Data Wrangling: From Raw Data to Networks

CSCI 5352: Network Analysis and Modeling

Notes written by Dr. Leto Peel
Université Catholique de Louvain

Creating networks from data

When creating networks from data we need to make a number of design descisions

- How will we collect the data?
- What type of entity (node) to use and how to extract it?
- What type of relationship or interaction do our links represent?
- What time period?
- Directed or undirected links?

Creating networks from data

When creating networks from data we need to make a number of design descisions

- How will we collect the data?
- What type of entity (node) to use and how to extract it?
- What type of relationship or interaction do our links represent?
- What time period?
- Directed or undirected links?

How we make these decisions depends on:

- the task we're trying to achieve
- the model and algorithm we are using

Motivating Example

Change-point detection

Aim: To understand how external events “shocks” are related to changes in network structure

Datasets

Two datasets

Enron

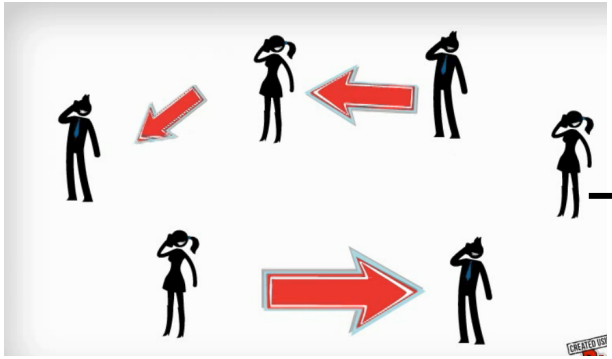
- Criminal investigation
- Single network
- Semi-structured

MIT Reality Mining

- Consenting participants
- Multiple networks
- Structured data

MIT Reality Mining dataset

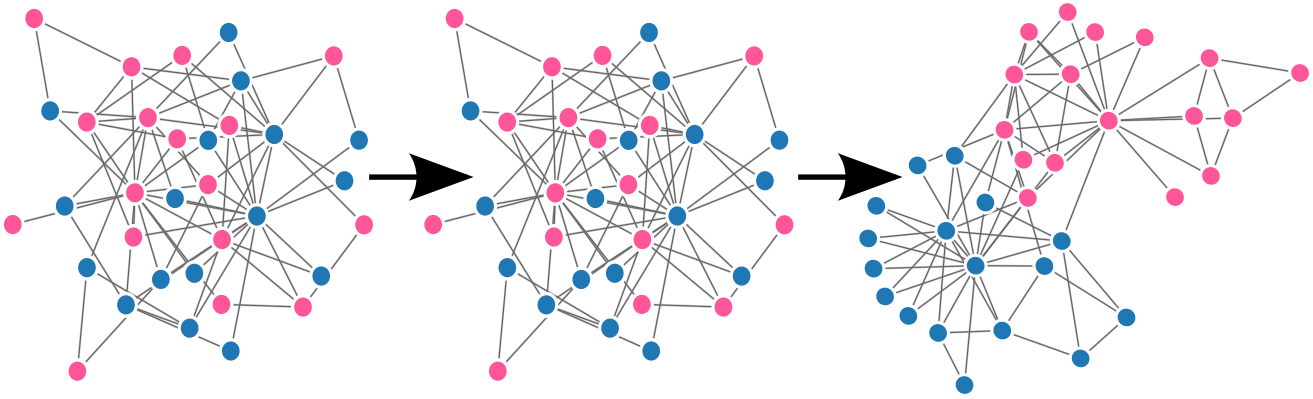
- 94 participants
  - 68 MIT Media Lab (90% graduate students, 10% staff)
  - 26 incoming Sloan business school students
- September 2004 and June 2005
- Rich dataset (phone data + survey data)
- Incentive: free use of exclusive phone



Real Interactions

Sensors

Data



Network Snapshots

Rich Data

Phone data

- Communication events (voice, sms)
- Phone charge status
- Phone active / on?
- Location (cell tower)
- Bluetooth devices
- App usage

Survey data

- Who are your friends?
- Have you travelled recently?
- Do you own a car?
- How long into the term did it take for your social circle to become what it is today?
- Preferred work/personal communication medium?

Which network?

Which network?

Friendship network

Which network?

Friendship network

Voice call network

SMS network

Which network?

Friendship network

Voice call network

SMS network

Physical proximity network

Noise

The dataset is very noisy.

Sources of noise / missing data:

Noise

The dataset is very noisy.

Sources of noise / missing data:

- phone left at home
- no battery (or being charged)
- sensor error
- date discrepencies (reset)

Link reciprocity

The bluetooth network is, in its raw form, a directed network.

Link reciprocity

The bluetooth network is, in its raw form, a directed network.

It doesn't make sense to have directed network of physical proximities.

