Scan Statistics on Enron Graphs

Carey E. Priebe

cep@jhu.edu

Johns Hopkins University

Department of Applied Mathematics & Statistics

Department of Computer Science

Center for Imaging Science

"The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus."

— Leopold Kronecker to Hermann von Helmholtz

Citation

C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park,
"Scan Statistics on Enron Graphs,"
Computational and Mathematical Organization Theory,
to appear.

SIAM International Conference on Data Mining
Workshop on Link Analysis, Counterterrorism and Security
Newport Beach, California
April 23, 2005

http://www.ams.jhu.edu/∼priebe/sseg.html

Outline

- Scan Statistics on Graphs, and on Time Series thereof
- Scan Statistics on Enron Graphs

Introduction

- Objective: *to develop and apply a theory of scan statistics on random graphs to perform change point / anomaly detection in graphs and in time series thereof.*
- Hypotheses:
  - H_0: homogeneity
  - H_A: local subregion of excessive activity

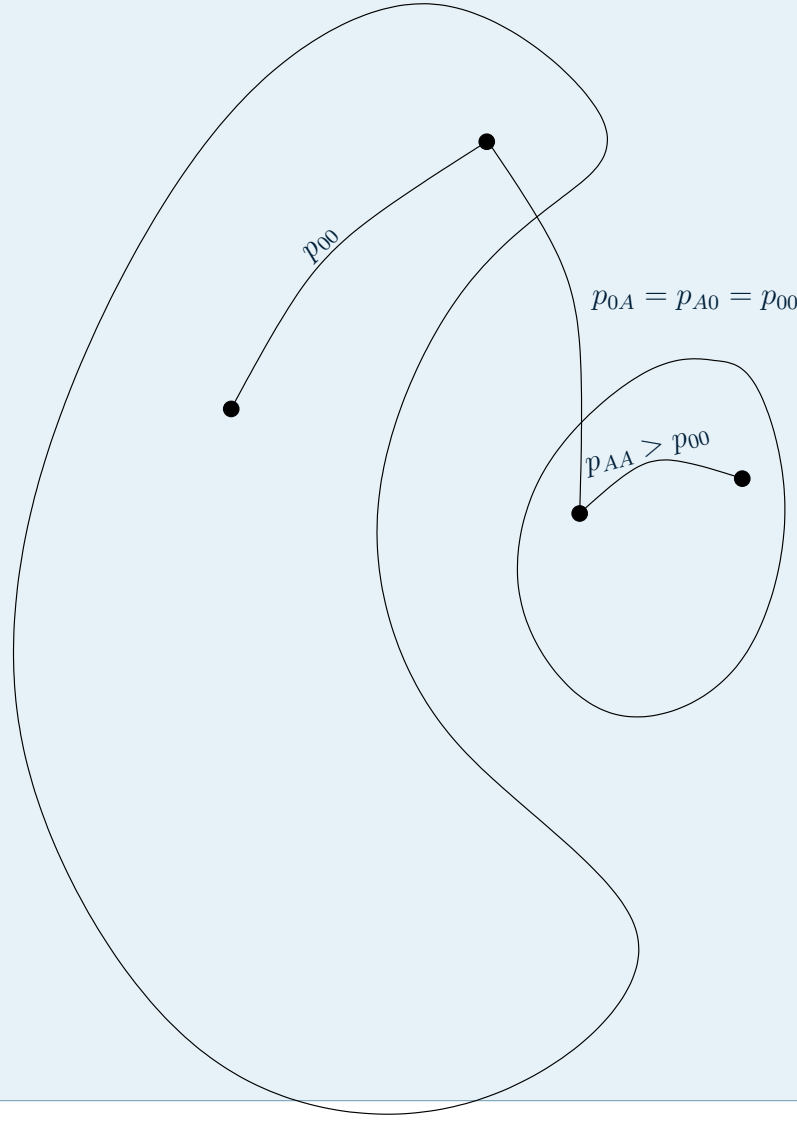
Scan Statistics

- "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 $\equiv$ 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.

Scan Statistics on Graphs

- directed graph (digraph): $D = (V, A)$
  - order: $|V(D)|$
  - size: $|A(D)|$
  - k-th order neighborhood of v:
    $N_k[v; D] = \{w \in V(D) : d(v, w) \leq k\}$
  - scan region (example: induced subdigraph): $\Omega(N_k[v; D])$
  - locality statistic (example: size): $\Psi_k(v) = |A(\Omega(N_k[v; D]))|$
  - "scale-specific" scan statistic: $M_k(D) = \max_{v \in V(D)} \Psi_k(v)$

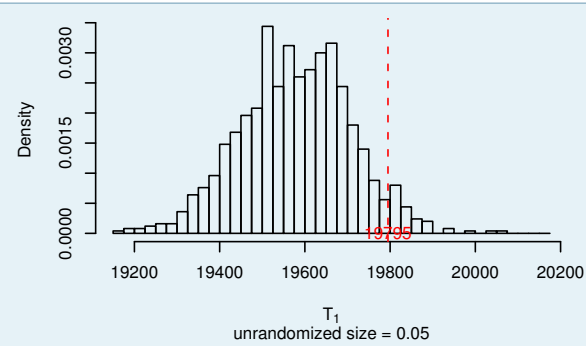
H_{A_1}

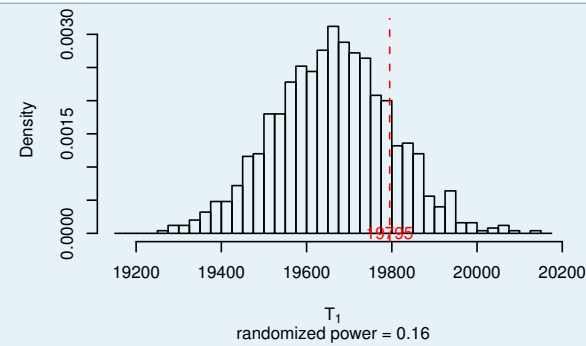


Monte Carlo Simulation

H_0 H_A

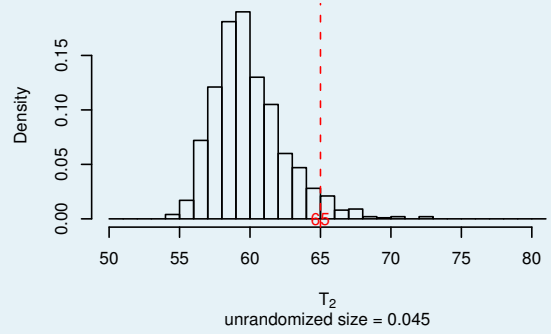
$T = size$

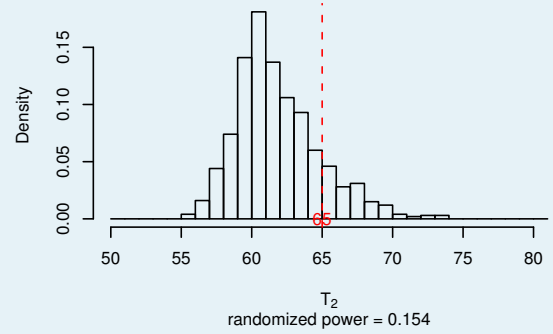




$\beta = 0.160$

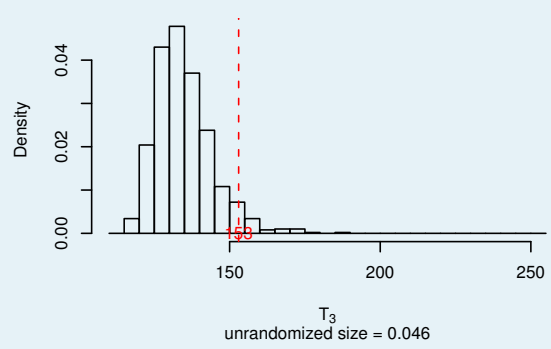
$T = scan0$





$\beta = 0.154$

$T = scan1$

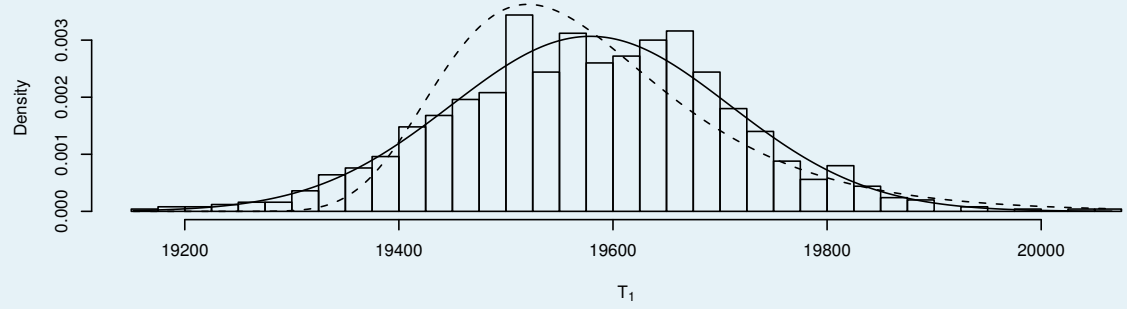


$\beta = 1.000$

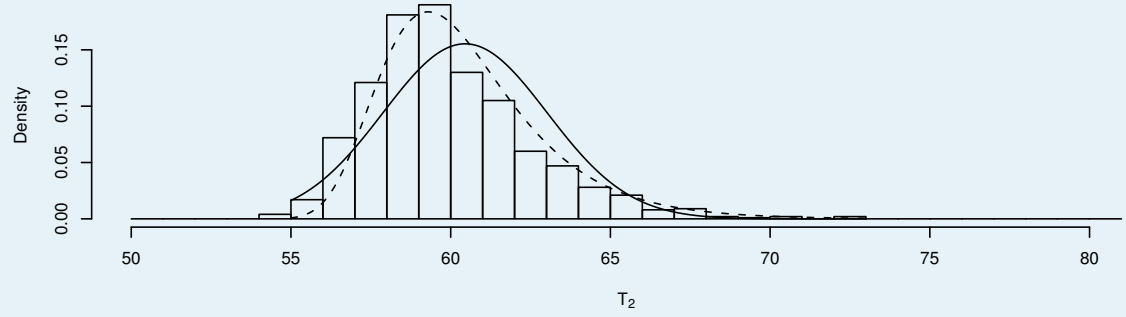
$$n = 1000, n' = 13, \alpha = 0.05, MC = 1000$$

