

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 decisions

- ► 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 decisions

- ▶ 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:

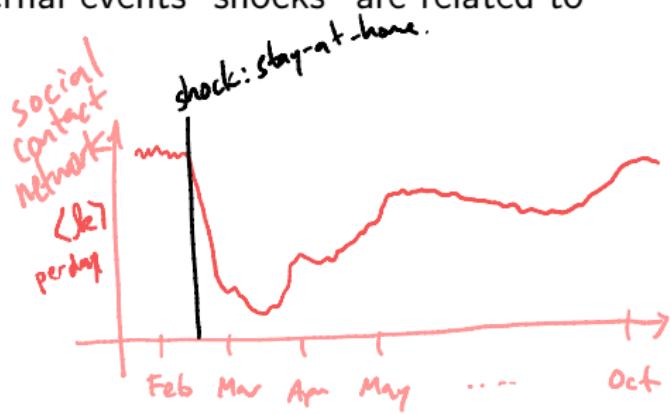
- ▶ the task we're trying to achieve *- the Q we're trying to answer.*
- ▶ the model and algorithm we are using

Motivating Example

Change-point detection (without knowing in advance)

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

Ex: social distancing.



Datasets

Two datasets

Enron

- ▶ Criminal investigation
- ▶ Single network
- ▶ Semi-structured

MIT Reality Mining

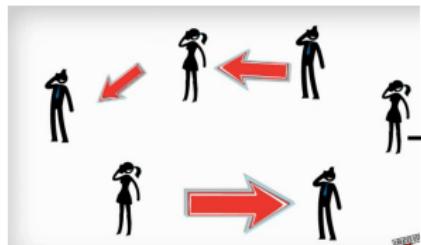
- ▶ Consenting participants
- ▶ Multiple networks
- ▶ Structured data

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

Reality



Real Interactions

Mining

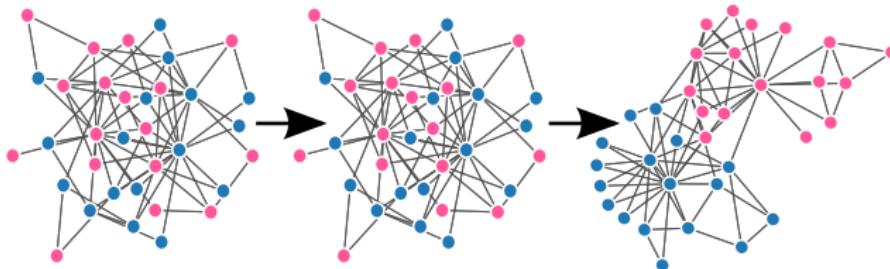


Sensors

Dataset



Data



Network Snapshots

Rich Data

Phone data

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

What did ppl do in 2004?

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?

Friends



Friendship network

Which network?



Friendship network



Voice call network



SMS network

diff.
perspective.

- WhatsApp
- Insta.
- iMessage
- SMS
- email

Which network?



Friendship network



Voice call network



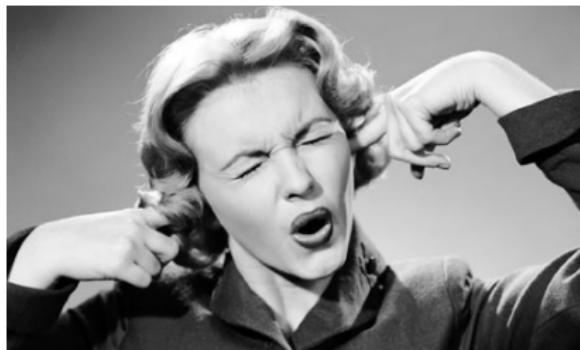
SMS network



Physical proximity network

Noise

The dataset is very noisy.



Sources of noise / missing data:

Noise

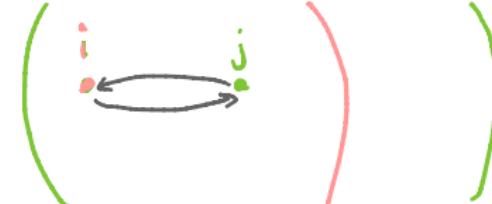
The dataset is very noisy.