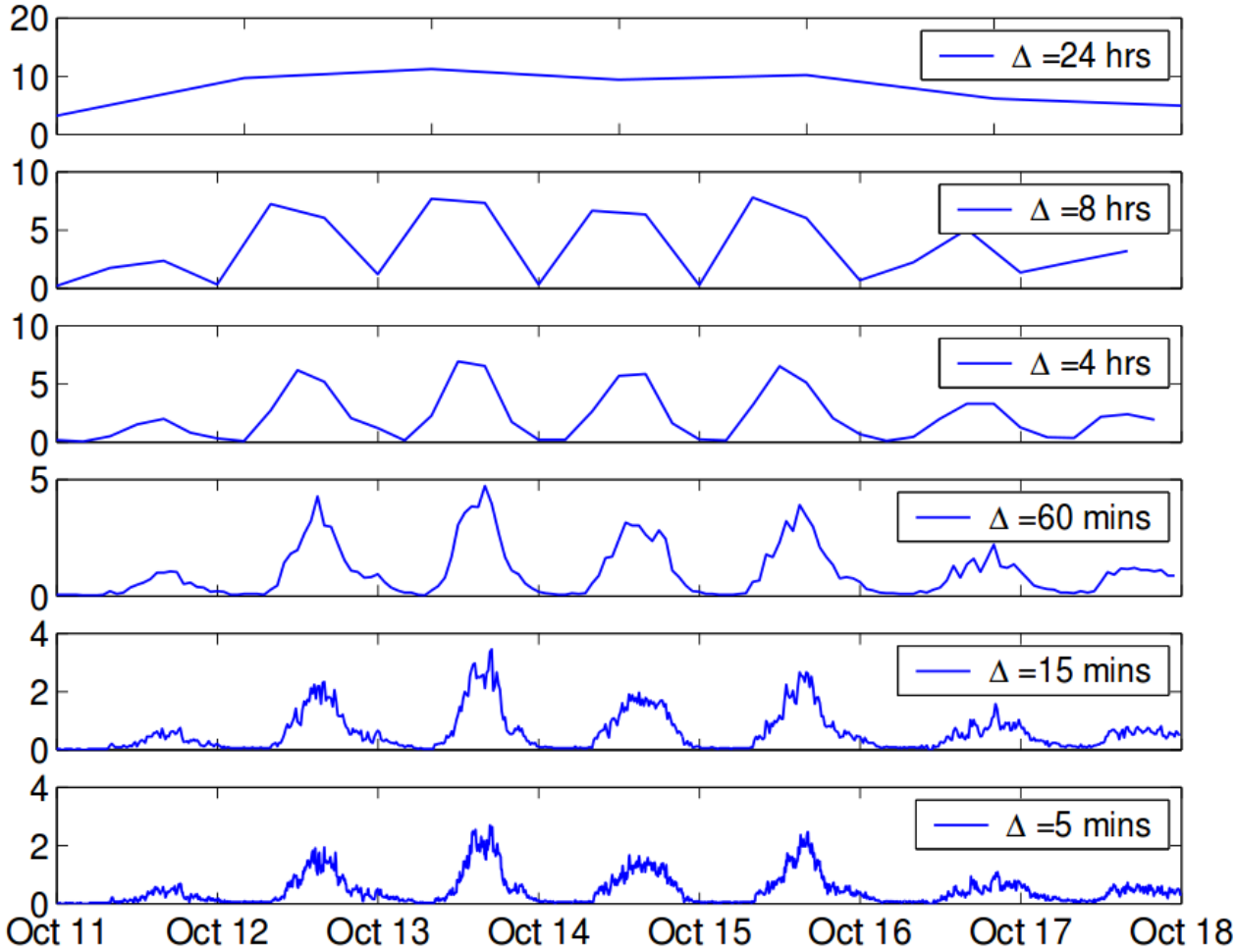
Link reciprocity

- Links that only exist in one direction indicate a mismatch between reality and sensor.
- Two choices: minimal or maximal set. In the minimal set, we discard all unreciprocated edges. In the maximal set, we include all unreciprocated edges.

Temporal resolution

- Bluetooth scans every 2.5 minutes
- What temporal resolution should we use?

Temporal resolution



Reality Mining Summary

- We have an enormous number of choices for how to even turn our data into a network.
- What time resolution?

Reality Mining Summary

- We have an enormous number of choices for how to even turn our data into a network.
- What time resolution?
- How should we deal with missing data?

Reality Mining Summary

- We have an enormous number of choices for how to even turn our data into a network.
- What time resolution?
- How should we deal with missing data?
- What about the Hawthorne Effect? (Behavior changes when people know they're being studied)

Reality Mining Summary

- We have an enormous number of choices for how to even turn our data into a network.
- What time resolution?
- How should we deal with missing data?
- What about the Hawthorne Effect? (Behavior changes when people know they're being studied)
- How do we avoid the forking paths problem, where we search over all analyses till we find something? How is this different from Exploratory Data Analysis?

Enron email dataset

- Largest supplier of natural gas to North America
- "America's Most Innovative Company" by the magazine Fortune from 1996 to 2001

