-zero activity: $\widetilde{\Psi}_{k,t}(v) \cdot I\{\widehat{\mu}_{0,t,\tau}(v) > c\}$
  - For $c = 1$, $v^* =$ `roy.hayslett` 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]

Anomaly Detection (another)

- `roy.hayslett` communicates with `sally.beck`, who is a $k = 0$ detection!

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]

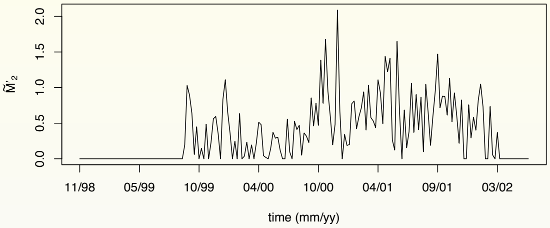
Anomaly Detection (chatter)

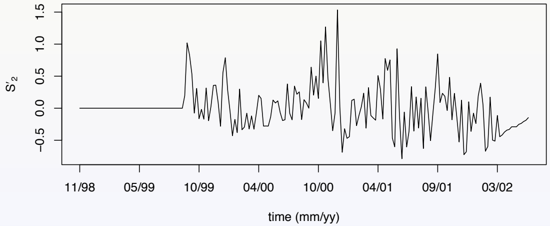
- Seek a detection in which the excess activity is due to chatter amongst the 2-neighbors!

$$\widetilde{\Psi}'_t(v) = \left(\widetilde{\Psi}_{2,t}(v) \cdot \mathcal{I}_{t,\tau}(v)\right) / \max(\gamma_t(v), 1)$$

$$\begin{aligned}
\mathcal{I}_{t,\tau}(v) =& I_1 \times I_2 \times I_3 \\
I_1 =& I\{\widehat{\mu}_{0,t,\tau} > c_1\}, \\
I_2 =& I\{\Psi_0(v) < \widehat{\sigma}_{0,t,\tau}(v) c_2 + \widehat{\mu}_{0,t,\tau}(v)\}, \\
I_3 =& I\{\Psi_1(v) < \widehat{\sigma}_{1,t,\tau}(v) c_3 + \widehat{\mu}_{1,t,\tau}(v)\}.
\end{aligned}$$

Anomaly Detection (chatter)





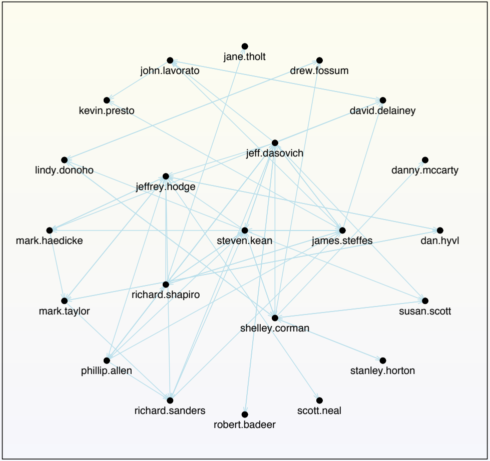
Anomaly Detection (chatter)

- $(v^*, t^*) = (\texttt{steven.kean}, 109)$

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]

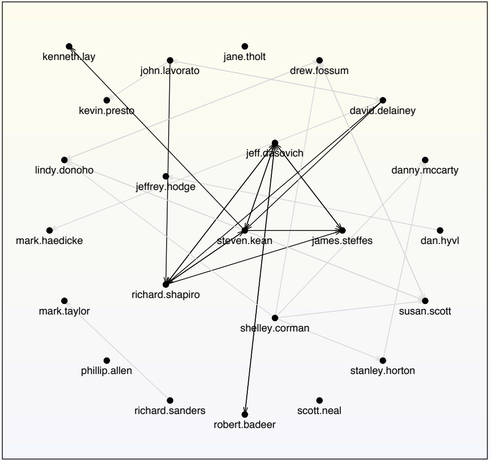
Anomaly Detection (chatter)

Ω_{109}



Anomaly Detection (chatter)