Sources of noise / missing data:

- 1) ► phone left at home
- 2) ► no battery (or being charged)
- 3) ► sensor error
- 4) ► date discrepancies (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

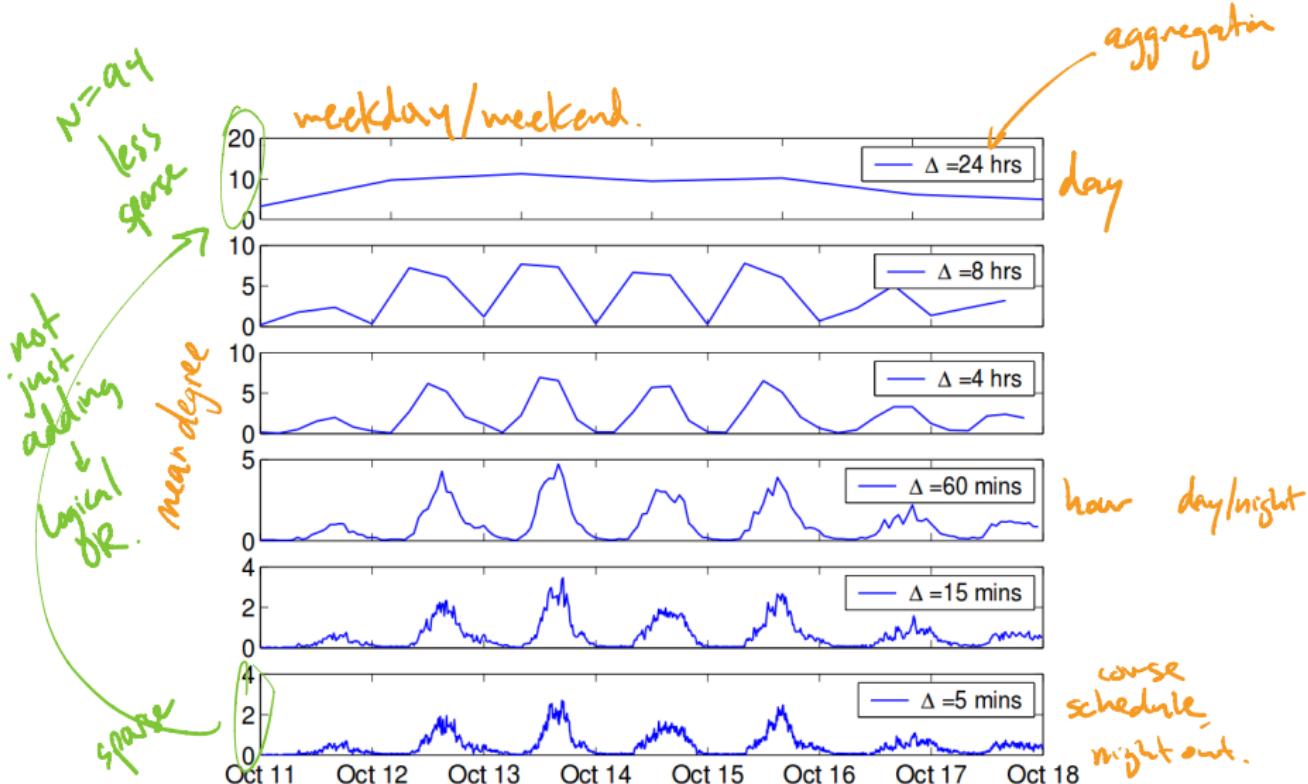
- ▶ Links that only exist in one direction indicate a mismatch between reality and sensor.
- ▶ Two choices: minimal or maximal set. In the minimal set, we discard all unreciprocated edges. In the maximal set, we include all unreciprocated edges.

Temporal resolution

↳ time-varying networks. (later)

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

Temporal resolution $A_{ij}(\Delta) = 1$ if $i \leftrightarrow j$ within Δ .



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?

$$A_{ij} \in \{0, 1\}$$

spurious edges (shouldn't be there)

missing edges (unobserved for some reason).

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)

↳ cases, late at night, phones would be switched off.
→ Hawthorne Effect in action.

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?

Preregistration



Open Science Foundation
OSF

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
- ▶ ~ 0.5 million messages, 1.4GB

Email data

mailedir

Home > Home Downloads enron_m... 20110402 mailedir lay-k

Places

- Recent
- Home
- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- Trash
- Devices
- LBFR Image
- LOS
- GWINRETO...
- DIAGS
- Computer
- Bookmarks
- X-nautilus...
- Network
- Browse N...
- Connect t...