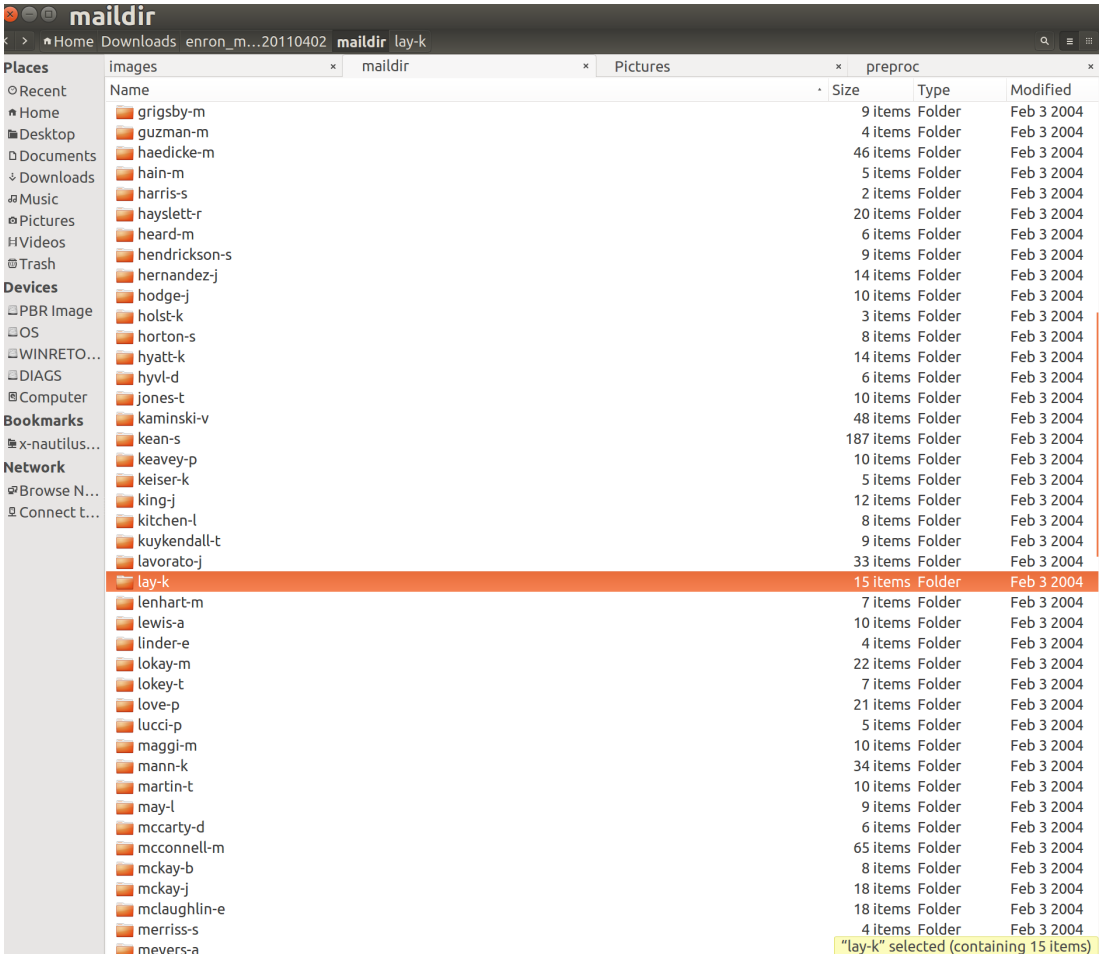
Enron email dataset

- Largest supplier of natural gas to North America
- "America's Most Innovative Company" by the magazine Fortune from 1996 to 2001
- Misrepresentation of earnings and unethical practises
- End of 2001: One of the largest bankruptcies in American history

