Scan Statistics on Enron Graphs

Kendall Giles

Department of Computer Science
Johns Hopkins University
and
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

UCLA IPAM, October 10, 2007

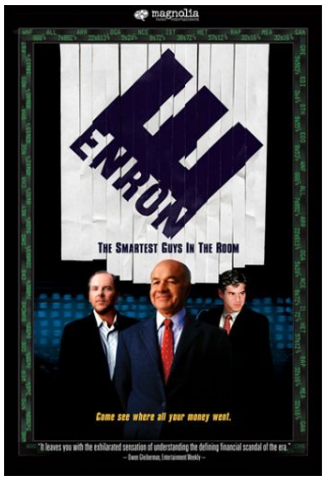
Outline

Outline

1. C.E. Priebe, J.M. Conroy, D.J. Marchette, and Y. Park, "Scan Statistics on Enron Graphs," Computational and Mathematical Organization Theory, Volume 11, Number 3, p229 - 247, October 2005, Springer Science+Business Media B.V.
2. 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.
3. Gina Kolata, "Enron Offers an Unlikely Boost to E-Mail Surveillance," New York Times„ Week in Review, May 22, 2005.
4. C.E. Priebe, "Scan Statistics on Enron Graphs," IPAM Summer Graduate School: Intelligent Extraction of Information from Graphs and High Dimensional Data„ UCLA, July 11-29, 2005.
5. C.E. Priebe, "Scan Statistics on Enron Graphs," 2005 Fall Department of Applied Mathematics and Statistics Seminars„ September 15, 2005, The Johns Hopkins University.

Outline

Enron: The Smartest Guys in the Room



- Enron was one of world's leading electricity, natural gas, pulp and paper, and communications companies
- Seventh largest U.S. company
- Revenue $\approx$ \$111 billion in 2000
- Fortune: "America's Most Innovative Company" for 6 consecutive years

Kenneth Lay, Enron chairman and CEO: "We are proud to receive this accolade for a sixth year. It reflects our corporate culture which is driven by smart employees who continually come up with new ways to grow our business."

Enron: The Collapse

- Discoveries of highly irregular (fraud) accounting procedures done throughout the 1990's by Enron and Arthur Andersen
- In November, 2001 the stock price dropped from \$90 to $\approx$ 30 cents
- Executives unloaded stock worth millions of dollars while encouraging others to buy
- In December, 2001 Enron filed for Bankruptcy

Enron: The Problem

- An Enron email dataset was made public by the U.S. Department of Justice
- Might detections of excessive activity in sent email indicate interesting events?

A Goal:
Develop and apply a theory of scan statistics on random graphs to perform change point/anomaly detection in graphs and in time series of graphs

Scan Statistics

- Used to investigate some random field X for possible presence of a local signal
- "moving window analysis"
  - scan a small "window" over the data
  - calculate some locality statistic for each window
    - average pixel value for an image
    - number of events for point pattern
- scan statistic $M(X) \equiv$ max of these local statistics

Hypotheses

- H_0: homogeneity
- H_A: local subregion with excess activity

Inference: $P_{H_0}[M(X) \geq c_\alpha] = \alpha$

If the maximum of the local statistics is large enough, then can infer that there exists a *local region of excessive activity*

Scan Statistics on Graphs

Directed graph: $D = (V, A)$

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

Variable-Scale Scan Statistic

- Let $K \subset \{1, \ldots, n-1\}$ be a collection of scales
- Let Ψ'_k denote be a scale-standardized version of Ψ_k
- Want $g_{k,\alpha}(\cdot)$ s.t. $\Psi'_k = g_{k,\alpha}(\Psi_k(v))$ satisfies $P[\Psi'_k \geq c_\alpha] \approx \alpha \; \forall \; v \in V(D), k \in K$

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

Reject for large values of $M_K(D)$

Outline

Enron Email Dataset

- Email to and from senior management, energy traders, executive assistants, etc.
- From about 1998 - 2002: 189 weeks
- 184 users
- 125,409 distinct messages
- minimal pre-processing done to correct integrity issues
- attachments and content were not used

Outline

Email Time-Series Digraphs

- For each week $t = 1, \ldots, 189$: $D_t = (V, A_t)$
- $|V| = 184$
- A_t: $(v, w) \in A_t \iff$
  v sends w at least one email during the t^{th} week
- assume short-time stationarity under the null

Time-dependent scale-k locality statistic:

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

$\Psi_{0,t}(v) \equiv \text{outdegree}(v; D_t)$

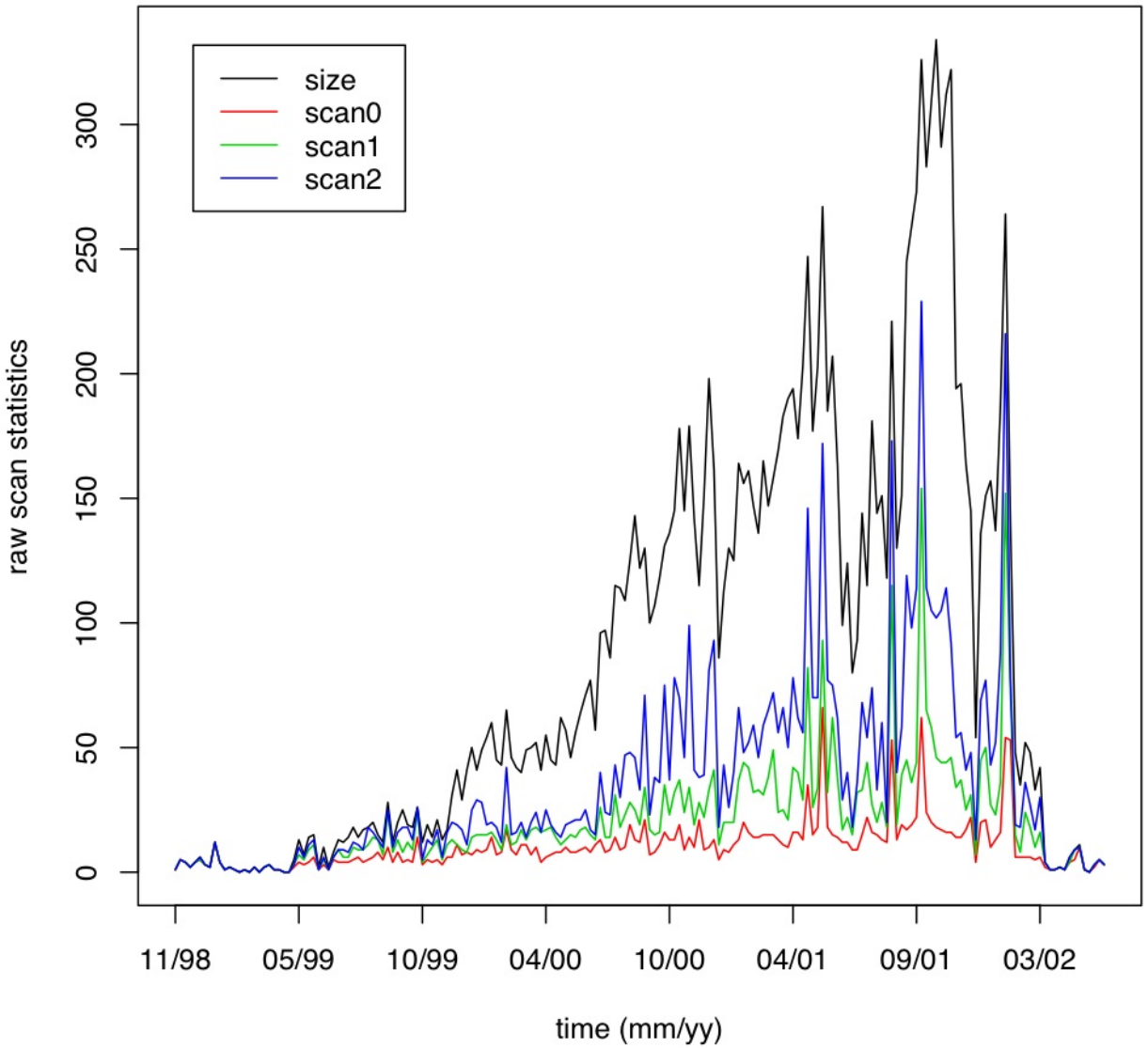
$M_{k,t} = \max_v \Psi_{k,t}(v)$

Vertex-Dependent Standardized Locality and Scan Statistics

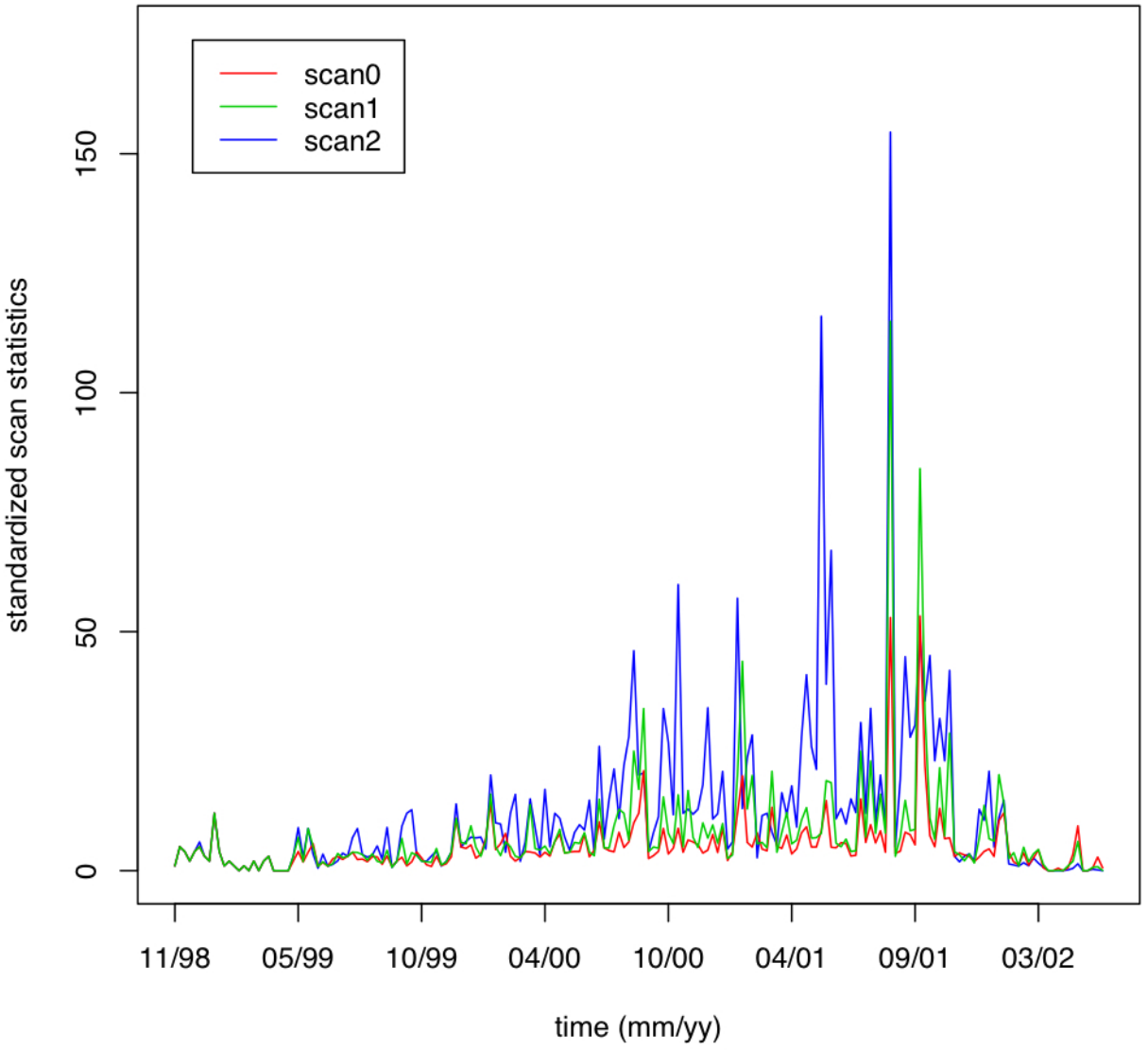
- $\widetilde{\Psi}_{k,t}(v) = (\Psi_{k,t}(v) - \widehat{\mu}_{k,t,\tau}(v))/\max(\widehat{\sigma}_{k,t,\tau}(v), 1)$
- $\widehat{\mu}_{k,t,\tau}(v)) = \frac{1}{\tau}\sum_{t'=t-\tau}^{t-1} \Psi_{k,t'}(v)$
- $\widehat{\sigma}^2_{k,t,\tau}(v) = \frac{1}{\tau-1}\sum_{t'=t-\tau}^{t-1}(\Psi_{k,\tau'}(v) -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

