

Scan Statistics on Enron Graphs

By Carey Priebe, John Conroy and
David Marchette (2005)

Presented by
Lisa Denogean
February 28, 2006

Introduction

- ◆ During 2001, Enron shares fell from over \$90 to 30 cents, executives unloading stock while encouraging others to buy
- ◆ CEO Kenneth Lay has been accused of selling over \$90 million worth of stock before bankruptcy evident in Nov 2001

Objective

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

Find point in time where there exists excessive activity among Enron email addresses

Outline

- ◆ Scan Statistics on Graphs
- ◆ Anomaly Detection for Enron Data
- ◆ Other Detections:
 - Aliasing
 - Non-Zero Baseline
 - “Chatter”
- ◆ Further Work

Scan Statistics

Moving Window Analysis

- ◆ Scan a small window over data, calculating some locality statistic for each window

Scan Statistic = $M(X)$

- ◆ The supremum or maximum of the locality statistics

Statistically significant signal (nonhomogeneity)

- ◆ Specify cutoff for “large” statistic, such that $P_{H_0}[M(X) \geq c_\alpha] = \alpha$

Scan Statistics on Graphs

For vertices v and w in directed graph $D=(V, A)$

- ◆ k^{th} order neighborhood:

$$N_k[v; D] = \{w \in V(D) : d(v, w) \leq k\}$$

- ◆ Scan region (induced subdigraph): $\Omega(N_k[v; D])$

- ◆ Locality Statistic (e.g. size) at scale k :

$$\Psi_k(v) = |A(\Omega(N_k[v; D]))|$$

- ◆ Scale-specific statistic (e.g. max):

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

- ◆ Variable scale scan statistic:

Let $K \subset \{1, \dots, n - 1\}$ be a collection of scales

Ψ'_k a scale-standardized statistic

Can find $g_{k,\alpha}(\Psi_k(v))$ where for all v and k

$$\Psi'_k(v) = g_{k,\alpha}(\Psi_k(v)) \text{ st } P[\Psi'_k(v) \geq c_\alpha] = \alpha$$

Then statistic is given by

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

Each locality statistic has same probability of rejection

Enron Data

- ◆ Graphs based on 184 users (unique email addresses) in 189 weeks, from 1998 to 2002
- ◆ 125,409 distinct messages, mostly executives
- ◆ Directed arc $A_t = \{(v, w) : v \text{ sends at least one email to } w \text{ during } t^{\text{th}} \text{ week}\}$
- ◆ Dataset processed as time series of digraphs D_1, \dots, D_{189}
- ◆ Assume short-time stationarity under null

Goal: Discover anomalous digraphs from recent past
(Subdigraphs with unusually high “chatter”)

Statistics and Time Series

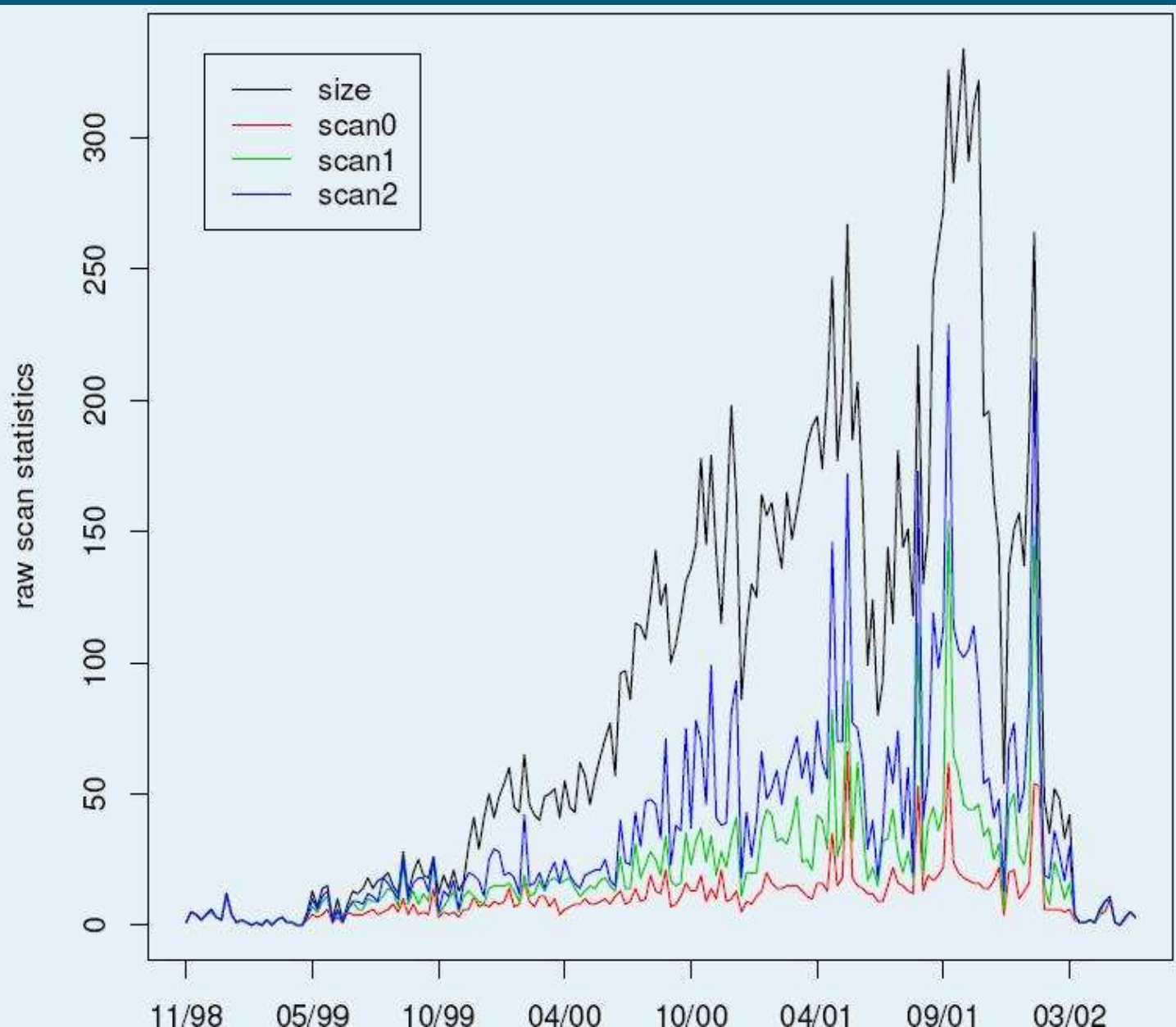
scale- k locality statistics: $\Psi_{k,t}(v) = |A(\Omega(N_k[v; D_t]))|$