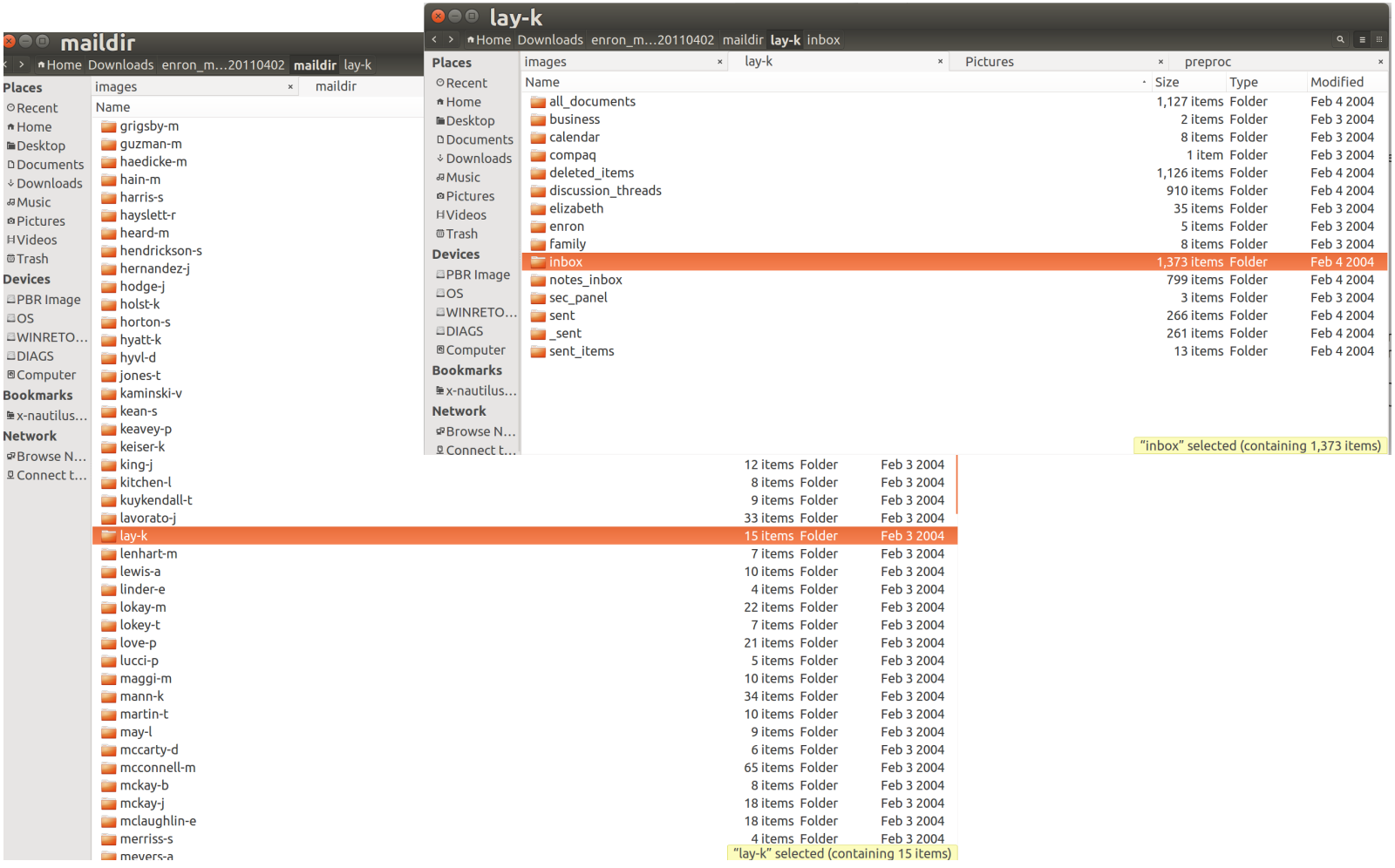
Enron email dataset

- Dataset publically released during the FERC investigation
- 151 Enron employee email accounts
- $\sim$ 0.5 million messages, 1.4GB