Gumbel Conjecture

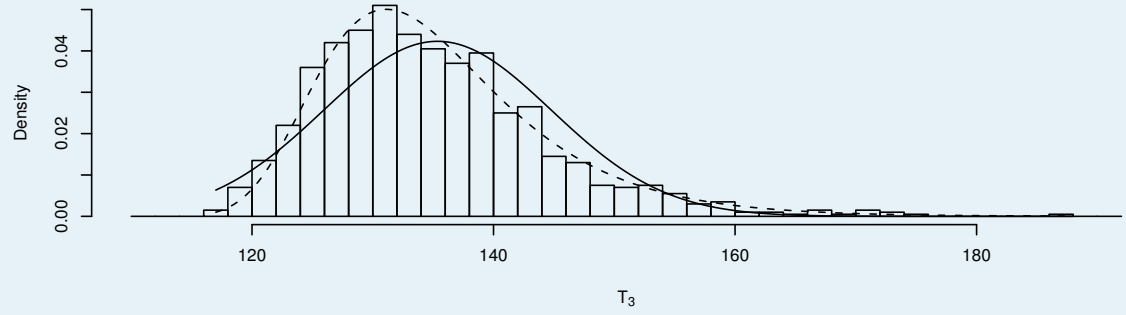
$T = size$



$T = scan0$

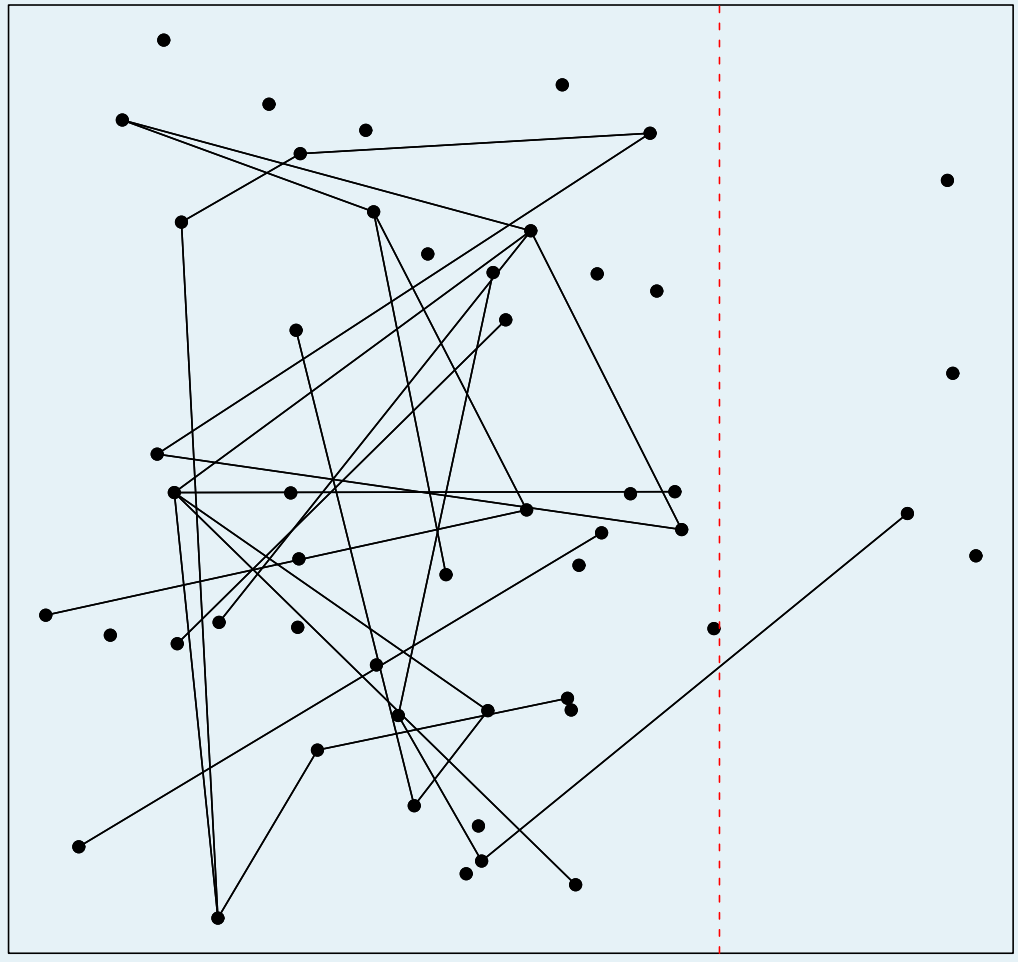


$T = scan1$



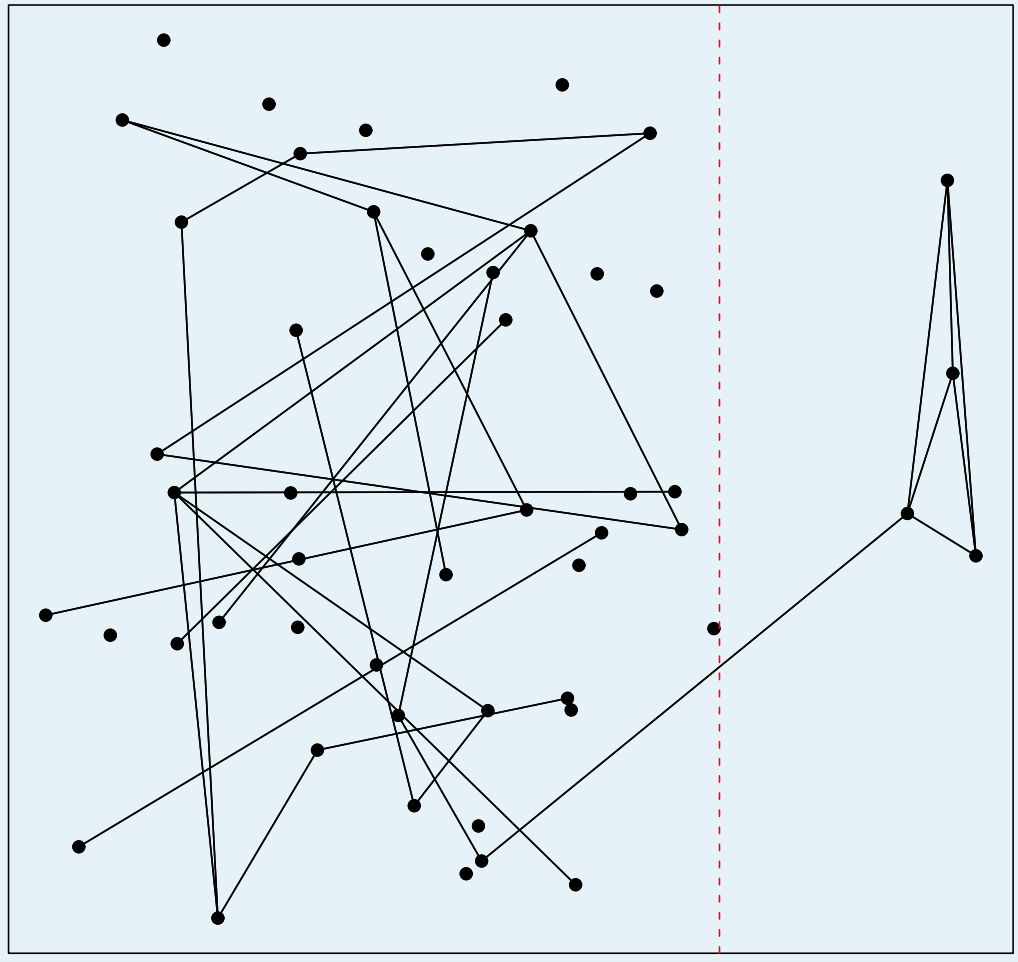
$n = 1000, MC = 1000$

Example: H_0



$$size = 26, scan0 = 5, scan1 = 5, scan2 = 11.$$

Example: H_{A_1}

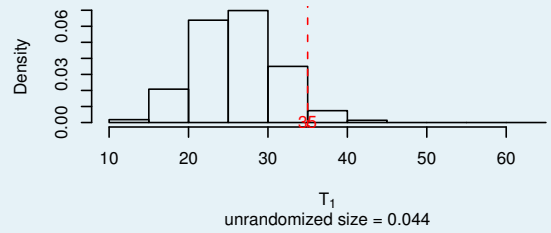


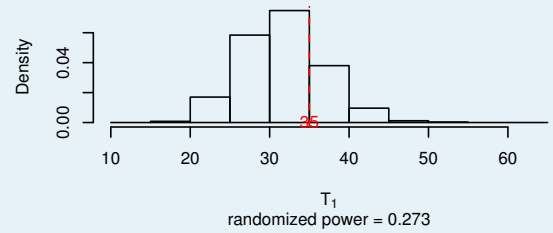
$$size = 32, scan0 = 5, scan1 = 7, scan2 = 11.$$