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

Raw scan statistics for $k = 0, 1, 2$



Standardized scan statistics for $k = 0, 1, 2$



Outline

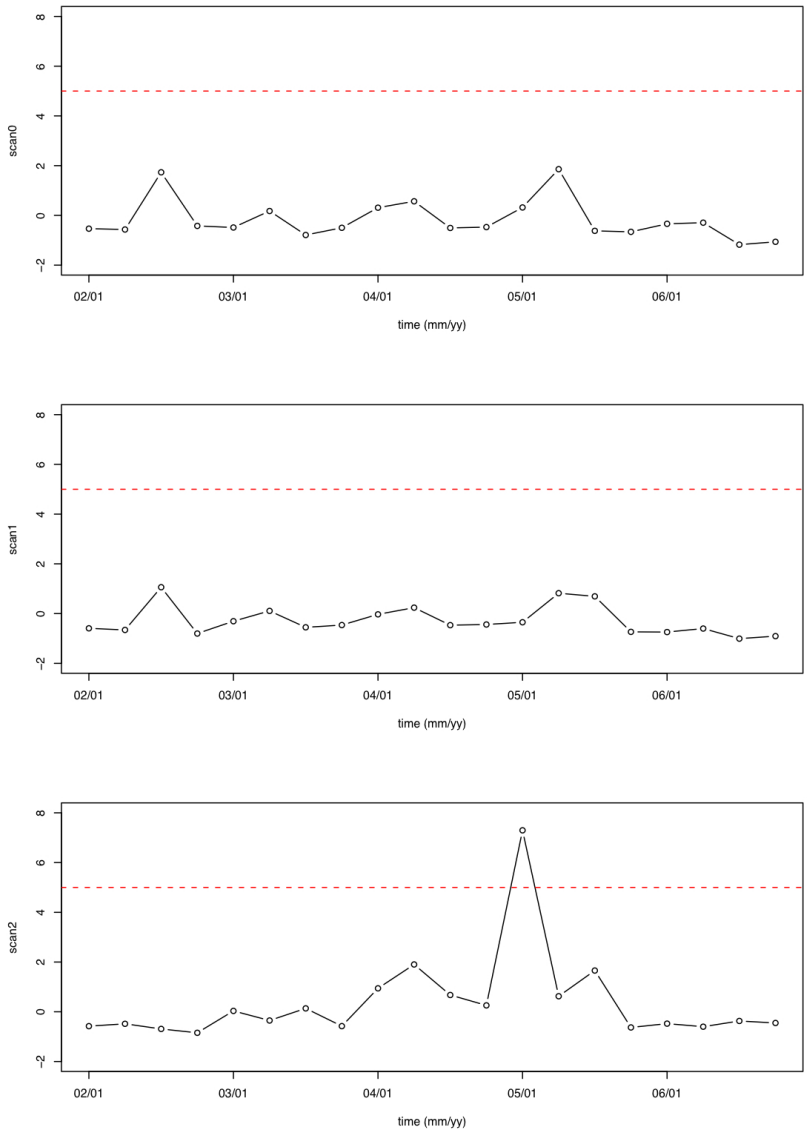
Anomaly Detection

temporally-normalized scan statistic:

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

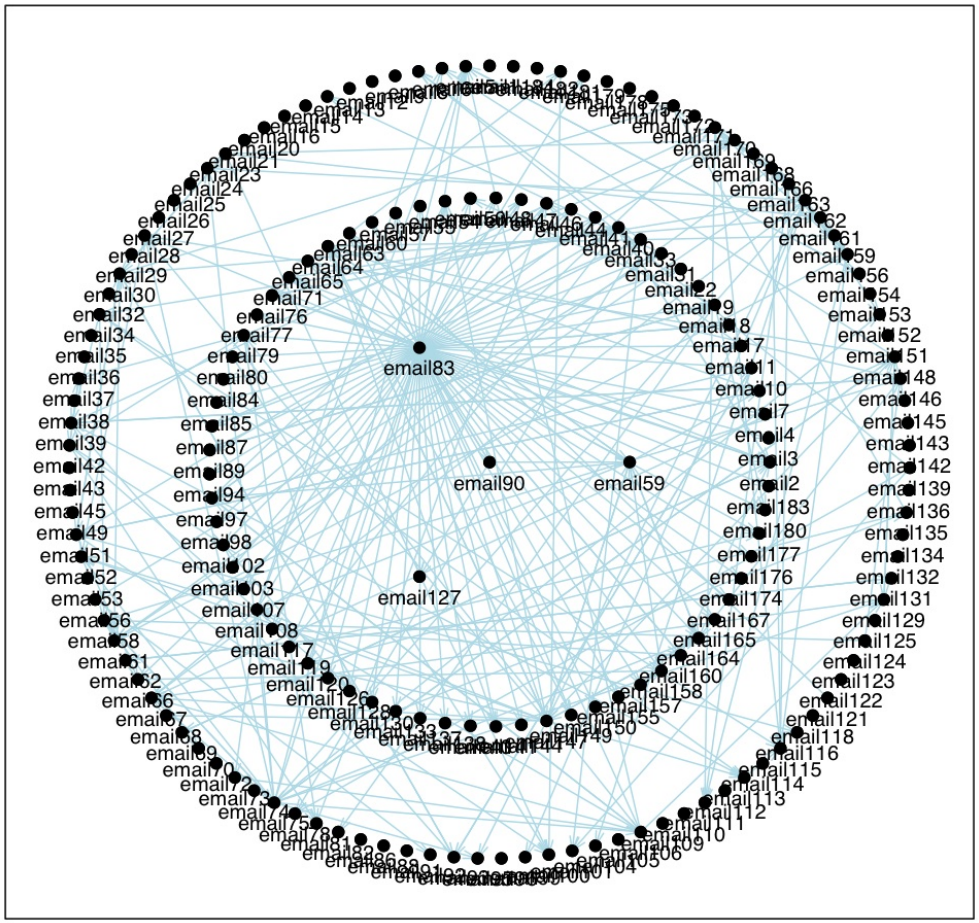
detection: time t such that $S_{k,t} > 5$

Anomaly Detection



$t^* = 132$ (May, 2001)