Email data

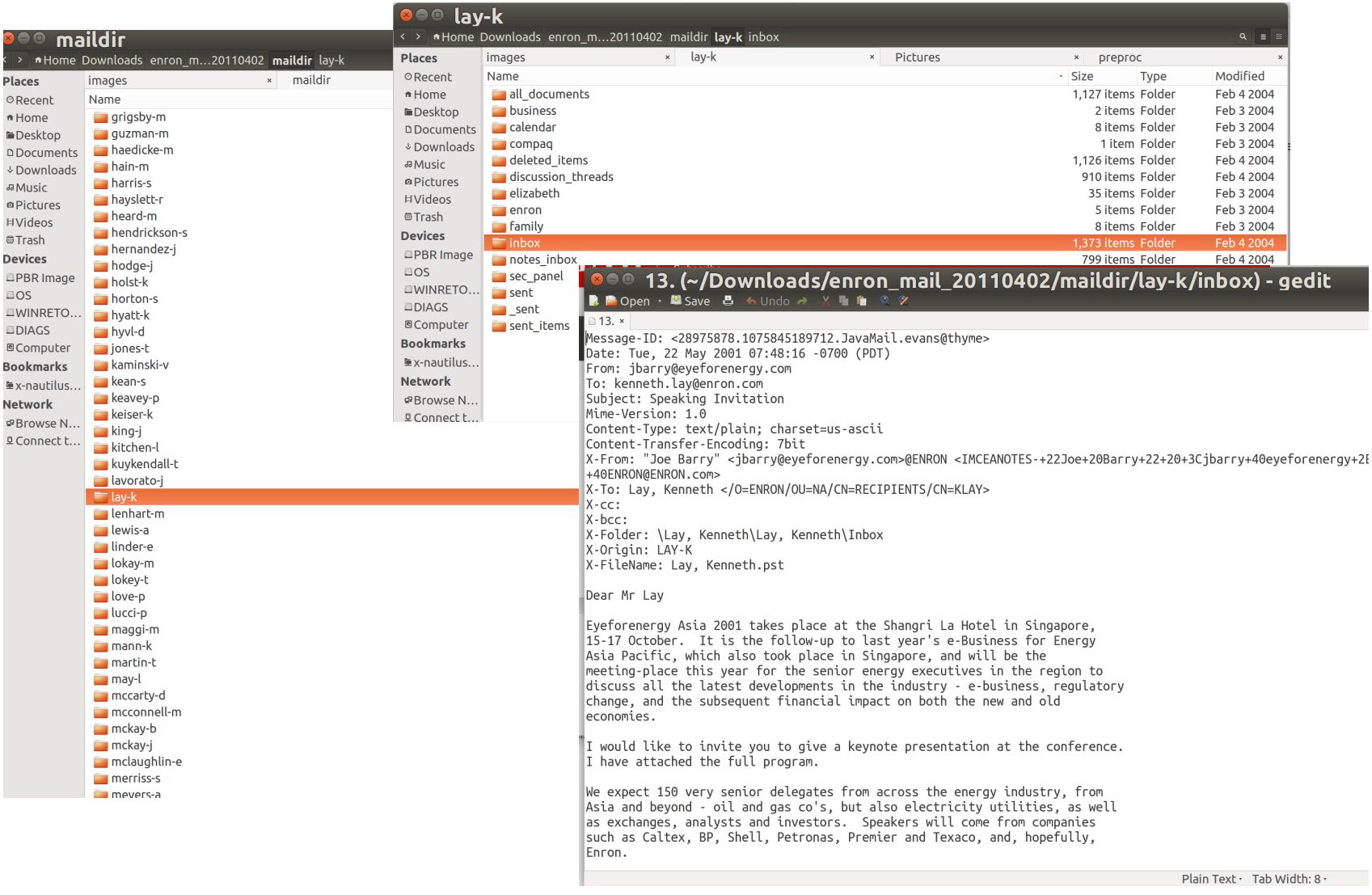

maildir
Home Downloads enron_m...20110402 maildir lay-k
Places
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Devices
PBR Image
OS
WINRETO...
DIAGS
Computer
Bookmarks
x-nautilus...
Network
Browse N...
Connect t...
images
maildir
Pictures
preproc
Name
Size
Type
Modified
grigsby-m 9 items Folder Feb 3 2004
guzman-m 4 items Folder Feb 3 2004
haedicke-m 46 items Folder Feb 3 2004
hain-m 5 items Folder Feb 3 2004
harris-s 2 items Folder Feb 3 2004
hayslett-r 20 items Folder Feb 3 2004
heard-m 6 items Folder Feb 3 2004
hendrickson-s 9 items Folder Feb 3 2004
hernandez-j 14 items Folder Feb 3 2004
hodge-j 10 items Folder Feb 3 2004
holst-k 3 items Folder Feb 3 2004
horton-s 8 items Folder Feb 3 2004
hyatt-k 14 items Folder Feb 3 2004
hyvl-d 6 items Folder Feb 3 2004
jones-t 10 items Folder Feb 3 2004
kaminski-v 48 items Folder Feb 3 2004
kean-s 187 items Folder Feb 3 2004
keavey-p 10 items Folder Feb 3 2004
keiser-k 5 items Folder Feb 3 2004
king-j 12 items Folder Feb 3 2004
kitchen-l 8 items Folder Feb 3 2004
kuykendall-t 9 items Folder Feb 3 2004
lavorato-j 33 items Folder Feb 3 2004
lay-k 15 items Folder Feb 3 2004
lenhart-m 7 items Folder Feb 3 2004
lewis-a 10 items Folder Feb 3 2004
linder-e 4 items Folder Feb 3 2004
lokay-m 22 items Folder Feb 3 2004
lokey-t 7 items Folder Feb 3 2004
love-p 21 items Folder Feb 3 2004
lucci-p 5 items Folder Feb 3 2004
maggi-m 10 items Folder Feb 3 2004
mann-k 34 items Folder Feb 3 2004
martin-t 10 items Folder Feb 3 2004
may-l 9 items Folder Feb 3 2004
mccarty-d 6 items Folder Feb 3 2004
mconnell-m 65 items Folder Feb 3 2004
mckay-b 8 items Folder Feb 3 2004
mckay-j 18 items Folder Feb 3 2004
mclaughlin-e 18 items Folder Feb 3 2004
merriss-s 4 items Folder Feb 3 2004
meyers-a
"lay-k" selected (containing 15 items)