- $k = 0$: $\Psi_{0,t}(v) = \text{outdegree}(v; D_t)$.
- **scan statistic:** $M_{k,t} = \max_v \Psi_{k,t}(v); k = 0, 1, 2$

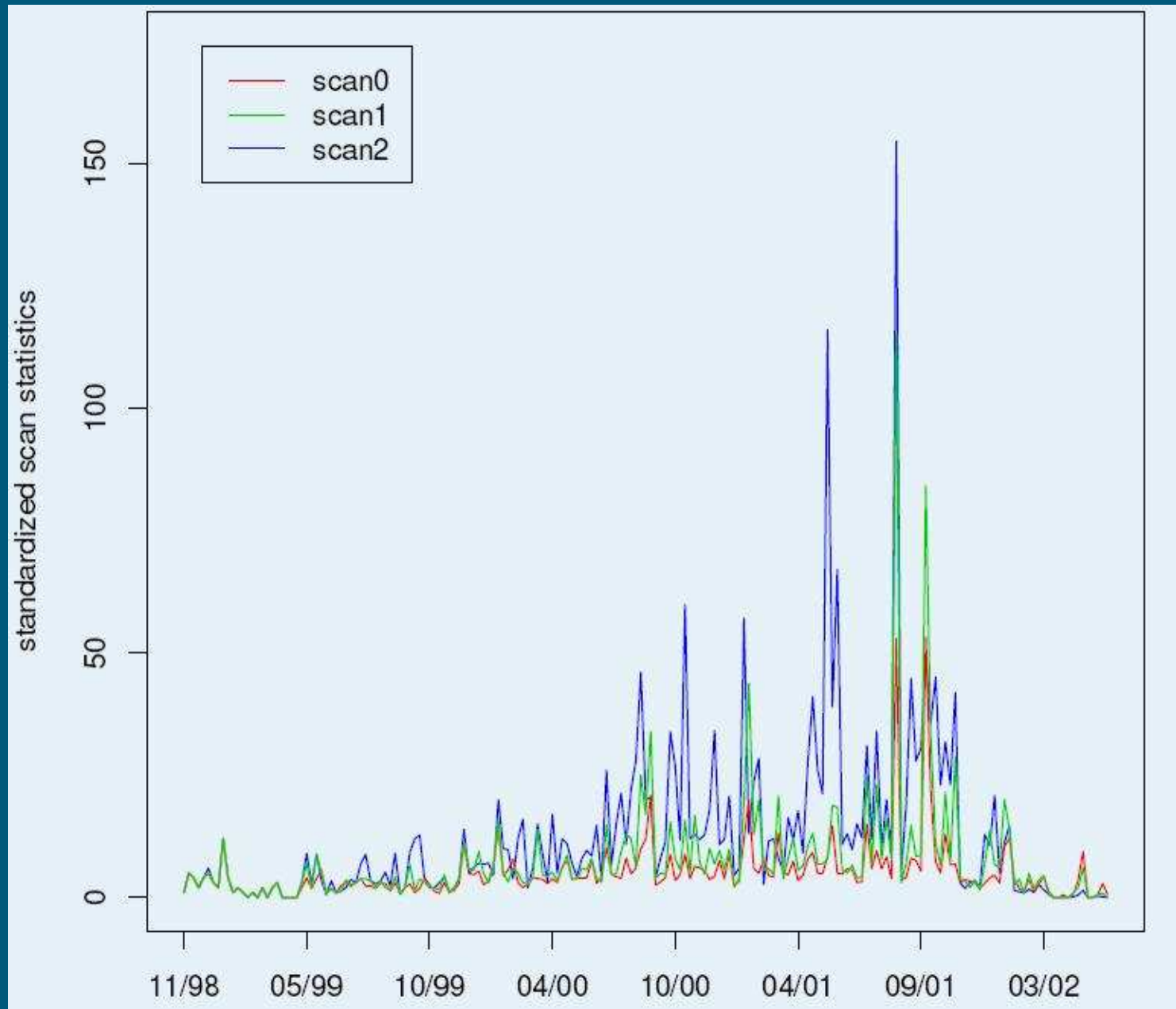
vertex-dependent standardized locality statistic:

- $\tilde{\Psi}_{k,t}(v) = (\Psi_{k,t}(v) - \hat{\mu}_{k,t,\tau}(v)) / \max(\hat{\sigma}_{k,t,\tau}(v), 1)$
- $\hat{\mu}_{k,t,\tau}(v) = \frac{1}{\tau} \sum_{t'=t-\tau}^{t-1} \Psi_{k,t'}(v)$ ($\tau = 20$)
- $\hat{\sigma}_{k,t,\tau}^2(v) = \frac{1}{\tau-1} \sum_{t'=t-\tau}^{t-1} (\Psi_{k,t'}(v) - \hat{\mu}_{k,t,\tau}(v))^2$
- **standardized scan statistic:** $\tilde{M}_{k,t} = \max_v \tilde{\Psi}_{k,t}(v)$