Example: Monte Carlo Simulation H_0 vs H_{A_1}

H_0 H_A

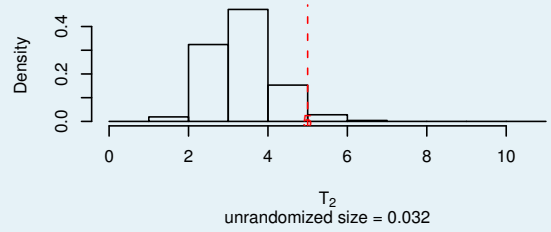
$T = size$

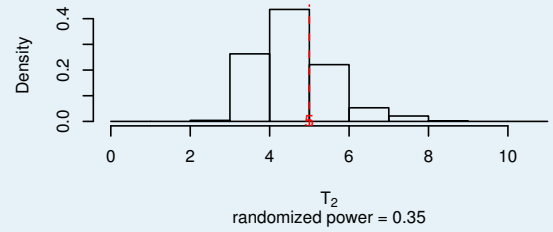




$\beta = 0.273$

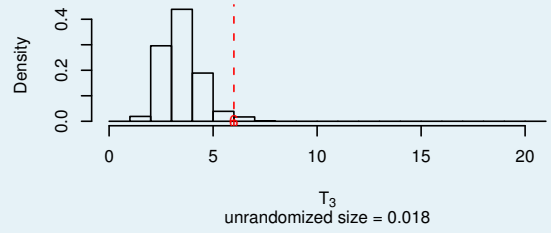
$T = scan0$

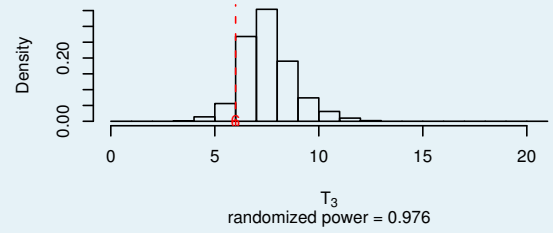




$\beta = 0.350$

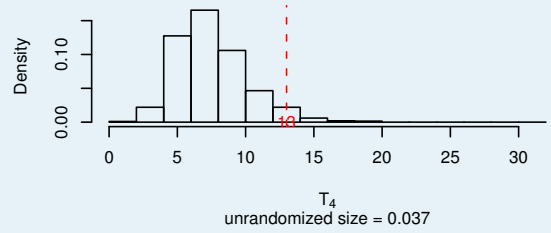
$T = scan1$

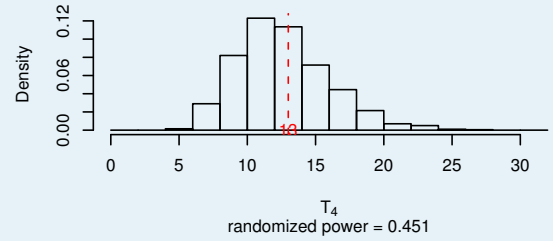




$\beta = 0.976$

$T = scan2$





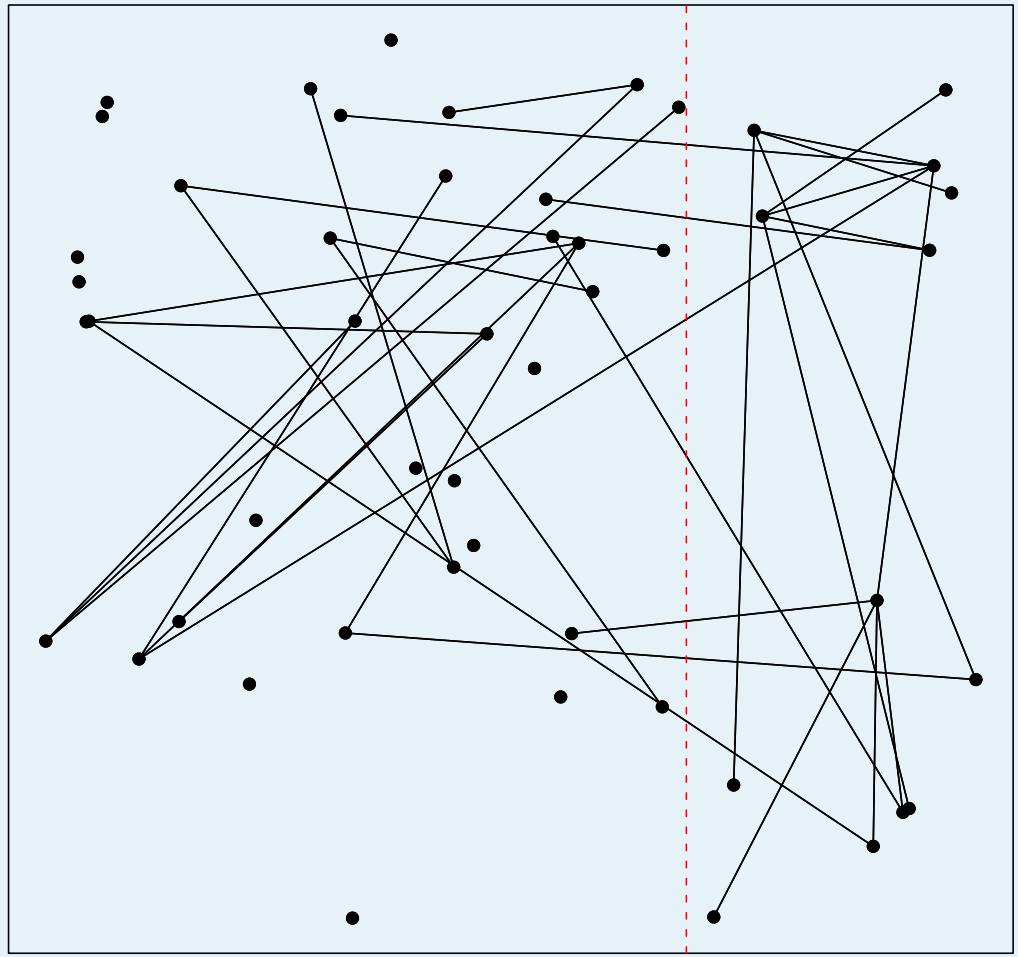
$\beta = 0.451$

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

H_{A_2}

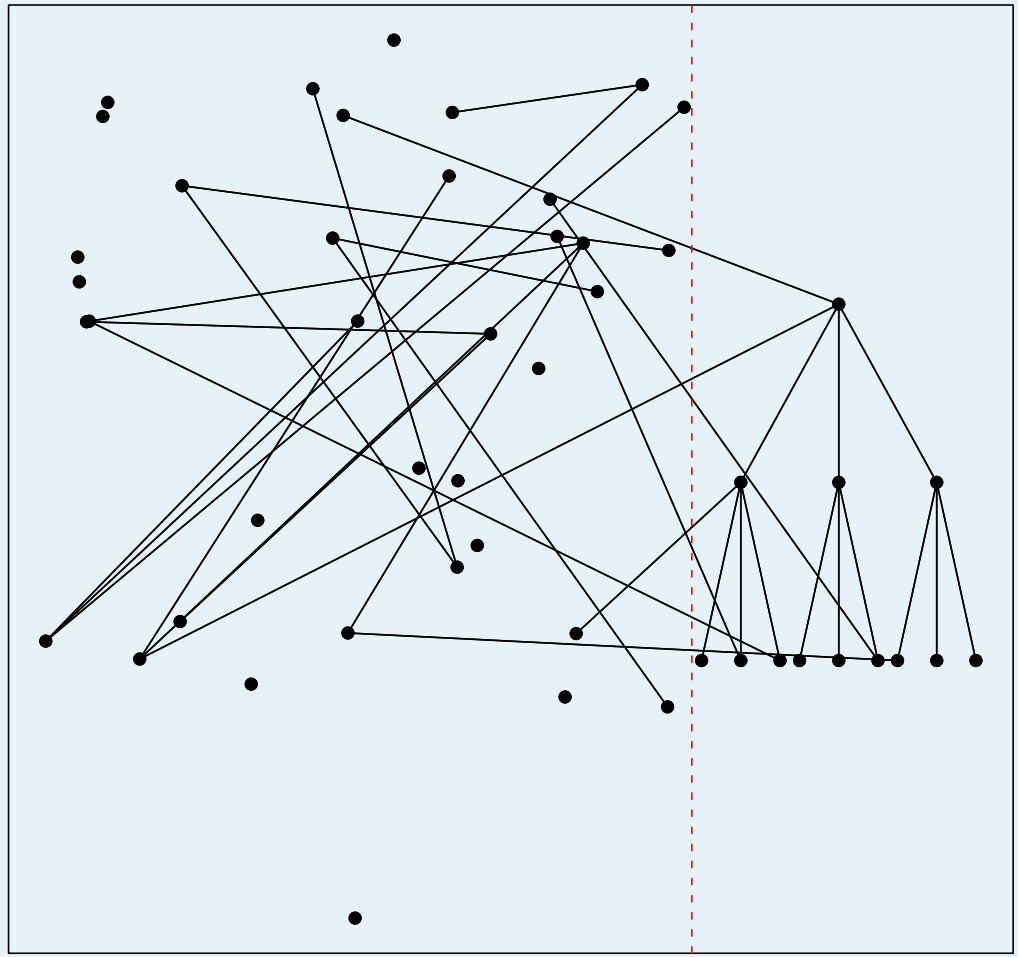
consider now a *structured* alternative . . .

Example: H_{A_2}



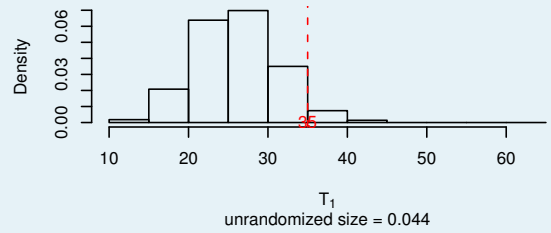
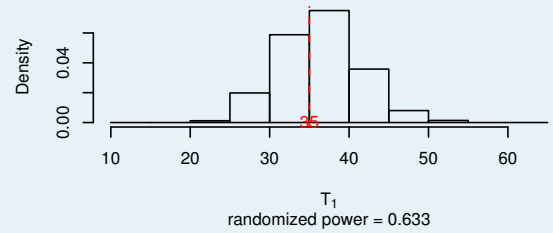
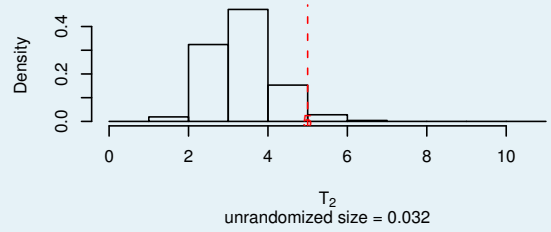
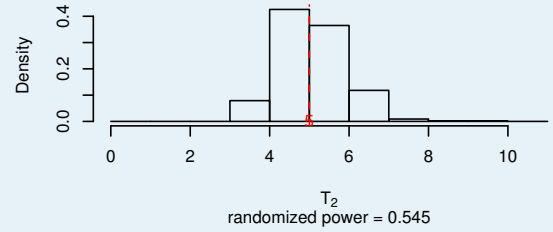
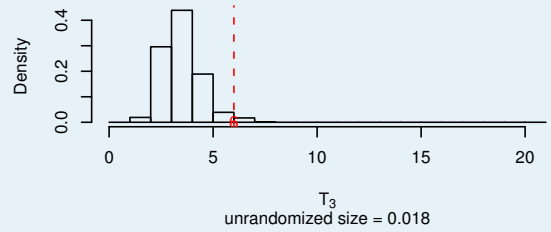
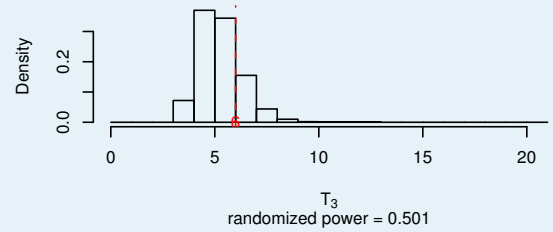
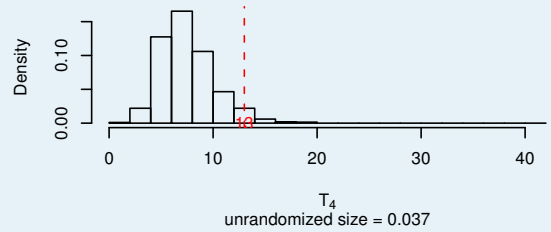
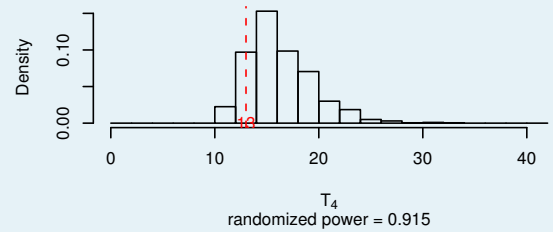
$$size = 34, scan0 = 5, scan1 = 5, scan2 = 17.$$