Email data

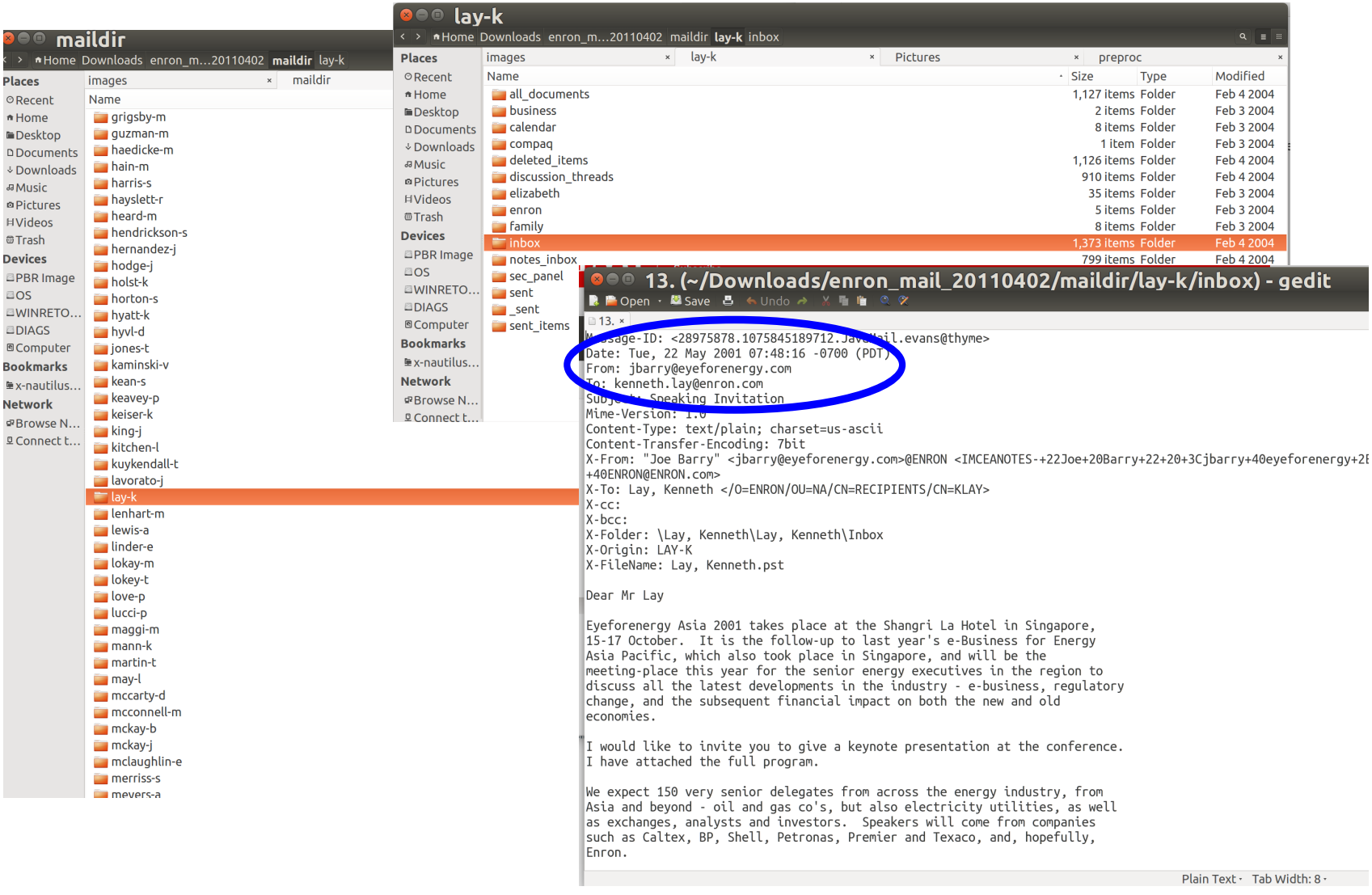


Custom email subfolders (name and depth)!

Email data


maldir
Home Downloads enron_m...20110402 maildir lay-k
Places
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Devices
PBR Image
OS
WINRETO...
DIAGS
Computer
Bookmarks
x-nautilus...
Network
Browse N...
Connect t...
images
maildir
Name
grigsby-m
guzman-m
haedicke-m
hain-m
harris-s
hayslett-r
heard-m
hendrickson-s
hernandez-j
hodge-j
holst-k
horton-s
hyatt-k
hyvl-d
jones-t
kaminski-v
kean-s
keavey-p
keiser-k
king-j
kitchen-l
kuykendall-t
lavorato-j
lay-k
lenhart-m
lewis-a
linder-e
lokay-m
lokey-t
love-p
lucci-p
maggi-m
mann-k
martin-t
may-l
mccarty-d
mcconnell-m
mckay-b
mckay-j
mclaughlin-e
merriss-s
meyers-a
lay-k
Home Downloads enron_m...20110402 maildir lay-k inbox
Desktop
Music
Pictures
Videos
Trash
PBR Image
OS
WINRETO...
DIAGS
Computer
x-nautilus...
Browse N...
Connect t...
images
lay-k
Pictures
preproc
Name
Size
Type
Modified
all_documents 1,127 items Folder Feb 4 2004
business 2 items Folder Feb 3 2004
calendar 8 items Folder Feb 3 2004
compaq 1 item Folder Feb 3 2004
deleted_items 1,126 items Folder Feb 4 2004
discussion_threads 910 items Folder Feb 4 2004
elizabeth 35 items Folder Feb 3 2004
enron 5 items Folder Feb 3 2004
family 8 items Folder Feb 3 2004
inbox 1,373 items Folder Feb 4 2004
notes_inbox 799 items Folder Feb 4 2004
sec_panel
sent
_sent
sent_items
13. (~/Downloads/enron_mail_20110402/maildir/lay-k/inbox) - gedit
Open
Save
Undo
13.
Message-ID: <28975878.1075845189712.JavaMail.evans@thyme>
Date: Tue, 22 May 2001 07:48:16 -0700 (PDT)
From: jbarry@eyeforenergy.com
To: kenneth.lay@enron.com
Subject: Speaking Invitation
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: "Joe Barry" <jbarry@eyeforenergy.com>@ENRON <IMCEANOTES-+22Joe+20Barry+22+20+3Cjbarry+40eyeforenergy+2E
+40ENRON@ENRON.com>
X-To: Lay, Kenneth </O=ENRON/OU=NA/CN=RECIPIENTS/CN=KLAY>
X-cc:
X-bcc:
X-Folder: \Lay, Kenneth\Lay, Kenneth\Inbox
X-Origin: LAY-K
X-FileName: Lay, Kenneth.pst
Dear Mr Lay
Eyeforenergy Asia 2001 takes place at the Shangri La Hotel in Singapore,
15-17 October. It is the follow-up to last year's e-Business for Energy
Asia Pacific, which also took place in Singapore, and will be the
meeting-place this year for the senior energy executives in the region to
discuss all the latest developments in the industry - e-business, regulatory
change, and the subsequent financial impact on both the new and old
economies.
I would like to invite you to give a keynote presentation at the conference.
I have attached the full program.
We expect 150 very senior delegates from across the energy industry, from
Asia and beyond - oil and gas co's, but also electricity utilities, as well
as exchanges, analysts and investors. Speakers will come from companies
such as Caltex, BP, Shell, Petronas, Premier and Texaco, and, hopefully,
Enron.
Plain Text
Tab Width: 8