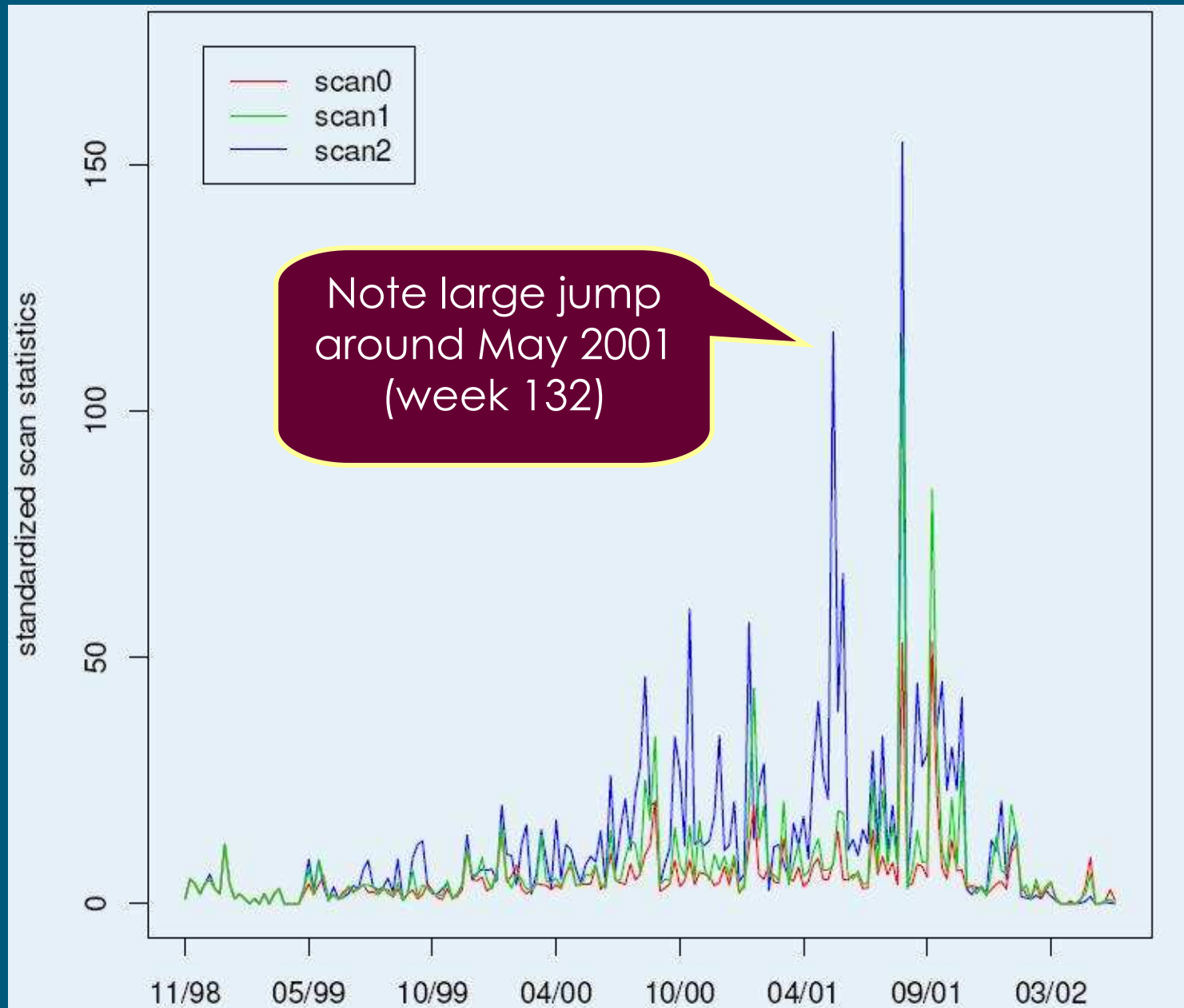
Raw scan statistics for $k=0, 1, 2$



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



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



Anomaly Detection

Temporally-normalized scan statistic, $S_{k,t}$:

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

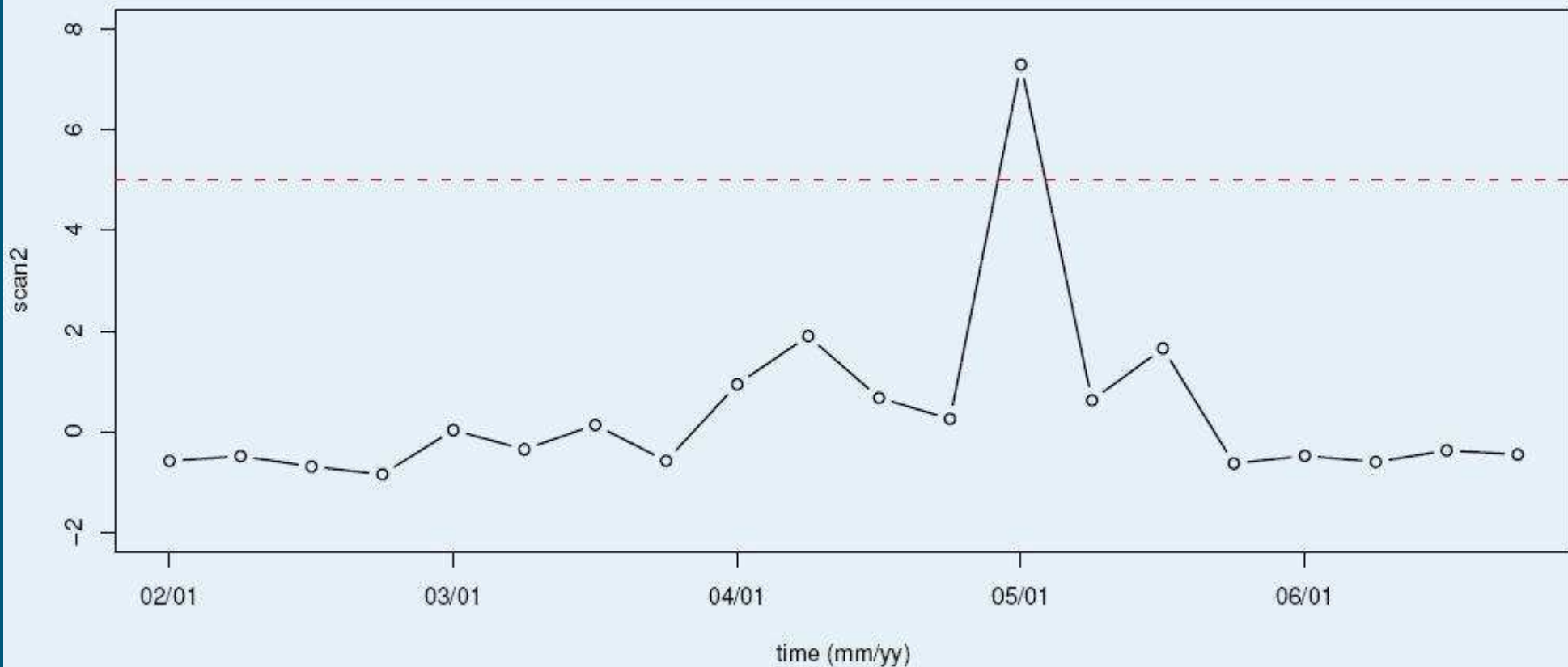
μ and σ are the running mean and standard deviation for $\ell = 20$ time steps

Detection for 5 std dev above mean:

Time t such that $S_{k,t} > 5$

normality assumption

Second order scan statistic indicates anomaly at $t^* = 132$ (May 2001), evident in prev graph



⇒ Identify $S_{2,132}$ as significant (7.3 std dev)