Images

mailedir

Pictures

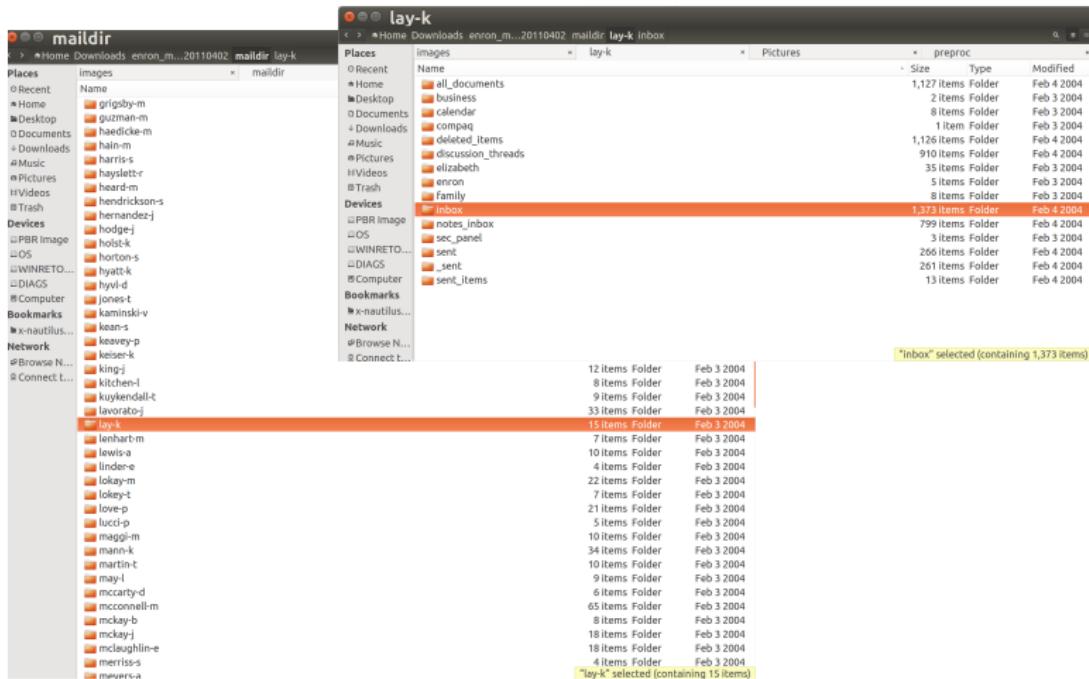
* prepro

	Name	Size	Type	Modified
	grigsby-m	9 items	Folder	Feb 3 2004
	guzman-m	4 items	Folder	Feb 3 2004
	haedrick-m	46 items	Folder	Feb 3 2004
	hain-m	5 items	Folder	Feb 3 2004
	harris-s	2 items	Folder	Feb 3 2004
	haylett-r	20 items	Folder	Feb 3 2004
	heards-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
	holt-k	3 items	Folder	Feb 3 2004
	horford-s	8 items	Folder	Feb 3 2004
	hyatt-k	14 items	Folder	Feb 3 2004
	hyvile	8 items	Folder	Feb 3 2004
	jones-t	10 items	Folder	Feb 3 2004
	kaminski-v	48 items	Folder	Feb 3 2004
	keans-s	187 items	Folder	Feb 3 2004
	keaveny-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-	9 items	Folder	Feb 3 2004
	lavordahl-	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
	lindse-e	4 items	Folder	Feb 3 2004
	lokay-m	22 items	Folder	Feb 3 2004
	lokay-t	7 items	Folder	Feb 3 2004
	love-p	21 items	Folder	Feb 3 2004
	lucci-p	5 items	Folder	Feb 3 2004
	maggio-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
	mcconnell-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
	merriess-s	4 items	Folder	Feb 3 2004
	meurens-a			

"lay-k" selected (containing 15 items)

CEO

Email data



Custom email subfolders (name and depth)!

Email data

maildir

Places Images maildir

- Recent
- Home
- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- Trash
- Devices
- PPR Image
- OS
- WINRETO...
- DIAGS
- Computer
- Bookmarks
- x-nautilus...
- Network
- Browse N...
- Connect t...
- lay-k
- lenhart-m
- lewis-a
- linder-e
- lokay-m
- lokay-t
- love-p
- lucid-p
- maggio-m
- mann-k
- martin-t
- may-l
- mccarthy-d
- mconnell-m
- mckay-b
- mckay-j
- mclaughlin-e
- merriess-s
- muvera-a

lay-k

Places Images lay-k Pictures prepoc

- Recent Name
- Home all_documents
- Desktop business
- Documents calendar
- Downloads compaq
- Music deleted_items
- Pictures discussion_threads
- Videos elizabeth
- Trash enron
- Family
- Devices inbox
- PPR Image notes_inbox
- OS sec_panel
- WINRETO... sent
- DIAGS sent
- Computer sent_items