Email data


maildir
Home Downloads enron_m...20110402 maildir lay-k
Places
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Devices
PBR Image
OS
WINRETO...
DIAGS
Computer
Bookmarks
x-nautilus...
Network
Browse N...
Connect t...
images
maildir
Name
grigsby-m
guzman-m
haedicke-m
hain-m
harris-s
hayslett-r
heard-m
hendrickson-s
hernandez-j
hodge-j
holst-k
horton-s
hyatt-k
hyvl-d
jones-t
kaminski-v
kean-s
keavey-p
keiser-k
king-j
kitchen-l
kuykendall-t
lavorato-j
lay-k
lenhart-m
lewis-a
linder-e
lokay-m
lokey-t
love-p
lucci-p
maggi-m
mann-k
martin-t
may-l
mccarty-d
mcconnell-m
mckay-b
mckay-j
mclaughlin-e
merriss-s
meyers-a
lay-k
Home Downloads enron_m...20110402 maildir lay-k inbox
Places
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Devices
PBR Image
OS
WINRETO...
DIAGS
Computer
Bookmarks
x-nautilus...
Network
Browse N...
Connect t...
images
lay-k
Pictures
preproc
Name
Size
Type
Modified
all_documents 1,127 items Folder Feb 4 2004
business 2 items Folder Feb 3 2004
calendar 8 items Folder Feb 3 2004
compaq 1 item Folder Feb 3 2004
deleted_items 1,126 items Folder Feb 4 2004
discussion_threads 910 items Folder Feb 4 2004
elizabeth 35 items Folder Feb 3 2004
enron 5 items Folder Feb 3 2004
family 8 items Folder Feb 3 2004
inbox 1,373 items Folder Feb 4 2004
notes_inbox 799 items Folder Feb 4 2004
sec_panel
sent
_sent
sent_items
13. (~/Downloads/enron_mail_20110402/maildir/lay-k/inbox) - gedit
Open
Save
Undo
13.
Message-ID: <28975878.1075845189712.JavaMail.evans@thyme>
Date: Tue, 22 May 2001 07:48:16 -0700 (PDT)
From: jbarry@eyeforenergy.com
To: kenneth.lay@enron.com
Subject: Speaking Invitation
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: "Joe Barry" <jbarry@eyeforenergy.com>@ENRON <IMCEANOTES-+22Joe+20Barry+22+20+3Cjbarry+40eyeforenergy+2E
+40ENRON@ENRON.com>
X-To: Lay, Kenneth </O=ENRON/OU=NA/CN=RECIPIENTS/CN=KLAY>
X-cc:
X-bcc:
X-Folder: \Lay, Kenneth\Lay, Kenneth\Inbox
X-Origin: LAY-K
X-FileName: Lay, Kenneth.pst
Dear Mr Lay
Eyeforenergy Asia 2001 takes place at the Shangri La Hotel in Singapore,
15-17 October. It is the follow-up to last year's e-Business for Energy
Asia Pacific, which also took place in Singapore, and will be the
meeting-place this year for the senior energy executives in the region to
discuss all the latest developments in the industry - e-business, regulatory
change, and the subsequent financial impact on both the new and old
economies.
I would like to invite you to give a keynote presentation at the conference.
I have attached the full program.
We expect 150 very senior delegates from across the energy industry, from
Asia and beyond - oil and gas co's, but also electricity utilities, as well
as exchanges, analysts and investors. Speakers will come from companies
such as Caltex, BP, Shell, Petronas, Premier and Texaco, and, hopefully,
Enron.
Plain Text
Tab Width: 8

Entity extraction

- To build a network we need to identify the nodes (i.e. the email addresses)

Entity extraction

- To build a network we need to identify the nodes (i.e. the email addresses)
- >15,000 unique email addresses
  - Only 151 employees part of the investigation
  - Spam?
  - Scalability issue?

Identifying the right entities

How to identify key employee email addresses?

Identifying the right entities

How to identify key employee email addresses?

- Custom folders so we can't check "Sent mail" folder for sender address

Identifying the right entities

How to identify key employee email addresses?

- Custom folders so we can't check "Sent mail" folder for sender address
- Similar issue with checking "Inbox" + this includes mailing lists

Identifying the right entities

How to identify key employee email addresses?

- Custom folders so we can't check "Sent mail" folder for sender address
- Similar issue with checking "Inbox" + this includes mailing lists
- Doesn't match the most frequently occurring emails

Metadata

- Often we use metadata (non-network data) as part of network analysis
- e.g. Comparing large-scale structure to node level information

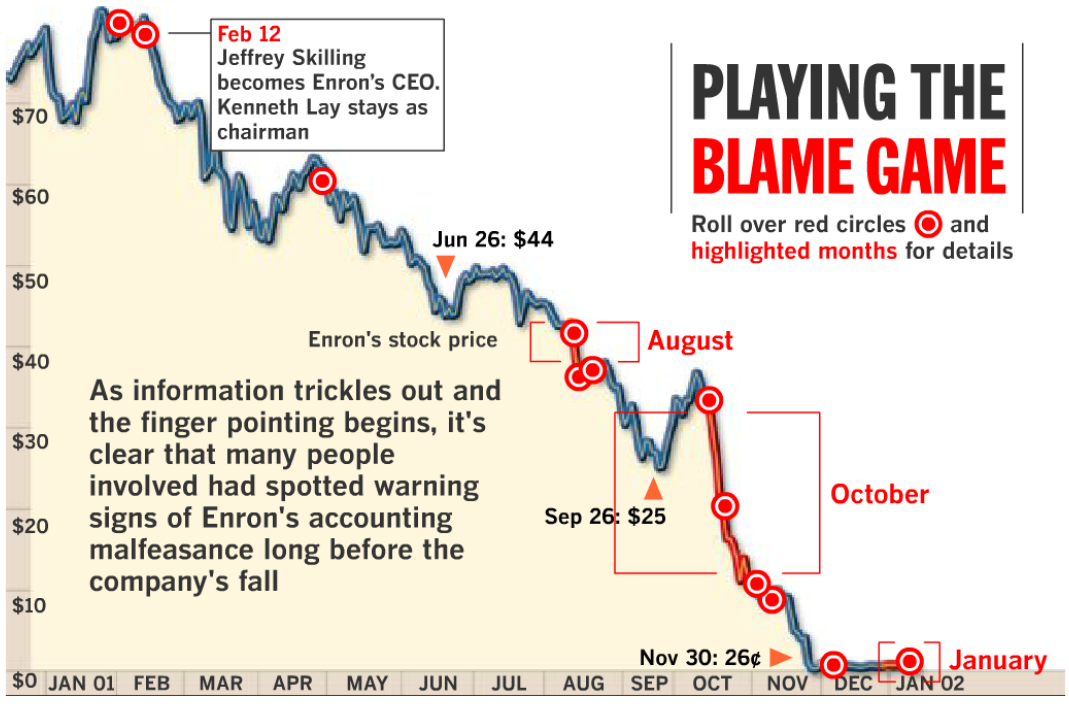
- For change-point detection we are interested in how changes relate to external events