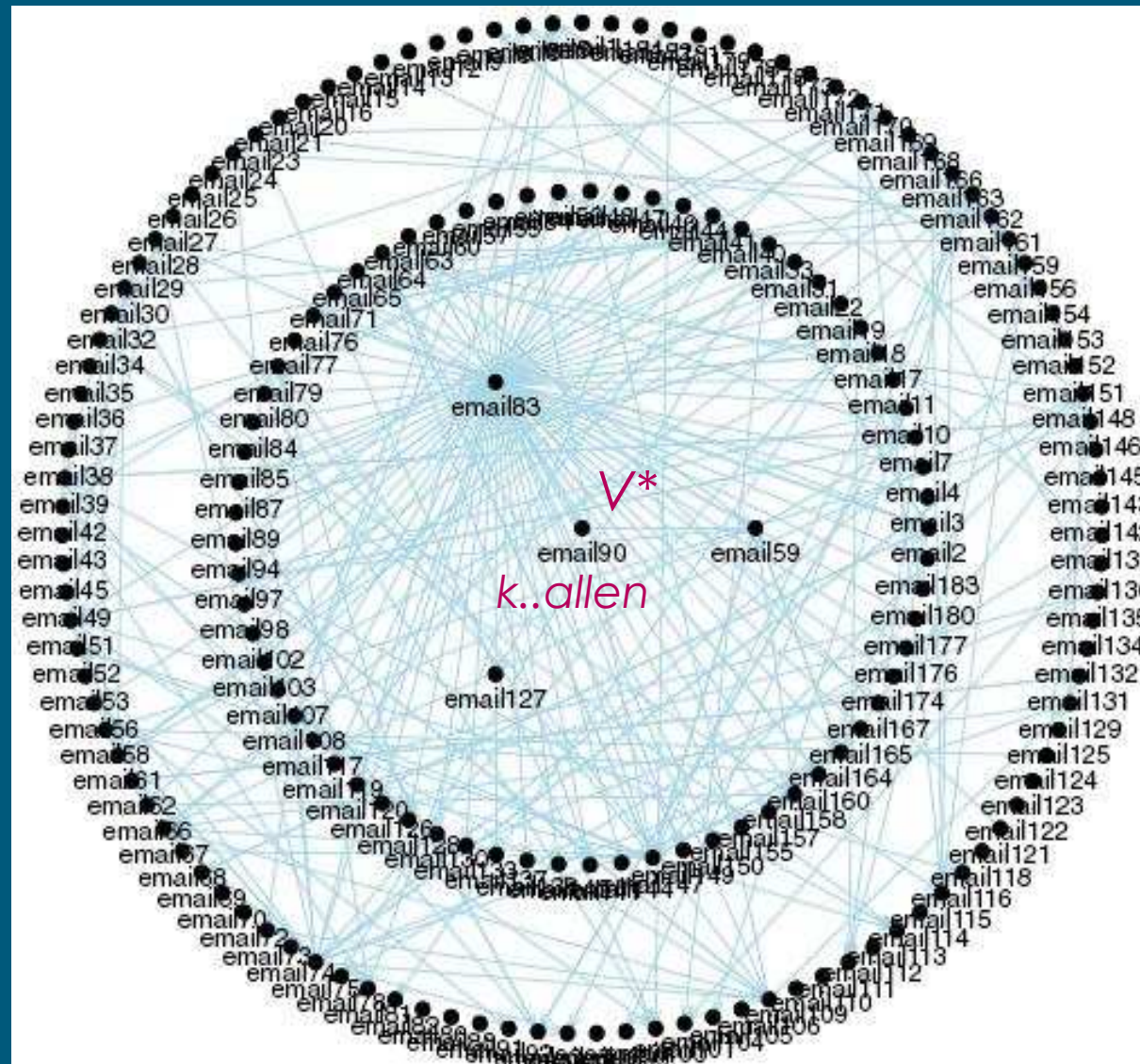
Critical Values of $S_{k,t}$

- ◆ Under normal, 7.3 standard deviations
p-value $< 10^{-10}$
- ◆ Under Gumbel, extreme value distribution
p-value $< 10^{-6}$

Using Bonferroni adjustment:

- ◆ $\tilde{\psi}_{k,t} \sim t_{19} \Rightarrow$ significant detection
- ◆ $\tilde{\psi}_{k,t} \sim$ Cauchy \Rightarrow may not be significant

Detection Graph D_{132}



Vertex of interest $v^* = \operatorname{argmax}_v \tilde{\psi}_{2,132}(v) = \text{email90} = k..allen$

Other Statistics

(not significant)

$$\operatorname{argmax}_v \psi_{0,132}(v) = \text{email83}$$

$$\operatorname{argmax}_v \psi_{1,132}(v) = \text{email83}$$

$$\operatorname{argmax}_v \psi_{2,132}(v) = \text{email147}$$

$$\operatorname{argmax}_v \tilde{\psi}_{0,132}(v) = \text{email147}$$

$$\operatorname{argmax}_v \tilde{\psi}_{1,132}(v) = \text{email75}$$

Detection for $v^* = \text{email90} = k..allen$ apparent only when using standardized second order scan statistic

Statistics for D_{132}

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

Scan statistics from raw locality statistics, vertex standardized statistics and temporally normalized version of M_k (S_k)

Statistics for D_{132}

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

Excessive activity in closed 2-neighborhood of v^* not accounted for by its outdegree or 1-neighborhood.
 v^* communicates with other vertices of high degree
 \Rightarrow excessive local activity

Aliasing

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

k..allen == phillip.allen?

- *k..allen* had no activity before $t^* = 132$.
- At $t^* = 132$, *phillip.allen* switched to *k..allen*.

Aliasing

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

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 $v \in V \setminus \{v^*\}$,

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

Number vertices with length 1 from v and v^* during time $t^* - \kappa$ to $t^* - 1$ ($\kappa \geq 5$)

$$phillip.allen = email.141 = \operatorname{argmax}_v s_{t^*, \kappa}$$

Another Detection

- ◆ Detection of *email90* due to inactivity prior to $t^* = 132$
- ◆ Interested in detection for increases from non-zero baseline
- ◆ Consider

$$\tilde{\Psi}_{k,t}(v) \cdot I\{\hat{\mu}_{0,t,\tau}(v) > c\}$$

Another Detection

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]

- Increase in activity of $v^* = \text{roy.hayslett=email152}$ for order 2 and 3 statistics at $t^* = 152$

Another Detection

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]

email154 = sally.beck

$\Psi_{k,t^*-5:t^*}(v)$
[3 , 2 , 0 , 2 , 3 , 62]
[3 , 3 , 0 , 3 , 6 , 154]
[4 , 3 , 0 , 37 , 11 , 229]
[4 , 3 , 0 , 98 , 16 , 267]

- ◆ Increase in activity of $v^* = \text{roy.hayslett} = \text{email152}$ for order 2 and 3 statistics at $t^* = 152$
- ◆ However, investigation shows activity due to communication between *email152* and *email154*, an order 0 detection at $t^* = 152$

“Chatter”

- ◆ Minimal level of recent activity required
- ◆ Order 0 and order 1 statistics do not yield detections
- ◆ Detect excess activity among 2-neighbors
- ◆ Detect “balanced chatter” by using an inhomogeneity penalty $\gamma_t(v) = \text{std dev of outdegrees of neighbors}$
(rules out *email152/email154* detection events)