Ω_{108}

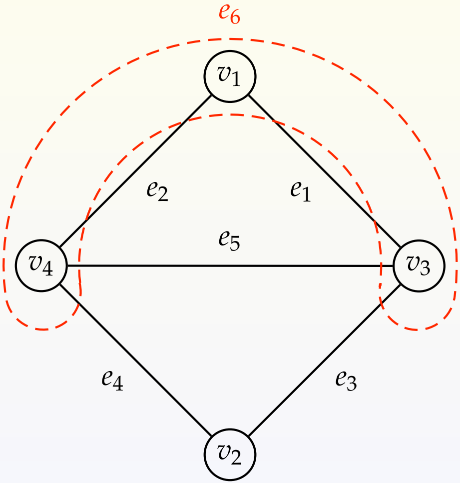


Definition of Hypergraph

- A graph in which generalized edges (called *hyperedges*) may connect more than two vertices.
  $\Rightarrow$ better suited to email data than a graph!
- 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 $k > 1$,

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

locality statistic: (vertex-dependent standardized)

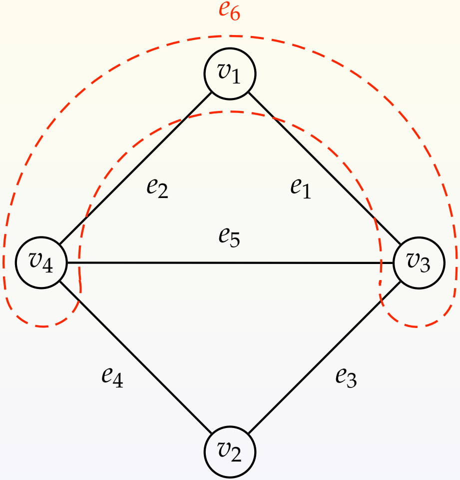
$$\widetilde{\Psi}_{k,t}(v, H) = \frac{\Psi_{k,t}(v, H) - \widehat{\mu}_{k,t,\tau}(v)}{\max(\widehat{\sigma}_{k,t,\tau}(v), 1)}$$

scan statistic: (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

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

Experiments 1
Detection by raw scan statistics

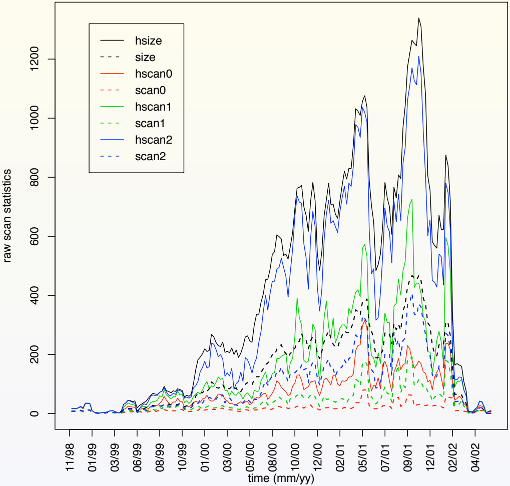


Figure: Time series scan statistics for weekly Enron email graphs.

Experiments 1

Detection by raw scan statistics

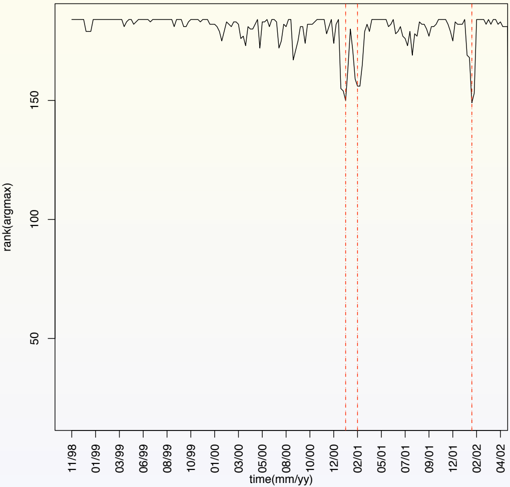


Figure: k_t vs. t for $\Psi_{1,t}(G)$ and $\Psi_{1,t}(H)$.