Detection Graph D_{132}



$$\arg\max_v \Psi_{0,132}(v) = email83$$

$$\arg\max_v \Psi_{1,132}(v) = email83$$

$$\arg\max_v \Psi_{2,132}(v) = email147$$

$$\arg\max_v \widetilde{\Psi}_{0,132}(v) = email147$$

$$\arg\max_v \widetilde{\Psi}_{1,132}(v) = email75$$

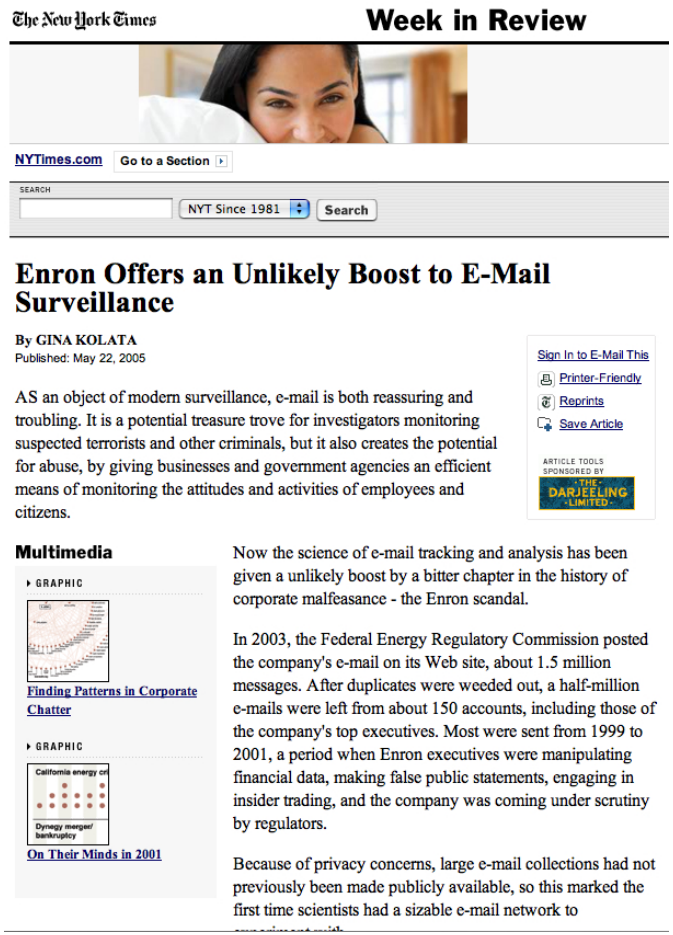
$$\arg\max_v \widetilde{\Psi}_{2,132}(v) = email90$$

Detection Graph Details

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)

- $v^* = \arg \max_v \widetilde{\Psi}_{2,132}(v) = email90$
- k..allen == phillip.allen?
  - k..allen had no activity before $t^* = 132$
  - at $t^* = 132$, phillip.allen switched to k..allen



The New York Times **Week in Review**

NYTimes.com Go to a Section

SEARCH NYT Since 1981 Search

Enron Offers an Unlikely Boost to E-Mail Surveillance

By GINA KOLATA
Published: May 22, 2005

Sign In to E-Mail This
Printer-Friendly
Reprints
Save Article

ARTICLE TOOLS SPONSORED BY THE DARJEELING LIMITED

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

GRAPHIC

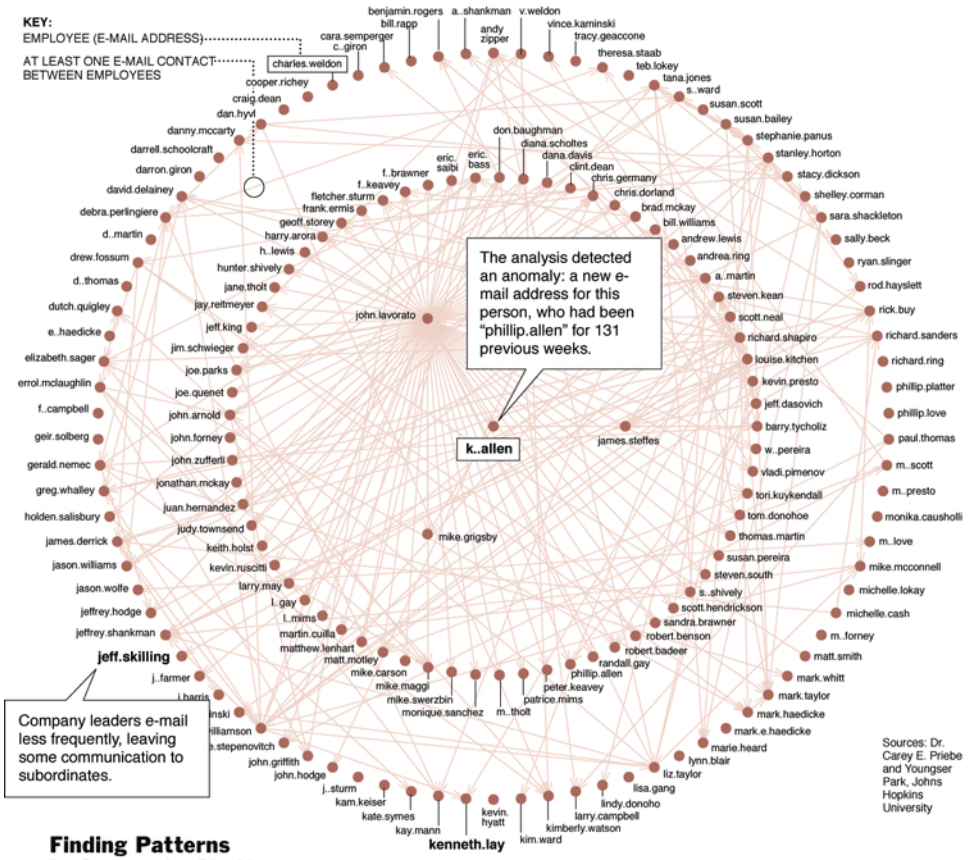
On Their Minds in 2001

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.