Detecting Chatter

Seek detection for activity due to chatter among 2-neighbors

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

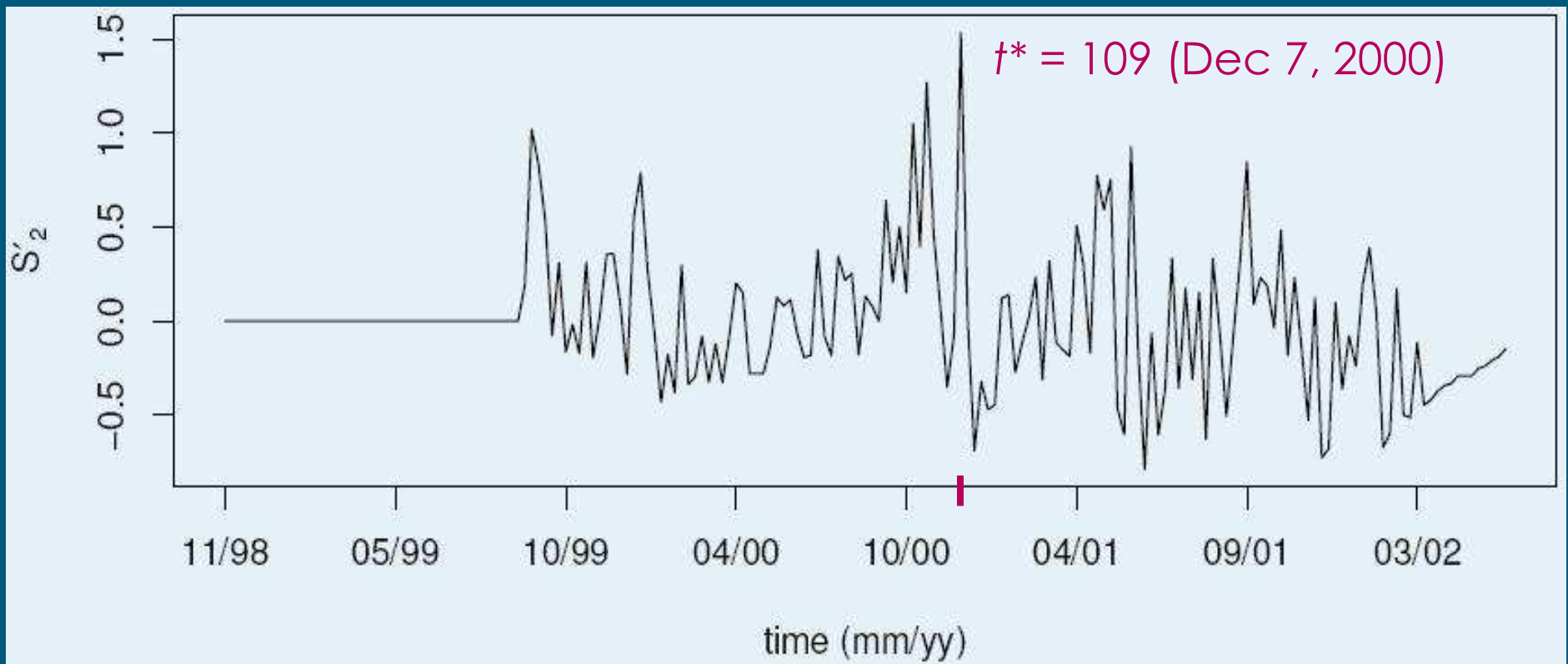
$$\mathcal{I}_{t,\tau}(v) = I_1 \times I_2 \times I_3$$

$$I_1 = I\{\hat{\mu}_{0,t,\tau} > c_1\},$$

$$I_2 = I\{\Psi_0(v) < \hat{\sigma}_{0,t,\tau}(v)c_2 + \hat{\mu}_{0,t,\tau}(v)\},$$

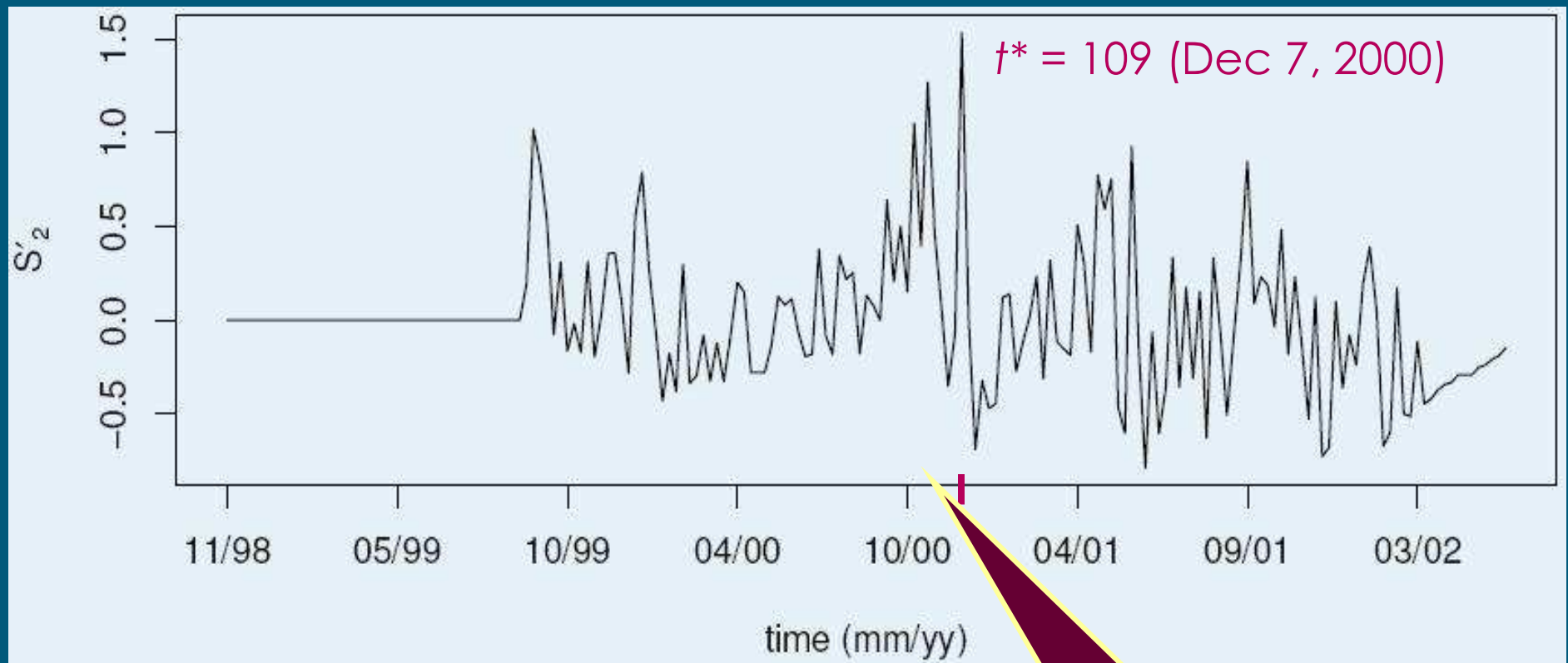
$$I_3 = I\{\Psi_1(v) < \hat{\sigma}_{1,t,\tau}(v)c_3 + \hat{\mu}_{1,t,\tau}(v)\}.$$