Example: H_{A_2}



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

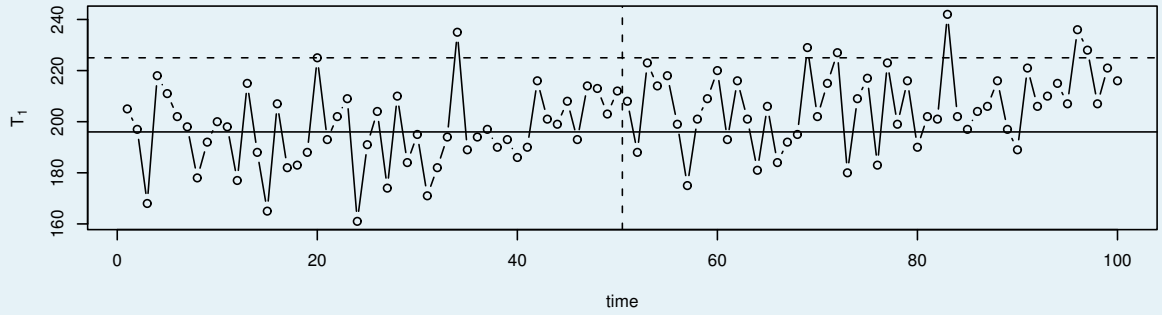
Example: Monte Carlo Simulation H_0 vs H_{A_2}

	H_0	H_A	
$T = size$			$\beta = 0.633$
$T = scan0$			$\beta = 0.545$
$T = scan1$			$\beta = 0.501$
$T = scan2$			$\beta = 0.915$

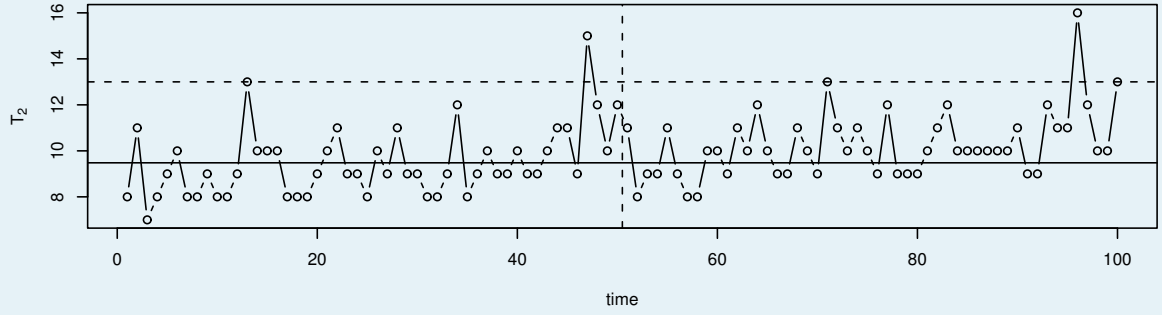
$$n = 50, n' = 13, \alpha = 0.05, MC = 1000$$