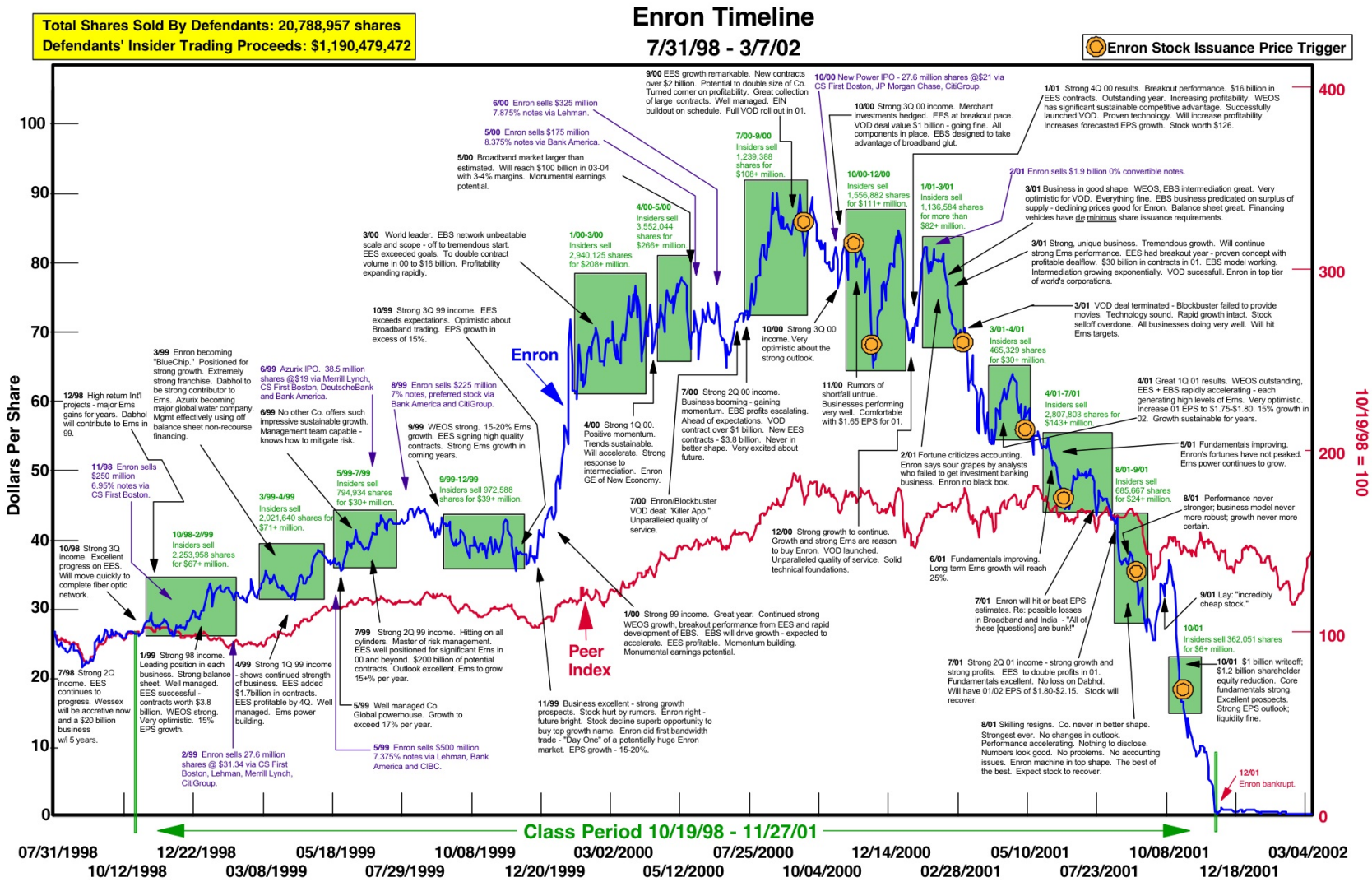
In 2003, the Federal Energy Regulatory Commission posted the company's e-mail on its Web site, about 1.5 million messages. After duplicates were weeded out, a half-million e-mails were left from about 150 accounts, including those of the company's top executives. Most were sent from 1999 to 2001, a period when Enron executives were manipulating financial data, making false public statements, engaging in insider trading, and the company was coming under scrutiny by regulators.