Experiments 2

Detection by normalized scan statistics

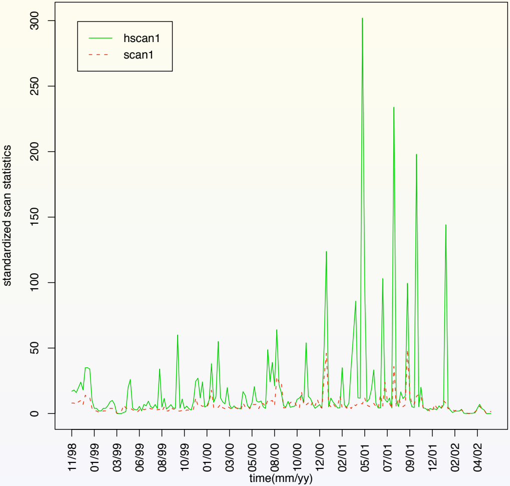


Figure: Time series of standardized scan statistics $\widetilde{M}_{1,t}(G)$ and $\widetilde{M}_{1,t}(H)$.

Experiments 2

Detection by normalized scan statistics

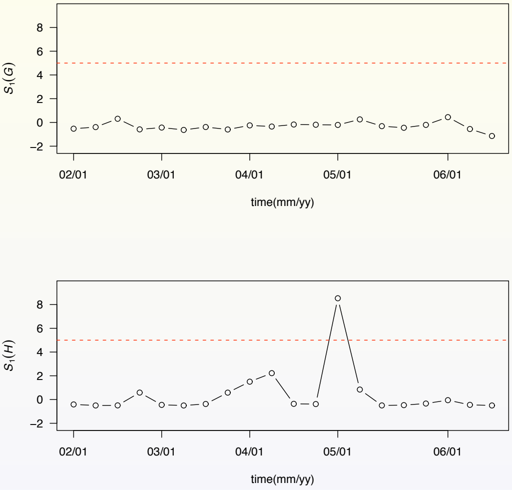


Figure: Time series of temporally-normalized scan statistics $S_{1,t}(G)$ and $S_{1,t}(H)$. It shows that $t^* = 130$ and $v^* = \arg\max_v \widetilde{\Psi}_{1,t^*=130}(H) = 76$.

Experiments 3

Comparison of scan statistics

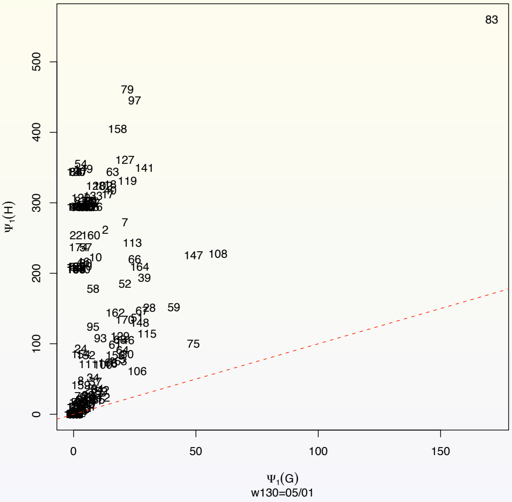
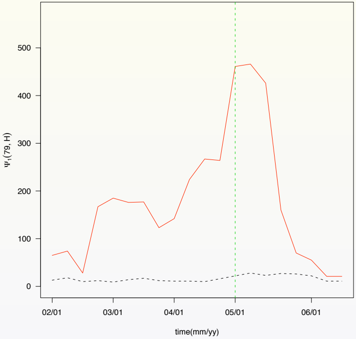


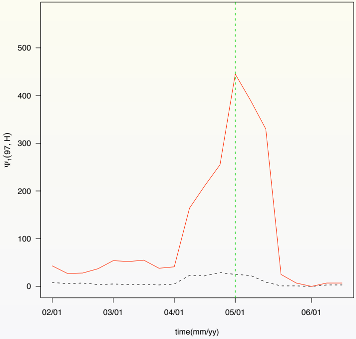
Figure: Locality statistics $\Psi_1(H)$ on hypergraph as a function of $\Psi_1(G)$ on graph at week 130 (= May, 2001).