Time Series

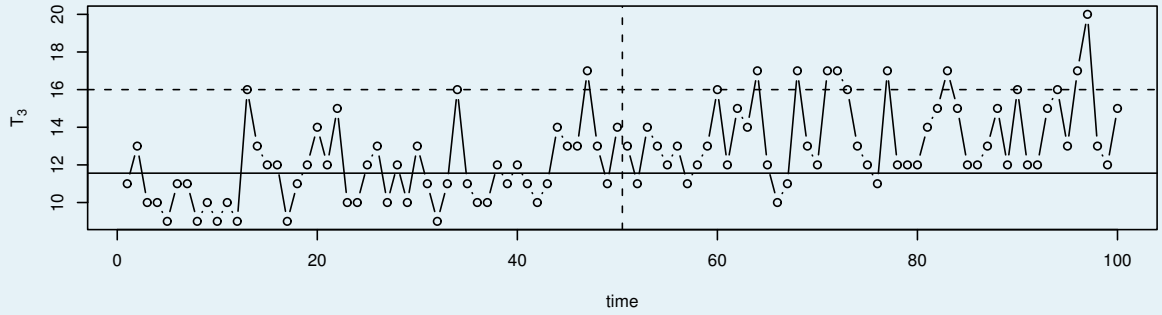
$T = size$



$T = scan0$



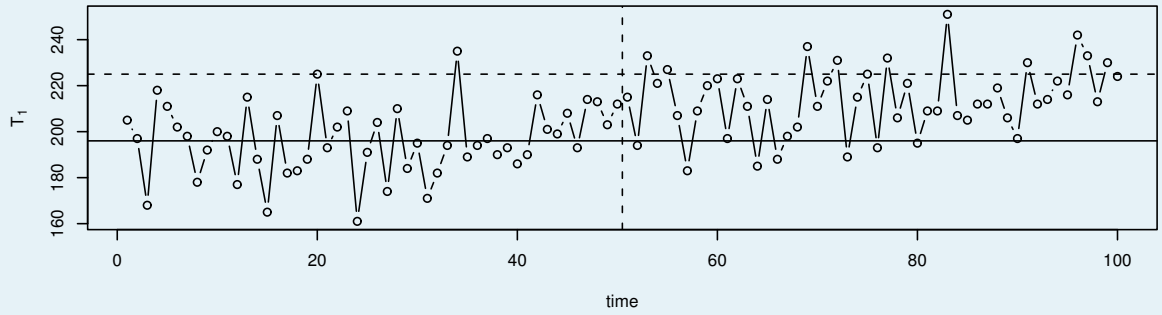
$T = scan1$



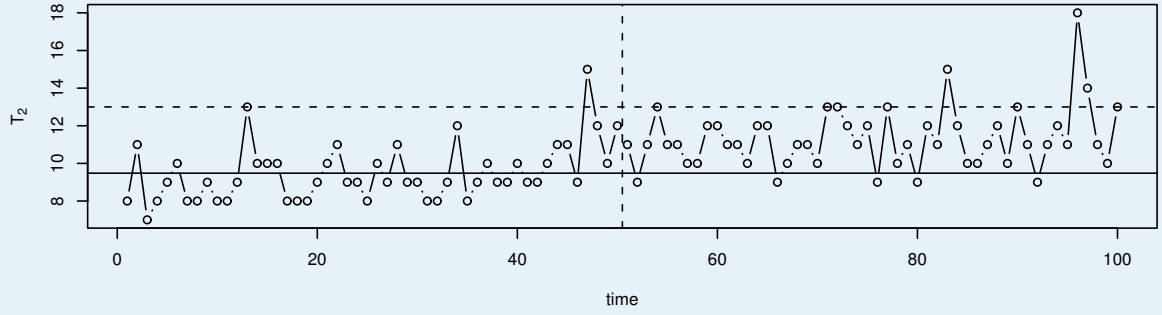
$$n = 100, n' = 4, \alpha = 0.05$$

Time Series

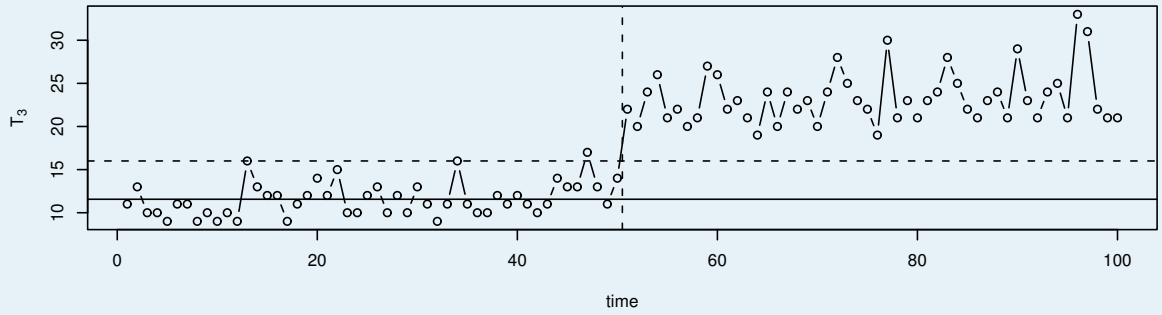
$T = size$



$T = scan0$



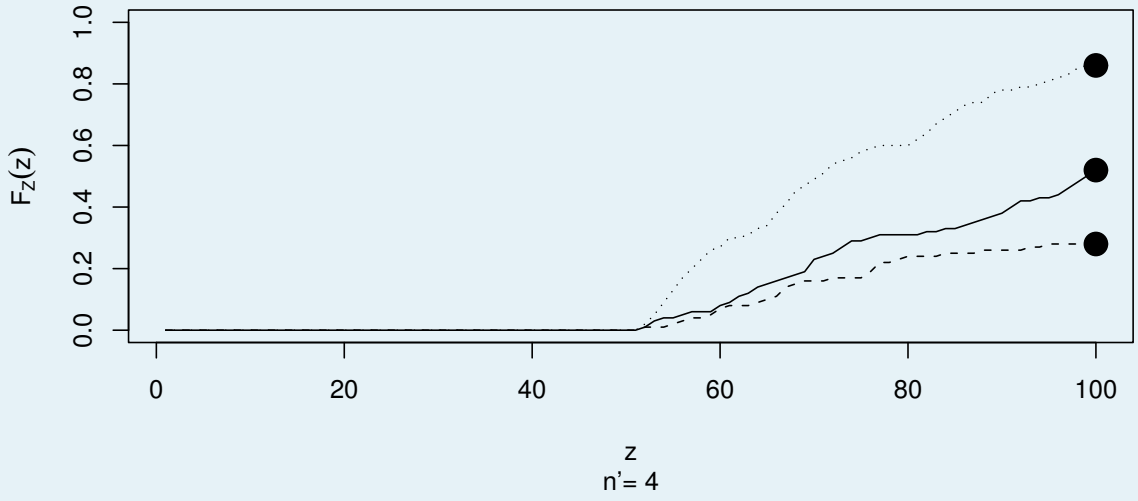
$T = scan1$



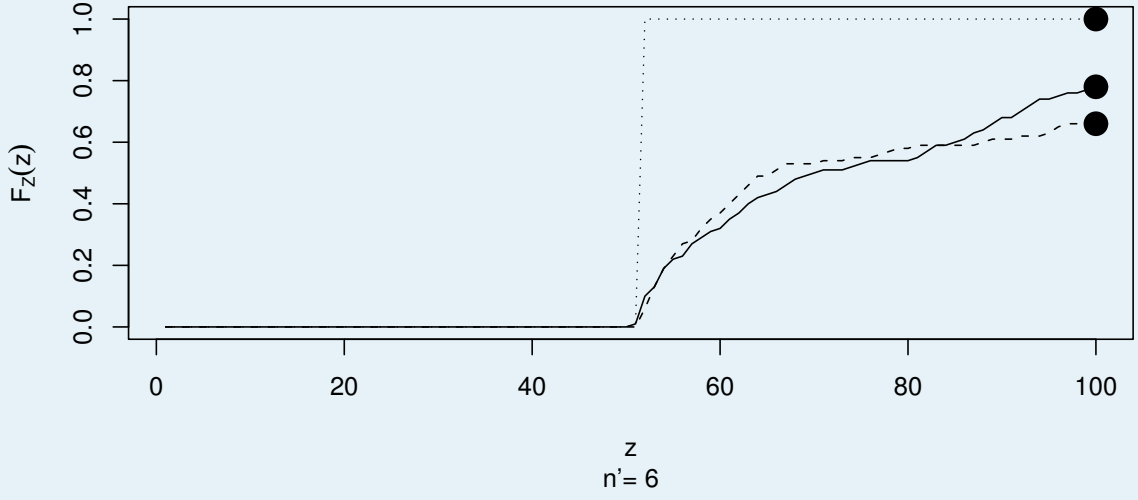
$$n = 100, n' = 6, \alpha = 0.05$$

Time Series

$n = 4$

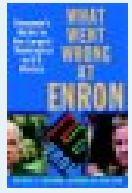


$n = 6$

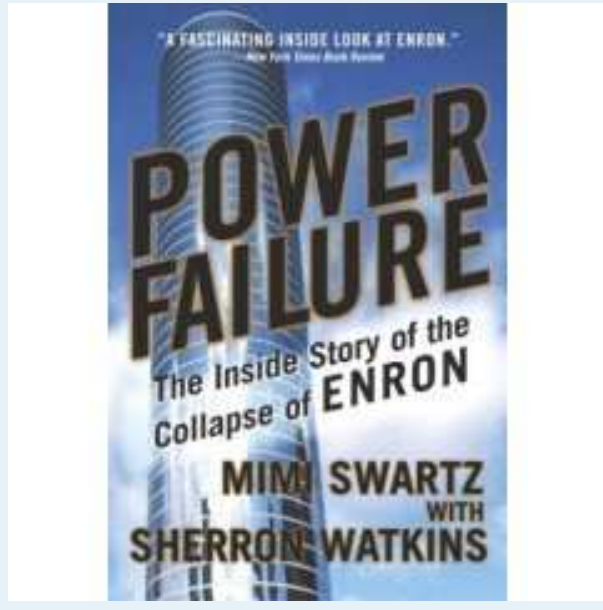


$$n = 100, n' = 4\&6, \alpha = 0.05$$

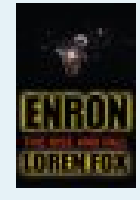
Enron


WHAT
WENT
WRONG
AT
ENRON


CONSPIRACY
OF FOOLS
KURT
EICHENWALD


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


Anatomy
of Greed


ENRON

Enron Data

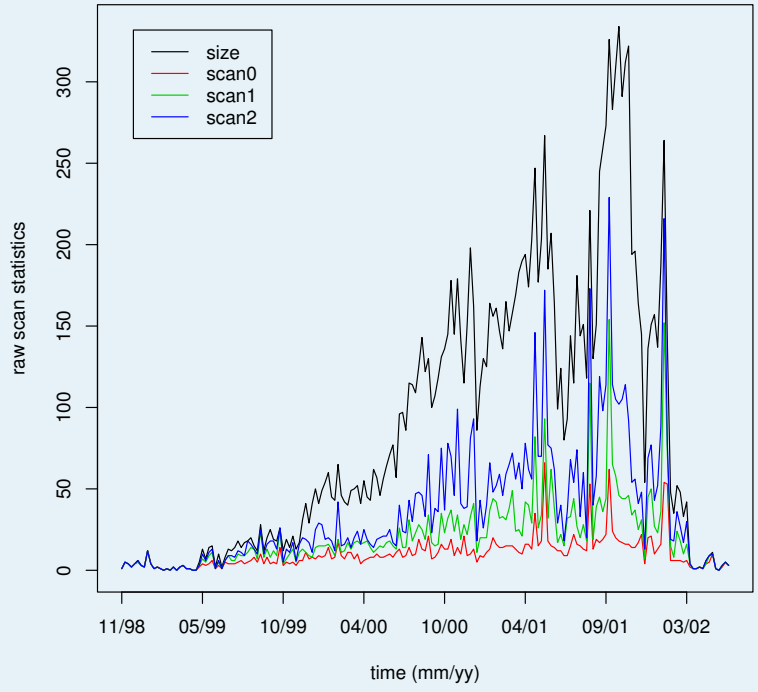
- 125,409 distinct messages from 184 unique “From” field, mostly Enron executives.
- 189 weeks, from 1988 through 2002.
- directed edges (arcs) $A_t = \{(v, w)$: vertex v sends at least one email to vertex w during the t-th week (“To”, “CC”, or “BCC”)$\}$

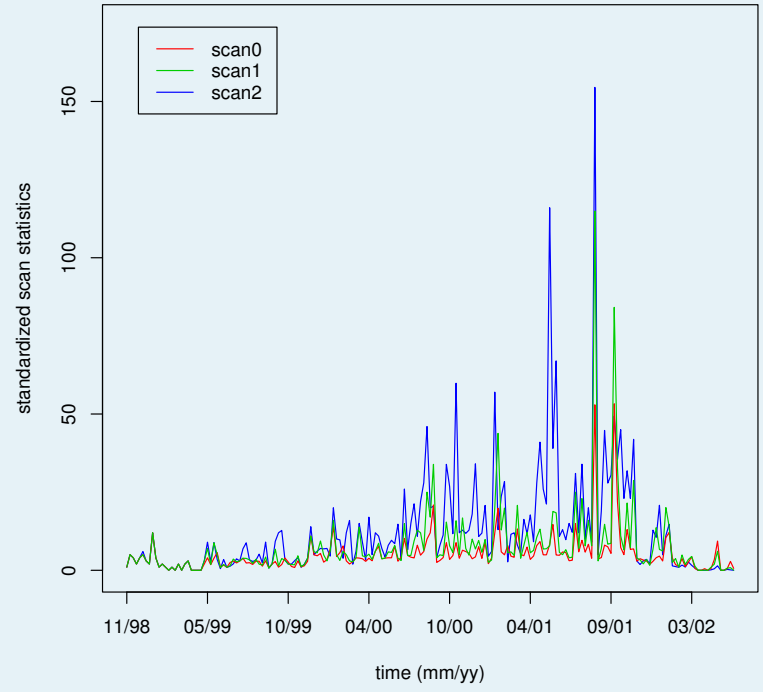
910948020 114 169	albert.meyers	Albert Meyers	Employee	Specialist
910948020 114 169	a..martin	Thomas Martin	Vice President	
911477940 114 123	andrea.ring	Andrea Ring	N/A	
911477940 114 123	andrew.lewis	Andrew Lewis	Director	
911481840 114 123	andy.zipper	Andy Zipper	Vice President	Enron Online
911481840 114 123	a..shankman	Jeffrey Shankman	President	Enron Global Mkts
911481840 114 123	barry.tycholiz	Barry Tycholiz	Vice President	
911481840 114 123	benjamin.rogers	Benjamin Rogers	Employee	Associate
911892180 114 38	bill.rapp	Bill Rapp	N/A	
911892180 114 38	bill.williams	xxx		

Statistics and Time Series

- scale-k locality statistics: $\Psi_{k,t}(v) = |A(\Omega(N_k[v; D_t]))|$
  - $k = 0$: $\Psi_{0,t}(v) =$ outdegree$(v; D_t)$.
  - scan statistic: $M_{k,t} = \max_v \Psi_{k,t}(v); k = 0, 1, 2$
- vertex-dependent standardized locality statistic:
  - $\widetilde{\Psi}_{k,t}(v) = (\Psi_{k,t}(v) - \widehat{\mu}_{k,t,\tau}(v)) \,/\, \max(\widehat{\sigma}_{k,t,\tau}(v), 1)$
  - $\widehat{\mu}_{k,t,\tau}(v) = \frac{1}{\tau} \sum_{t'=t-\tau}^{t-1} \Psi_{k,t'}(v)$
  - $\widehat{\sigma}^2_{k,t,\tau}(v) = \frac{1}{\tau-1} \sum_{t'=t-\tau}^{t-1} (\Psi_{k,t'}(v) - \widehat{\mu}_{k,t,\tau}(v))^2$
  - standardized scan statistic: $\widetilde{M}_{k,t} = \max_v \widetilde{\Psi}_{k,t}(v)$