S'_t - temporally normalized $\max_v \tilde{\psi}'_t(v)$



$$\operatorname{argmax} \tilde{\psi}'_t(v) = (\text{email164}, 109)$$

S'_t - temporally normalized $\max_v \tilde{\psi}'_t(v)$



$\operatorname{argmax} \tilde{\psi}'_t(v) = (\text{email164}, 109)$

Kenneth Lay
begins selling
shares

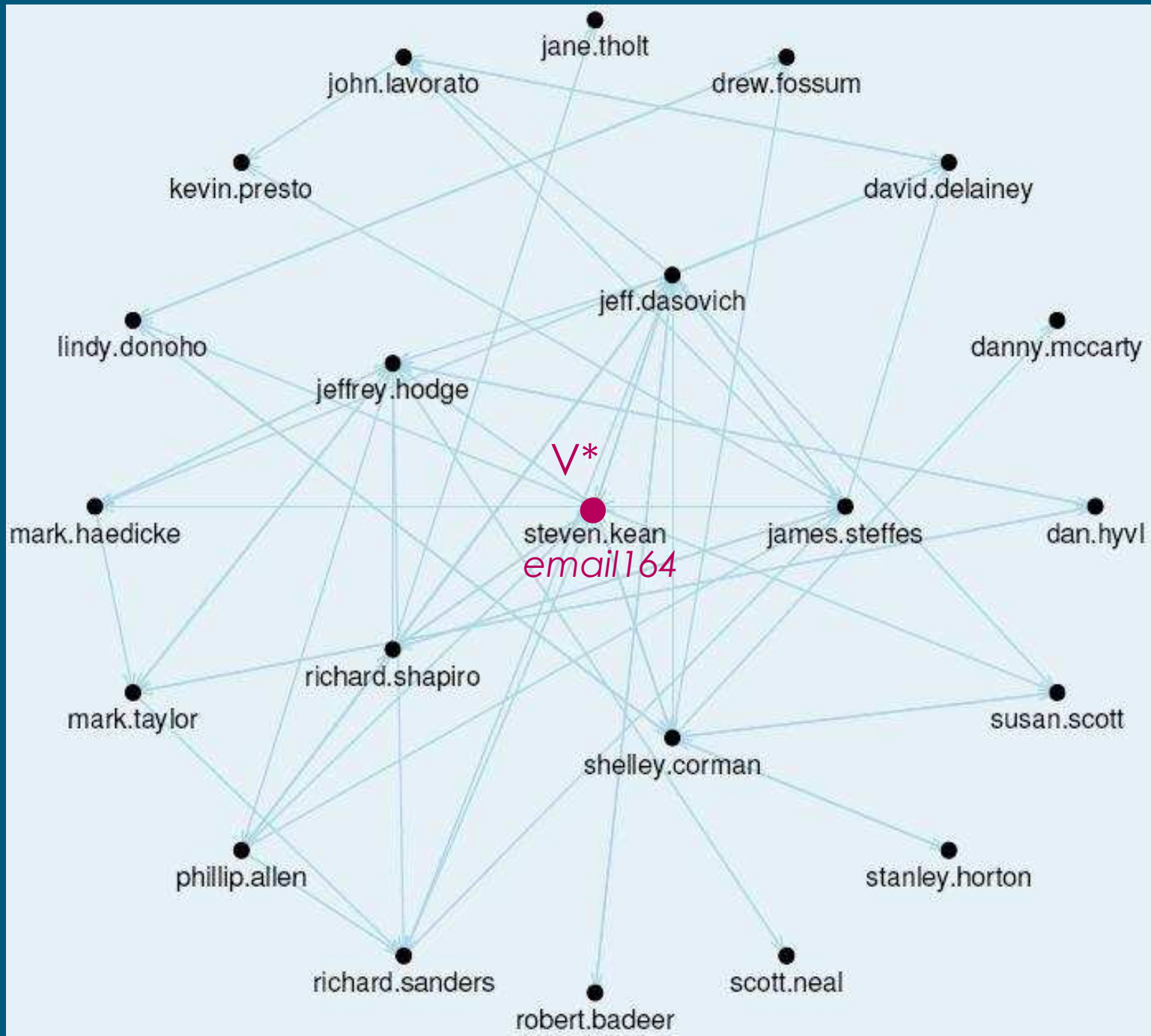
Raw Statistics for *email164=steven.kean* at $t^* = 109$

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]