Because of privacy concerns, large e-mail collections had not previously been made publicly available, so this marked the first time scientists had a sizable e-mail network to

New York Times



New York Times



www.enronfraud.com

Anomaly Detection (Chatter)

Want detection with excess activity due to chatter amongst the 2-neighbors:

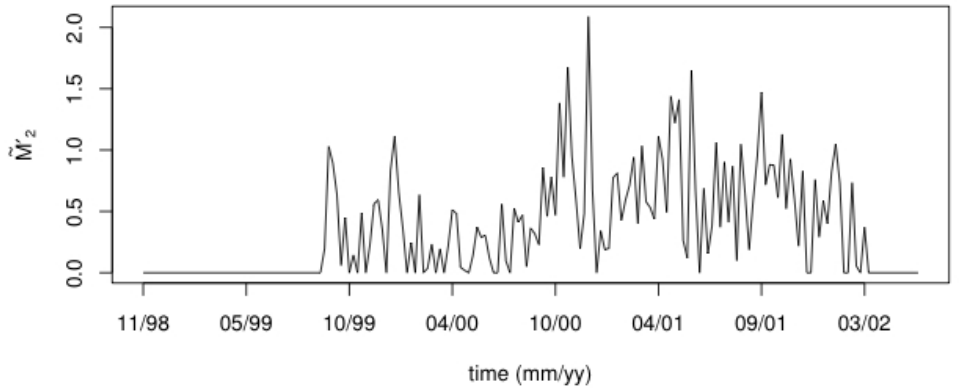
$$\widetilde{\Psi}'_t(v) = (\widetilde{\Psi}_{2,t}(v) \cdot \Im_{t,\tau}(v))/\max(\gamma_t(v), 1)$$

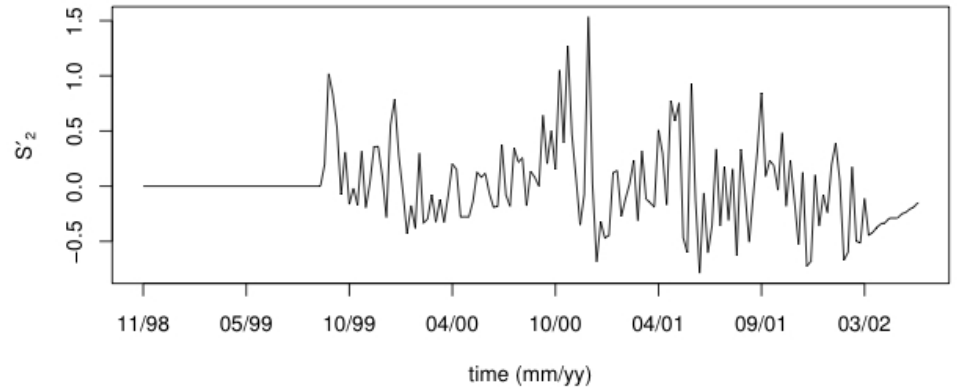
$\Im_{t,\tau}(v) = I_1 \times I_2 \times I_3$

- $I_1 = I\{\widehat{\mu}_{0,t,\tau} > c_1\}$
- $I_2 = I\{\Psi_0(v) < \widehat{\sigma}_{0,t,\tau}(v)c_2 + \widehat{\mu}_{0,t,\tau}(v)\}$
- $I_3 = I\{\Psi_1(v) < \widehat{\sigma}_{1,t,\tau}(v)c_3 + \widehat{\mu}_{1,t,\tau}(v)\}$

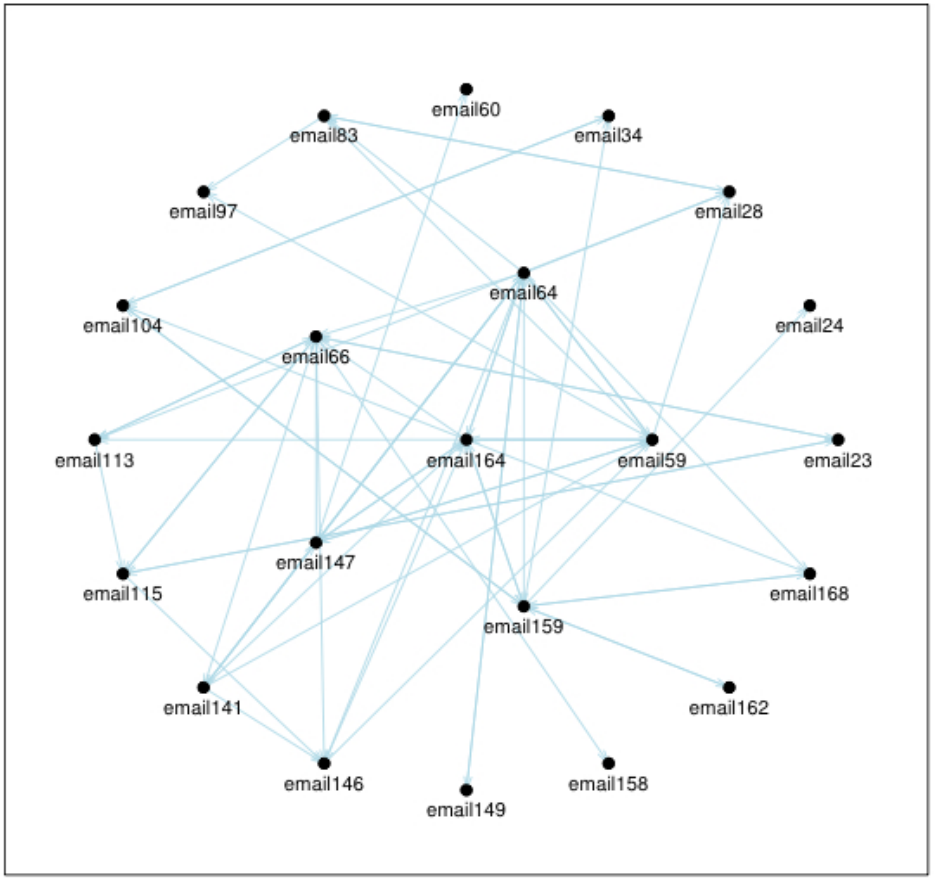
$\gamma_t(v)$ is an "inhomogeneity penalty"