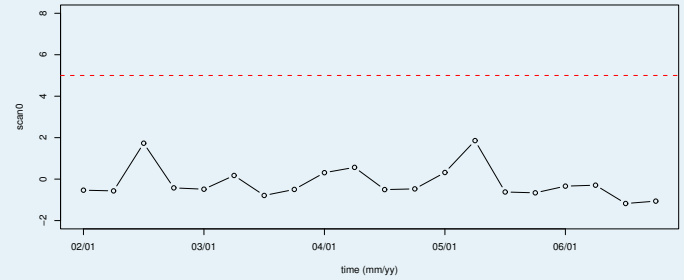
Statistics and Time Series

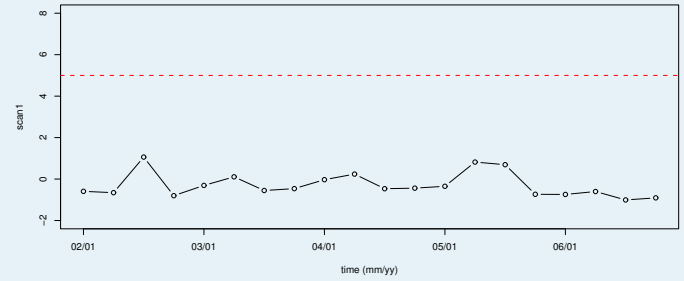


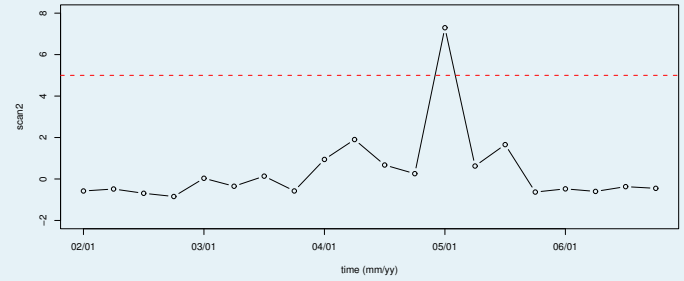


Anomaly Detection

- temporally-normalized scan statistics: $S_{k,t} = (\widetilde{M}_{k,t} - \widetilde{\mu}_{k,t,\ell})/\max(\widetilde{\sigma}_{k,t,\ell}, 1)$
- detection: time t such that $S_{k,t} > 5$
- $t^* = 132$ (May, 2001)







Detection Graph D_{132}

$$\arg\max_v \Psi_{0,132}(v) = john.lavorato$$

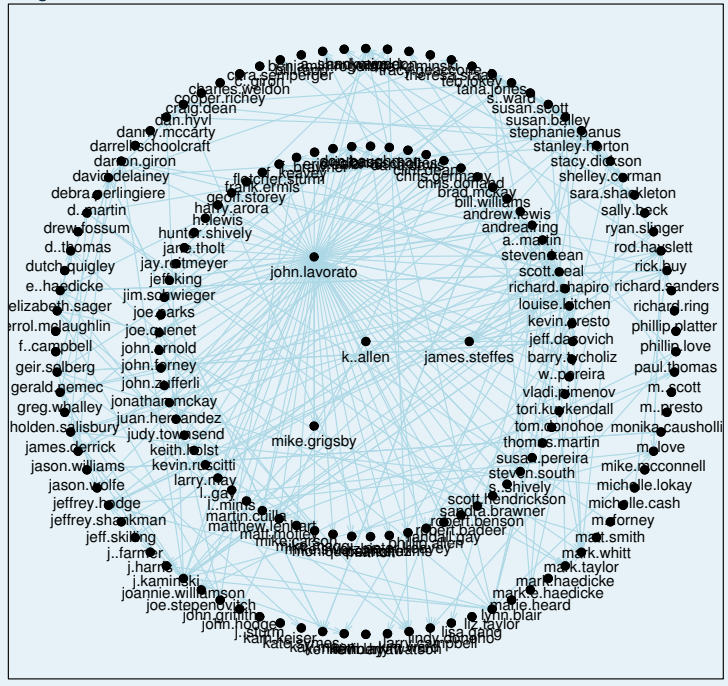
$$\arg\max_v \Psi_{1,132}(v) = john.lavorato$$

$$\arg\max_v \Psi_{2,132}(v) = richard.shapiro$$

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

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

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



< > − +

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		

Anomaly Detection (*Aliasing*)

- $v^* = \arg\max_v \widetilde{\Psi}_{2,132}(v) = k..allen$
- *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

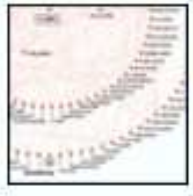
By GINA KOLATA
Published: May 22, 2006

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

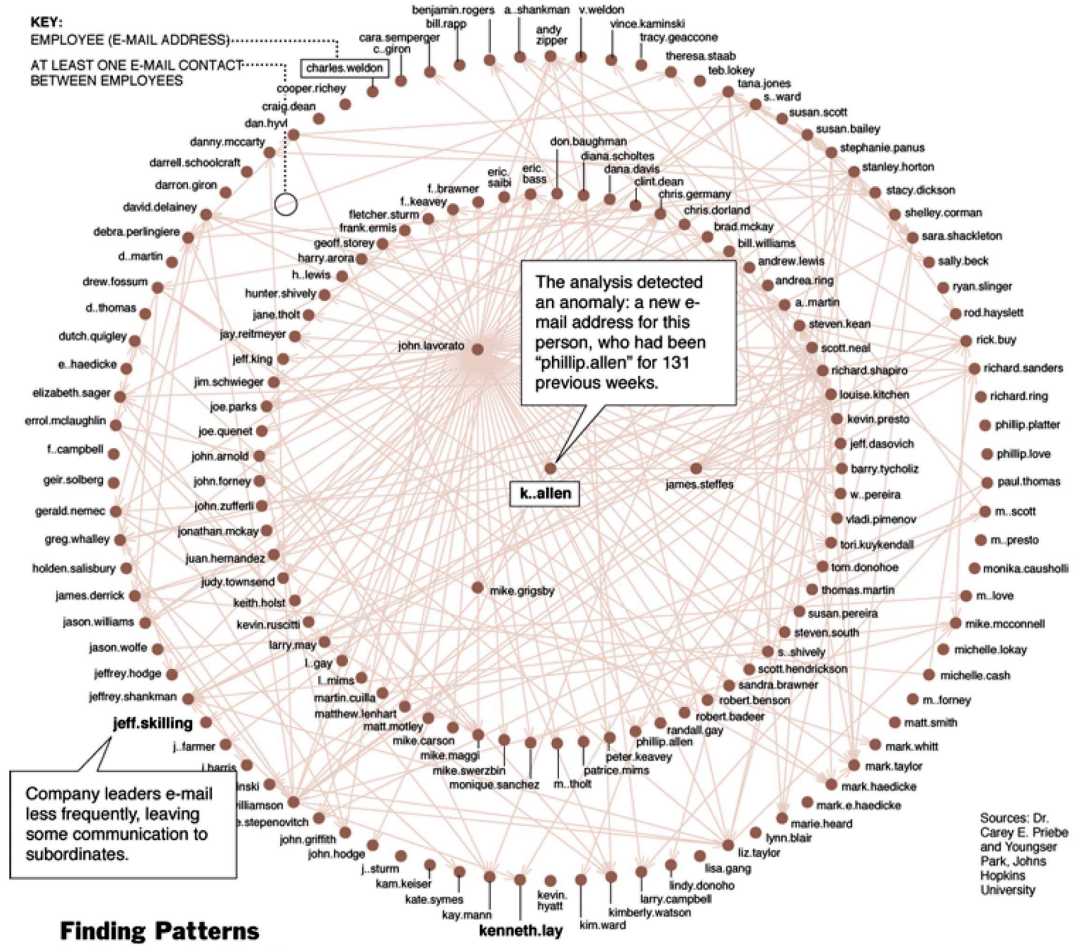
Now the science of e-mail tracking and analysis has been given a unlikely boost by a bitter chapter in the history of corporate malfeasance - the Enron scandal.

In 2003, the Federal Energy

< > − +

New York Times (May 22, 2005)