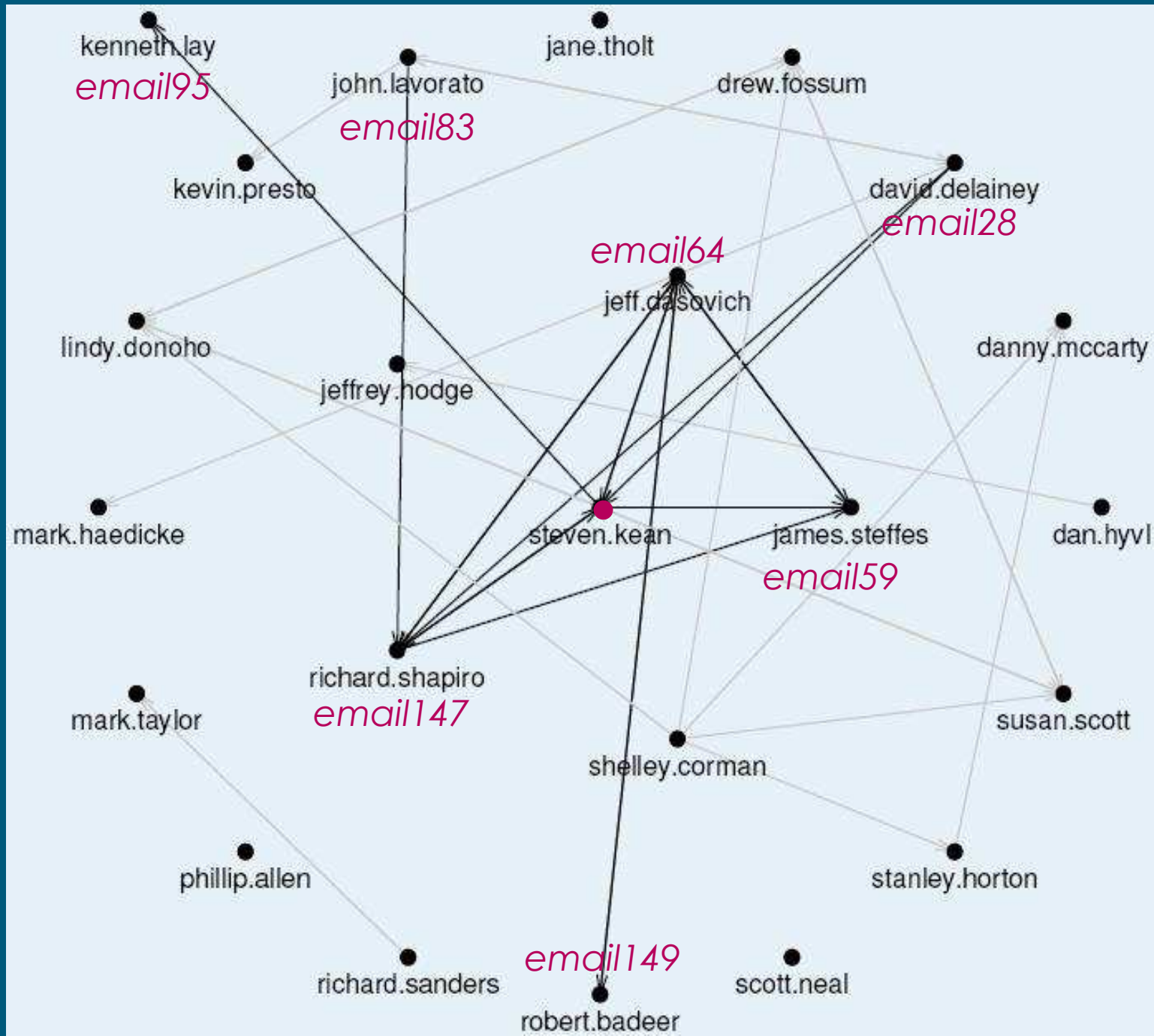
Inhomogeneity penalty $\gamma_t(v) = 1.7$

Consider $t^* - 1 = 108$ and *email164*'s 2-neighborhood at $t = 108$ and $109 \dots$

Subdigraph at $t^* = 109$ ($v^* = \text{steven.kean} = \text{email164}$)



Subdigraph at $t = 108$ ($v^* = \text{steven.kean} = \text{email164}$)



Chatter Results

- ◆ $v^* = \text{email164} = \text{steven.kean}$ has five neighbors with outdegrees 6, 6, 6, 7, 10 at $t^* = 109$
- ◆ Detection is due to v^* communicating with a moderate subset of vertices, each of whom communicates with another moderate subset
- ◆ Vertices in the neighborhood at t^* , while active at t^*-1 have increased their activity
- ◆ Detection is not only due to v^* joining a new group, the group itself is more active as well
- ◆ This suggests detection is robust – insensitive to small changes in the graph

Further Analysis - Clustering

Direct additional analysis, such as clustering, toward potentially informative events found via scan statistics

- ◆ Find emails with similar content based on terms that occur
- ◆ Restrict attention to *signature terms*:
 - Terms occurring more than expected
 - Based on mutual information
 - Dunning 1993, Hovy & Lin 2000
- ◆ Two documents connected if significant overlap in signature terms

Questions

- ◆ Why are 2-neighbors so significant?
- ◆ Other statistics...
 - emails that get forwarded multiple times
 - tight-knit groups
- ◆ What about the distribution of other statistics (only considered $S_{k,t}$) to accurately determine p -values?

Further Work

- ◆ Exponential smoothing, detrending and variance stabilization may be appropriate for time series
- ◆ Multivariate time series (one series for each vertex) could be used with vector autoregressive models
- ◆ Extend scan statistics to weighted graphs and hypergraphs for detection of anomalous number and type of messages
- ◆ Perform in a streaming data environment (sliding one-week window)

Other Applications - Biochatter

- ◆ Time-series microarray data, t columns are time points for g genes
- ◆ For each t , perform univariate or multivariate model-based clustering on the g genes
- ◆ The g genes are vertices and undirected edges are drawn if two genes belong to same cluster
- ◆ Generate temporally normalized scan statistics $S_{k,t}$ to find anomalous gene activity

Related Work

- ◆ “Content and Scan Statistics for Enron” – John Conroy et al.
- ◆ “Random Dot Product Graphs” – Ed Scheinerman et al.
- ◆ “Biochatter” – Jeff Solka, Chris Overall, Jennifer Weller
- ◆ Spatial scan statistics: Naus 1965, Adler 1984, Loader 1991, Chen and Glaz 1996
- ◆ p -value approximation: Cressie 1993, Naiman and Priebe 2001 (importance sampling)