Events

Enron's Collapse

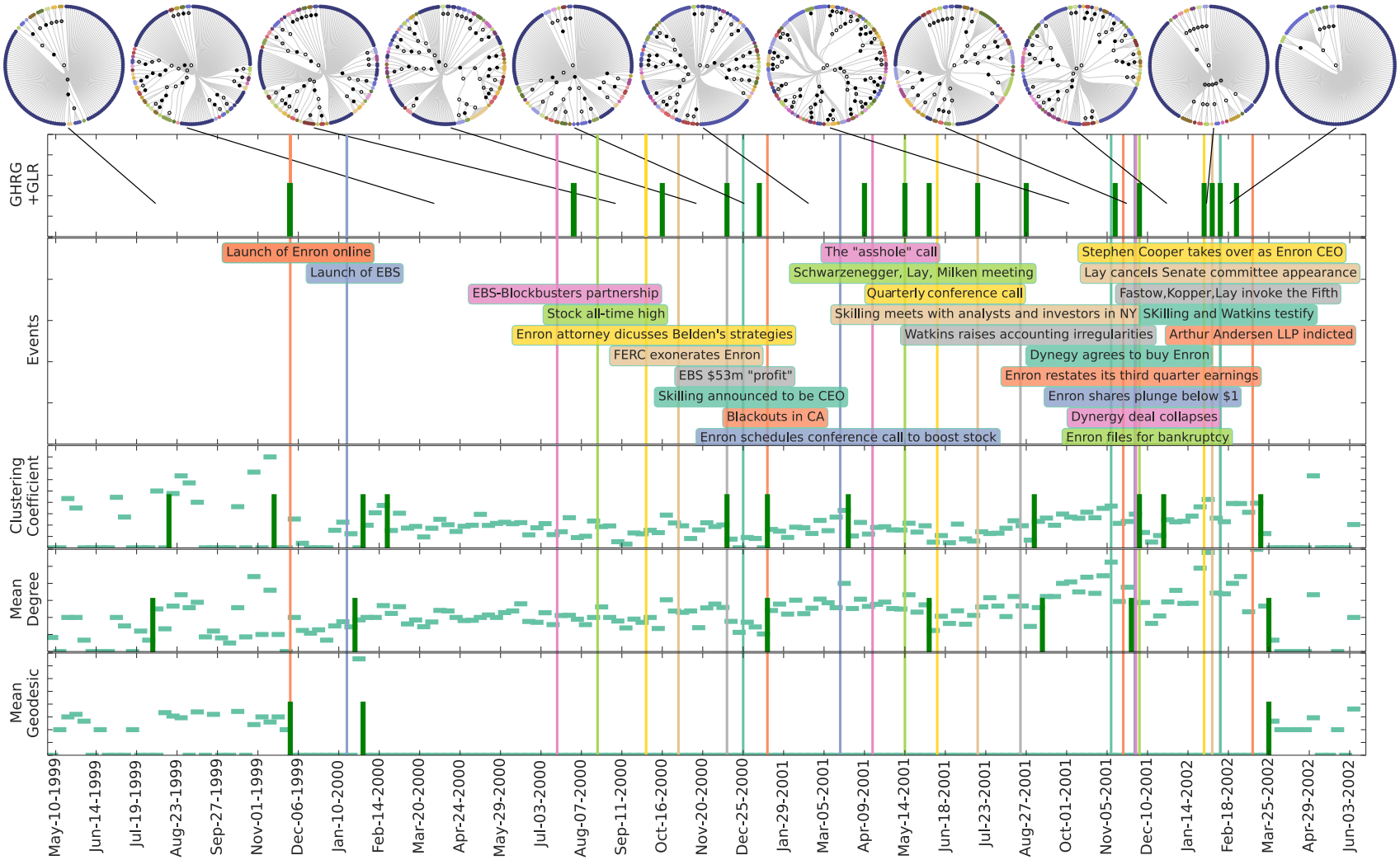
Digg Facebook Twitter Linkedin Email

A month-by-month look at Enron's collapse



Source: time.com

Detecting change points



How does the method work?

- Fit the network using a generalized hierarchical random graph model.
- Use a Bayesian hypothesis test to quantitatively determine if, when, and precisely how a change point has occurred. (In short: when do the data switch from supporting one model to supporting another?)
- Method validated by analyzing the detectability of change points using synthetic data with known change points of different types and magnitudes.

Resources

Further reading:

Datasets:

- Enron emails: `https://www.cs.cmu.edu/~./enron/`
- MIT Reality Mining: `http://realitycommons.media.mit.edu/realitymining.html`

Code:

- Enron parser: `https://piratepeel.github.io/datasets.html`

Software:

- Python
- Matlab
- Octave