The New York Times May 21, 2005



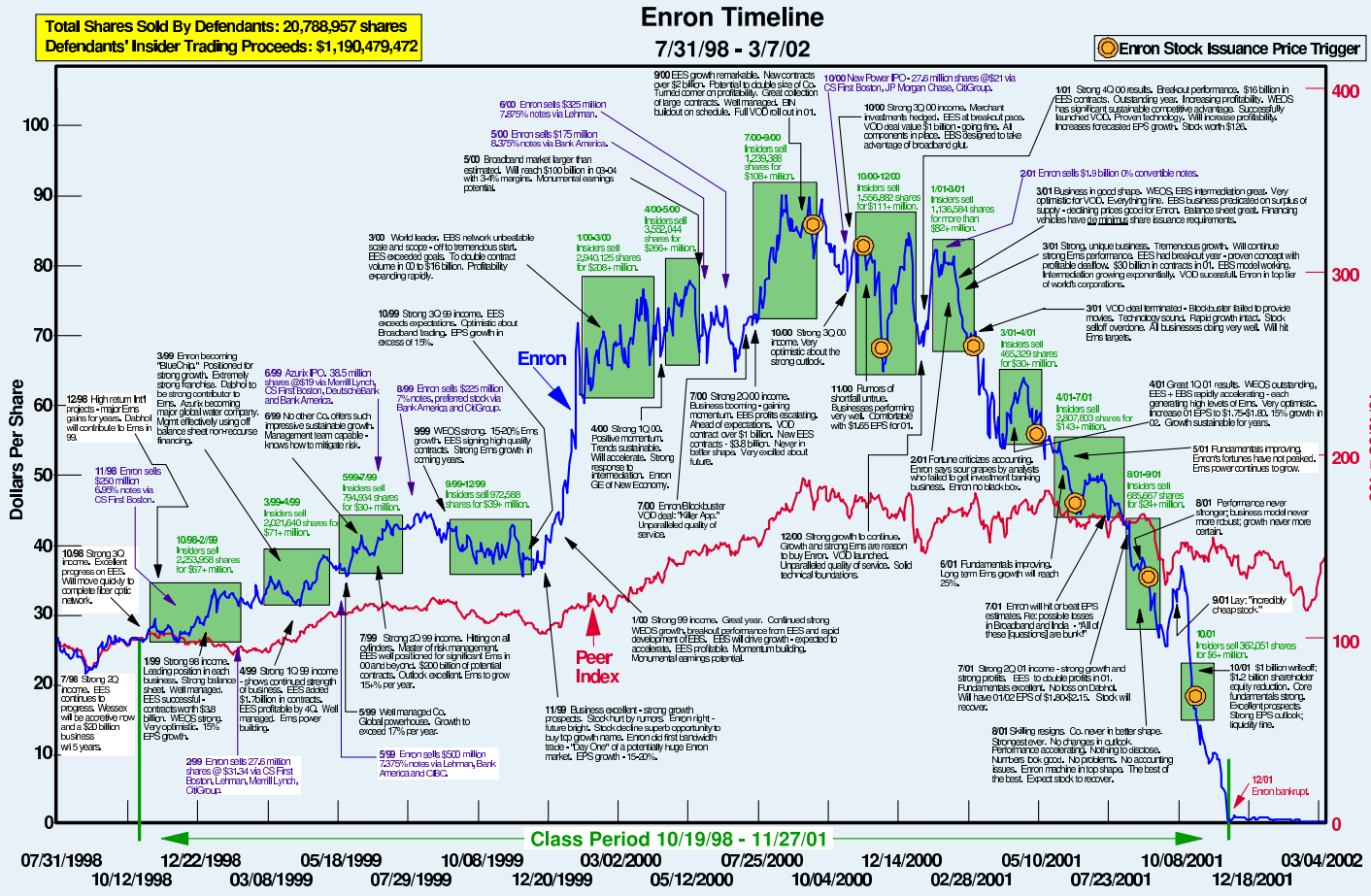
Finding Patterns In Corporate Chatter

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]

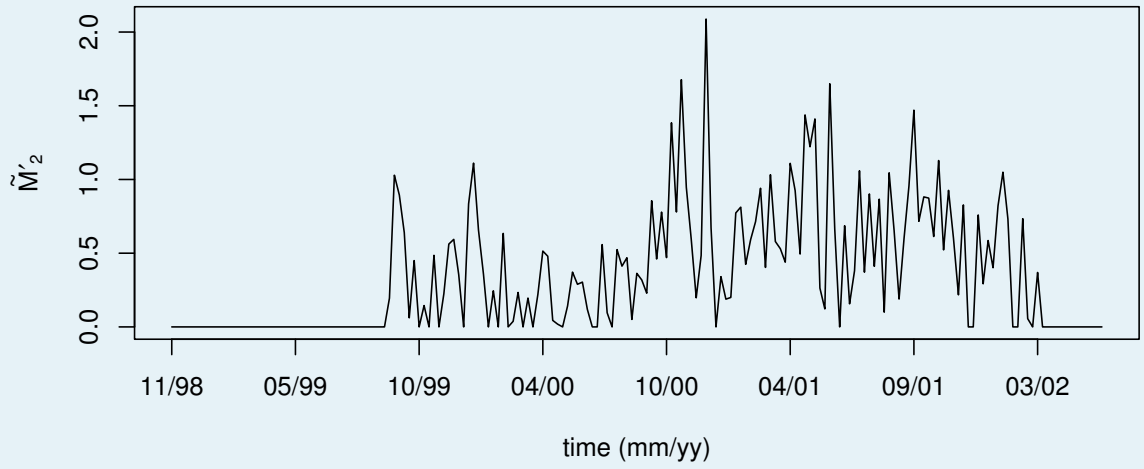
Anomaly Detection (*chatter*)

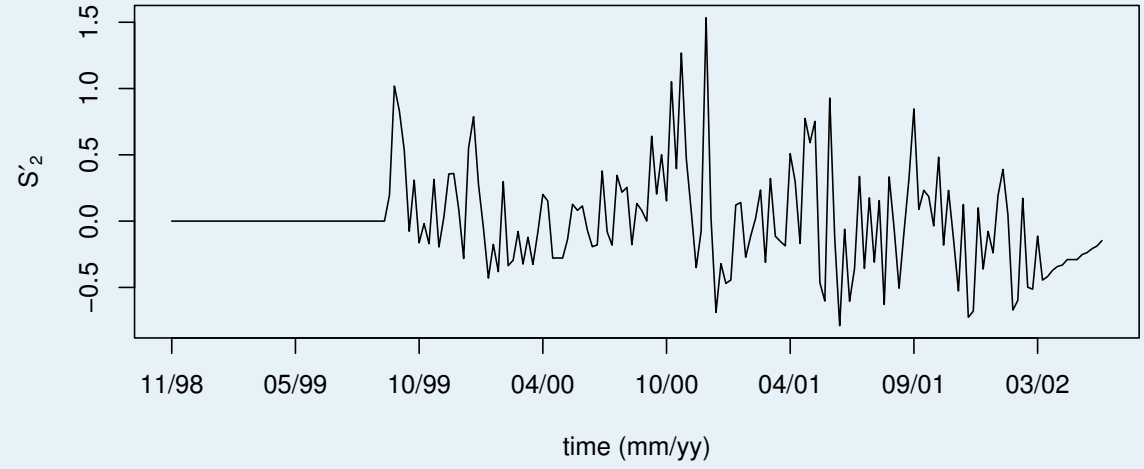
- Seek a detection in which the excess activity is due to chatter amongst the 2-neighbors!

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

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

Anomaly Detection (*chatter*)



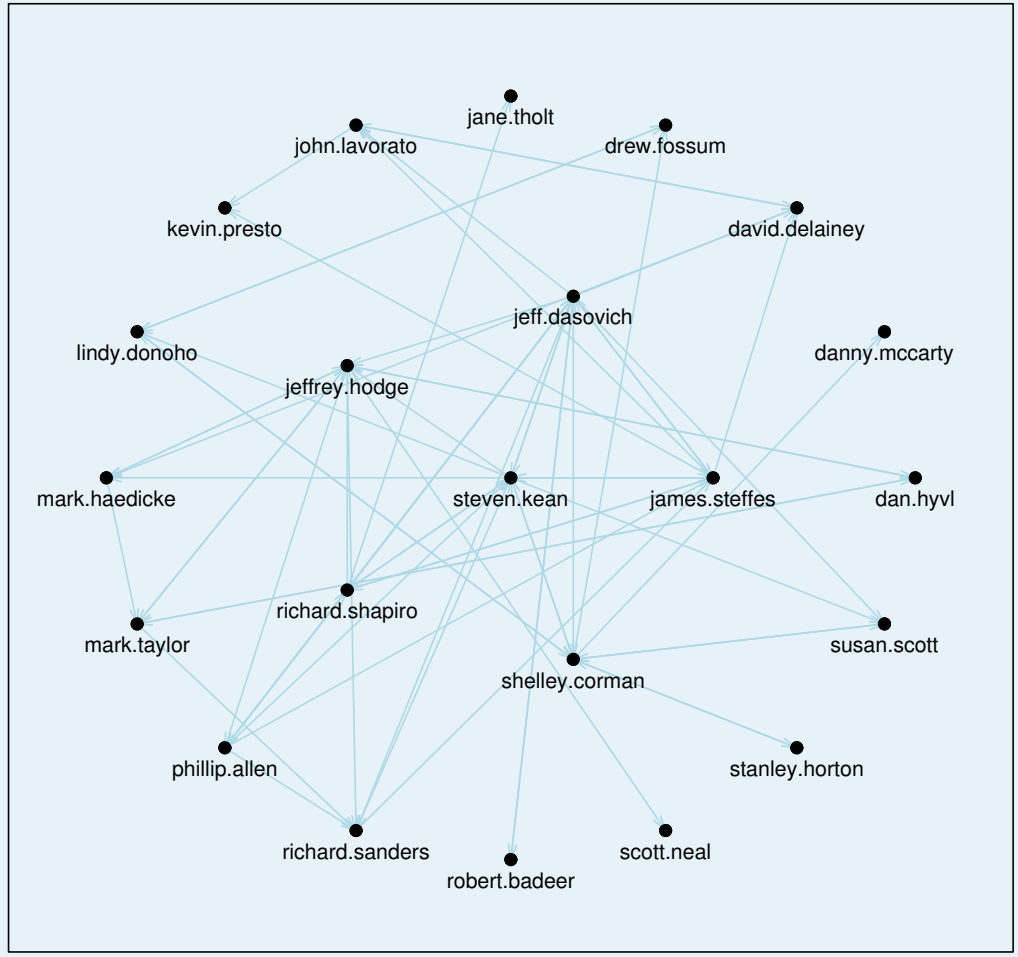


Anomaly Detection (*chatter*)