Anomaly Detection (Chatter)



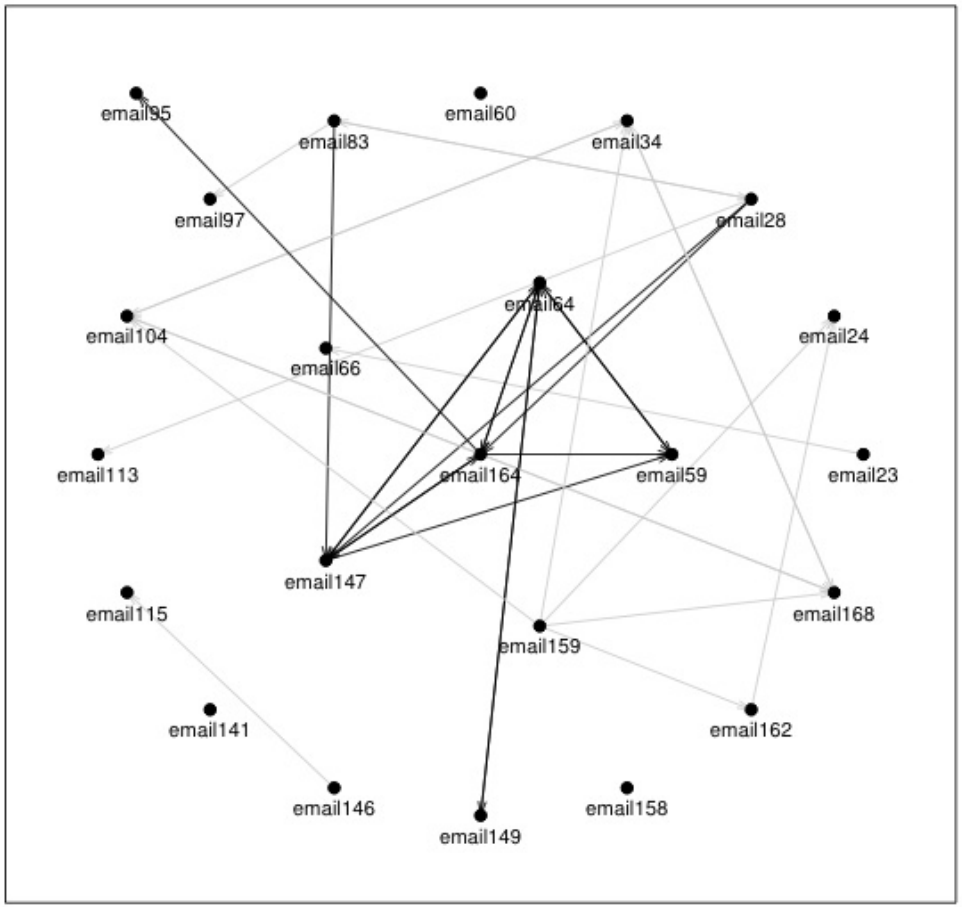


Anomaly Detection (Chatter)



Ω_{109}

Anomaly Detection (Chatter)



Ω_{108}

Conclusions

- scan statistics seems to show promise for detecting anomalies in time series of graphs
- many extentions
- look for upcoming work on *content* and scan statistics for Enron graphs
- work being done using scan statistics for anomaly detection in genetic networks and other application areas

For Futher Information

Including access to the Enron datasets:

- http://www.cis.jhu.edu/ parky/Enron/enron.html

Thank you.

Scan Statistics on Enron Graphs

Kendall Giles

kgiles@cs.jhu.edu

www.kendallgiles.com