Experiments 3

Comparison of scan statistics



employee #79



employee #97

Experiments 3

Comparison of scan statistics

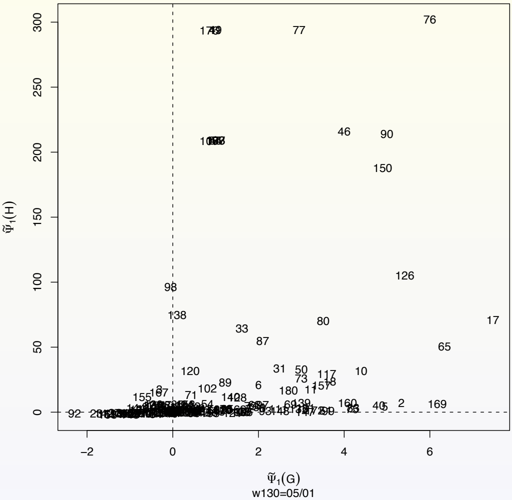
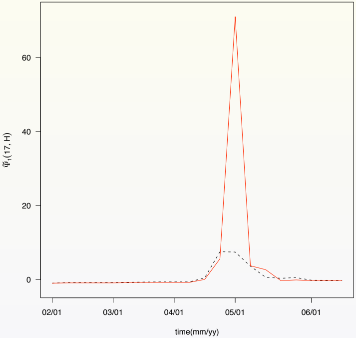


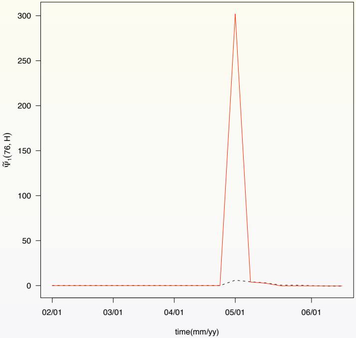
Figure: Standardized scan statistics $\widetilde{\Psi}_1(H)$ on hypergraph as a function of $\widetilde{\Psi}_1(G)$ on graph at week 130 ($=$ May, 2001).

Experiments 3

Comparison of scan statistics



employee #17



employee #76

Discussion / Future Works

- Scan Statistics offers promise for detecting anomalies in time series of graphs and hypergraphs.
- Weighted/Directed Hypergraph
- Content Analysis
- Real-time Data (streaming graphs)

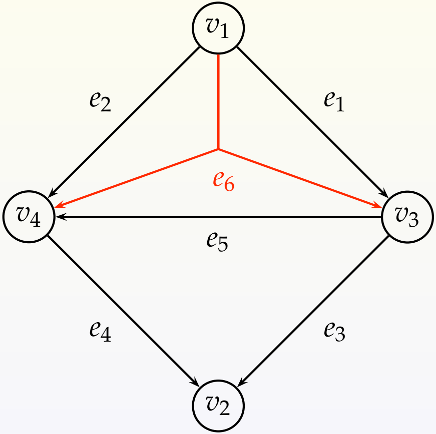
- http://www.cis.jhu.edu/~parky/Enron

Definition of Directed Hypergraph

- A hypergraph $H^D = (V, \mathcal{A})$ consists of a set of vertices $V = \{v_1, \cdots, v_n\}$ and a set of directed hyperedges $\mathcal{A} = \{e_1, \cdots, e_m\}$, with $e_i \neq \emptyset$ and $e_i \subseteq V$ for $i = 1, \cdots, m$.
- A *directed* hyperedge or *hyperarc* is an ordered pair, $e_i = (T, H)$, of disjoint subsets of vertices; T is the *tail* while H is the *head* of e_i.
- Incidence matrix of H^D is a $n \times m$ matrix $[a_{ij}]$:

$$a_{ij} = \begin{cases} 1 & \text{if } v_i \in T(e_j), \\ -1 & \text{if } v_i \in H(e_j), \\ 0 & \text{otherwise.} \end{cases}$$

Directed 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

- $size(H^D) = |\mathcal{A}| = 6$,
- $\Psi_0(v, H^D) =$ outdegrees $=$ $\{\#$ of 1's$\} = \{3, 0, 2, 1\}$,
- $\Psi_k(v, H^D) = |\Omega(N_k(v, H^D))|$ for $k > 0$.