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

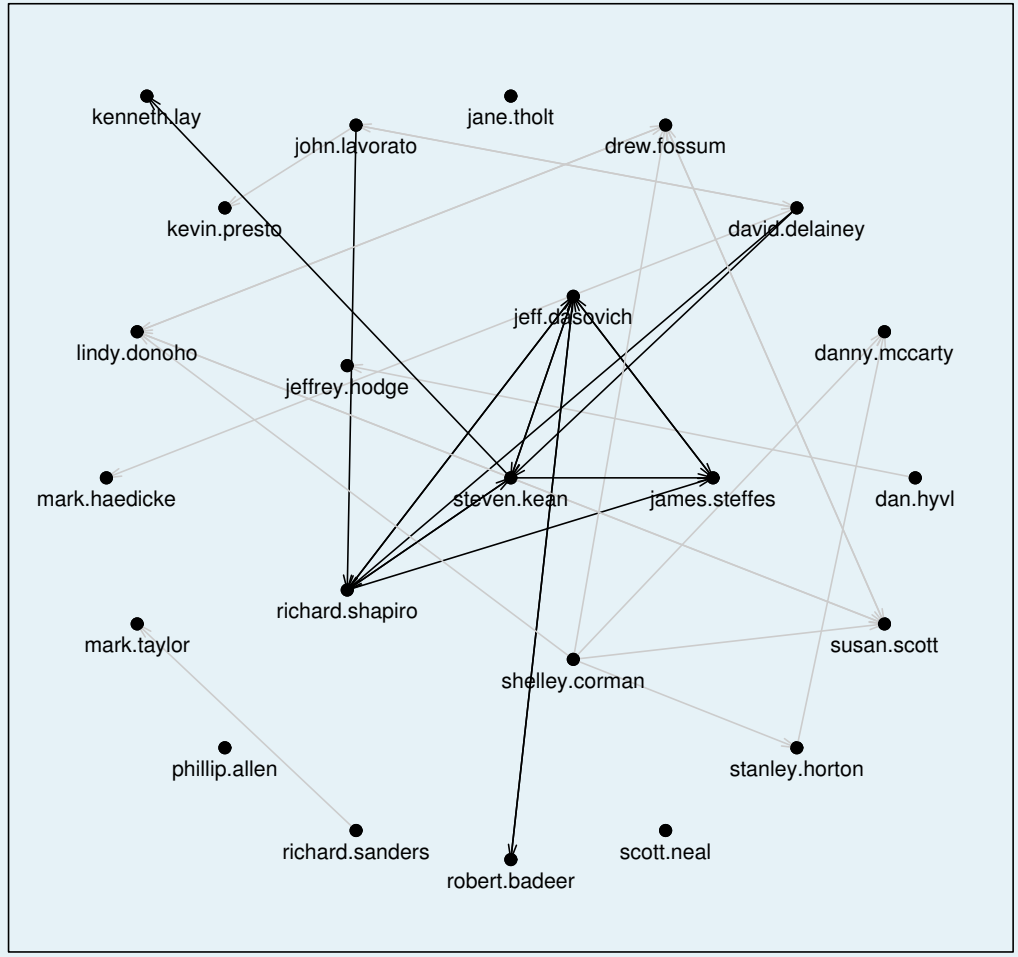
scale k	$\Psi_{k,t^*-5:t^*}(v^*)$
0	[3 , 5 , 4 , 5 , 4 , 5]
1	[11 , 13 , 10 , 10 , 11 , 18]
2	[14 , 35 , 21 , 38 , 13 , 65]

Anomaly Detection (*chatter*)



Ω_{109}

Anomaly Detection (*chatter*)



Ω_{108}

Text Data for Week 109 Detection

- 1092 transactions among 22 users
- 343 files
- 91 unique messages
- Counts and Subject Lines:

  9
  5 Analysis of Joskow / Hogan Papers
  4 FERC Request
  3 Data on Monthly Generation for SCE
  3 Draft Talking points about California Gas market
  3 EnronOnline question
  3 Presentations from GA Meeting on December 8
  2 California Price Issues
  2 Conectiv / Delmarva
  2 Davis, Hoecker and Richardson
  2 FYI-Edison wants Reregulation

Clustering Based on Content

- Find emails with similar content based on *terms* that occurs.
- Terms: space-delimited string of characters from $\{a, b, c, \cdots, z\}$, after text is lower cased and all other characters and stop words are removed.
- Need to restrict our attention to *(signature terms)*.
  - terms that occur more often then expected.
  - based on mutual information.
  - Dunning 1993, Hovy & Lin 2000.

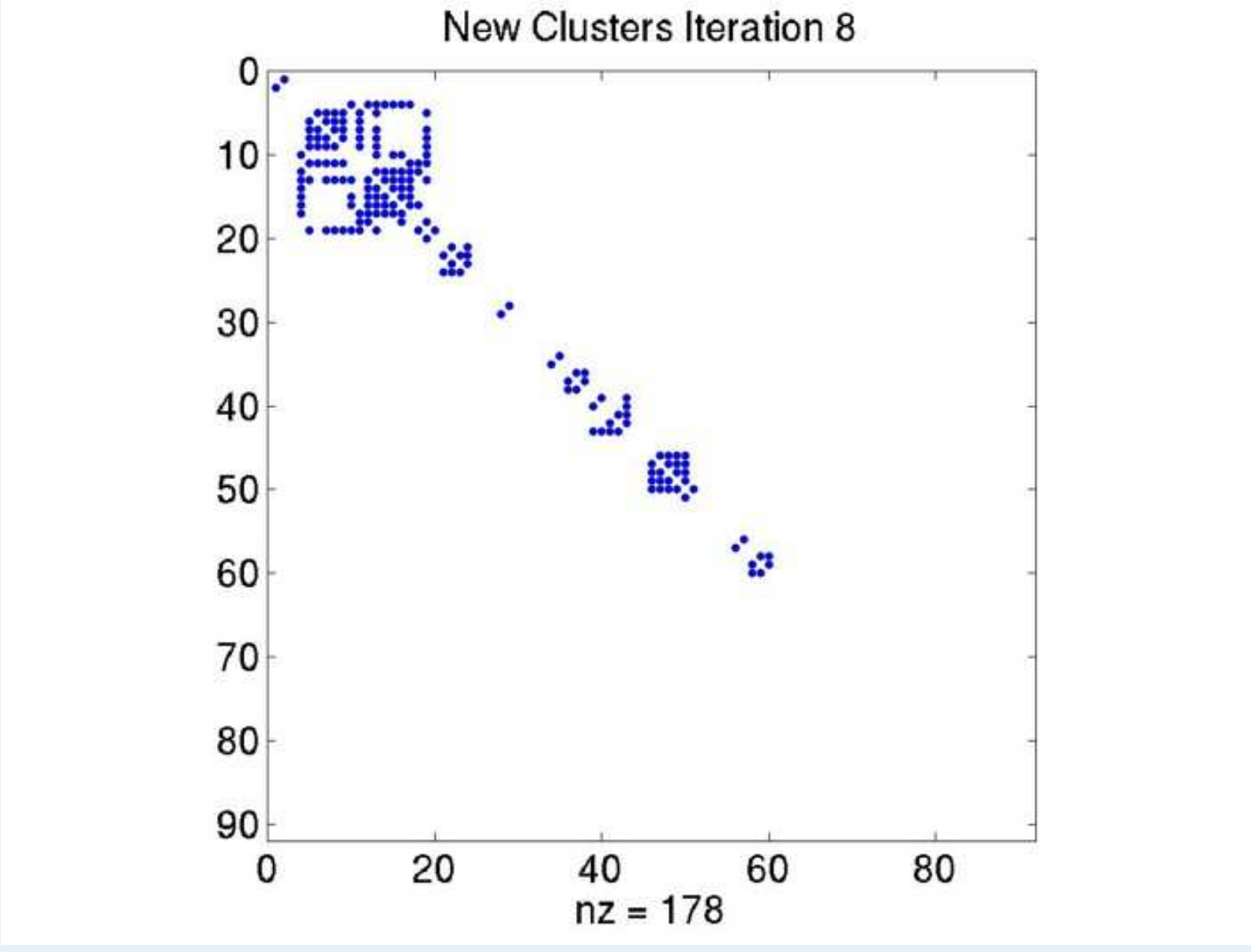
Example Message

Subject: Re: Analysis of Joskow / Hogan Papers
Sounds very good.
Might be useful to get a "reputable" economist to write a paper that 1) describes traditional means for defining, identifying and mitigating market power, 2) compares those with the "new" means folks are coming up with these days, and 3) comments on the "split" in the academic community over the issues.
When Steve Kean and I discussed the notion initially, thought it might be a good idea to gently "pile on" to the public discussion with the goal of making clear 1) just how complex this issue is and 2) how important it will be to have a thorough analysis (say, about 12+ months worth?) before rushing to judgment on anything Joskow might allege in his paper. Thoughts?
Best, Jeff
Signature Terms: analysis, california, com, economists, enron, hogan, joskow, kahn, market, na, paper, power

Simple Clustering

- Preprocessing: remove any line with 2 or more @'s
- Compute signature terms for each message.
- Form an $n \times d$ matrix F, where
  $F(i, j) =$ number of sigterm i occurs in doc j.
- Calculate $R =$ corrcoef(F) and $P = d \times d$ matrix of p-values for R.
- Form a graph $G(P < \tau)$, that is, two documents are connected if there is a significant overlap in their signature terms!

Sample Clusters



Your mother is near?

Subject: Organizational Changes
———————- Forwarded by Richard Shapiro/NA/Enron on 12/08/2000=
Subject: Re: Analysis of Joskow / Hogan Papers
Having read the Hogan paper, I think that the "academic" community is ...
paper by three prominent economists done for San Diego Gas and Electric. The ...
paper by John D. Chandley, Scott M. Harvey, and William W. Hogan argues that
Subject: Hogan-California Market Power
FYI. Not sure if you had seen this. Hogan makes many of the arguments about
Subject: Re: Draft Talking points about California Gas market
Given the way the numbers came out, I guess we don't need the talking points,
Subject: Re: FERC Request
Drew is okay with this. I will email the list to FERC.
Subject: Update on FERC California Gas/Electric Matters
into the California market last summer....
Various Enron units continue to receive informal data requests from FERC ...
staff regarding current California gas/electric

Related News Item

January 13, 2001 Leading economists Paul Jaskow and Edward Kahn conclude that high wholesale prices observed in summer 2000 [in California] cannot be explained as the natural outcome of ‘market fundamentals in competitive markets since there is a very significant gap between actual market prices and competitive benchmark prices. (Source: CATO Policy Analysis)

http://cantwell.senate.gov/news/releases/2002_04_18_consumer.html

Discussion

- scan statistics offers promise for
  detecting anomalies in time series of graphs.
- extensions:
  - weighted graphs (# of messages)
  - coloured hypergraphs (“To”, “CC”, or “BCC”)
  - sliding window (online analysis)
  - exponential smoothing, detrending, variance stabilization
  - ...
- “Content and Scan Statistics for Enron” — John Conroy, et al.
- “Random Dot Product Graphs” — Ed Scheinerman, et al.

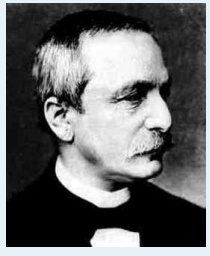
Kronecker Quote

"The wealth of your practical experience

with sane and interesting problems

will give to mathematics

a new direction and a new impetus."



– Leopold Kronecker to Hermann von Helmholtz –