13. (~/Downloads/enron_mail_20110402/maildir/lay-k/inbox) - gedit

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 <IMCANEOTES+223oe+20Barry+22+28+3Cjarry+4@eyeforenergy+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

mailedir

Places Images maildir

- Recent
- Home
- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- Trash
- Devices
- PPR Image
- OS
- WINRETO...
- DIAGS
- Computer
- Bookmarks
- x-nautilus...
- Network
- Browse N...
- Connect t...
- lay-k
- lenhart-m
- lewis-a
- linder-e
- lokay-m
- lokey-t
- love-p
- lucid-p
- maggio-m
- mann-k
- martin-t
- may-l
- mccarthy-d
- mconnell-m
- mckay-b
- mckay-j
- mclaughlin-e
- merriess-s
- mueller-a

lay-k

Places Images lay-k Pictures preproc

- Recent Name
- Home all_documents
- Desktop business
- Documents calendar
- Downloads compaq
- Music deleted_items
- Pictures discussion_threads
- Videos elizabeth
- Trash enron
- Family
- Devices inbox
- PPR Image notes_inbox
- OS sec_panel
- WINRETO... sent
- DIAGS sent
- Computer Computer
- Bookmarks x-nautilus...
- Network Network
- Browse N...
- Connect t...
- sent_items

13. (~/Downloads/enron_mail_20110402/maildir/lay-k/inbox) - gedit

metadatata

Message-ID: <28975878-1075845189712.JN@sil.evans@thyme>
Date: Tue, 22 May 2001 07:48:16 -0700 (PDT)
From: jbarry@eyeforenergy.com
To: kenneth.lay@enron.com
Subject: EyeEnergy Asia 2001 - Making 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+223oe+20Barry+22+28+3Cjarry+48eye@energy+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

Eyeenergy 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.

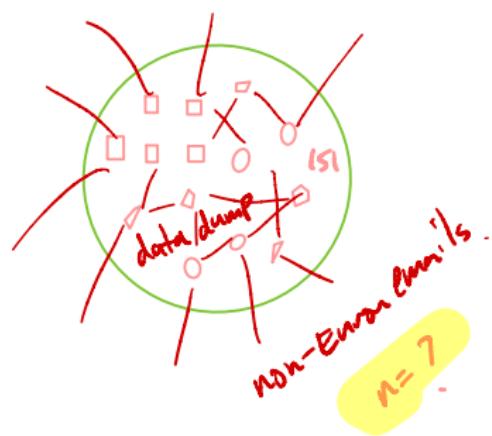
I would like to invite you to give a keynote presentation at the conference. I have attached the full program.

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

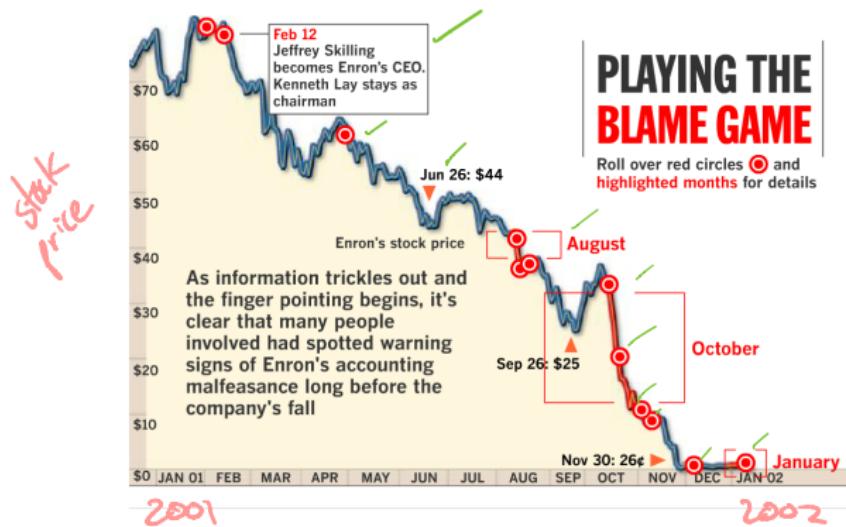
- ▶ Often we use metadata (non-network data) as part of network analysis
- ▶ e.g. Comparing large-scale structure to node level information
- ▶ For change-point detection we are interested in how changes relate to external events

Events

Enron's Collapse

[Digg](#)[Facebook](#)[Twitter](#)[Linkedin](#)[Email](#)

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

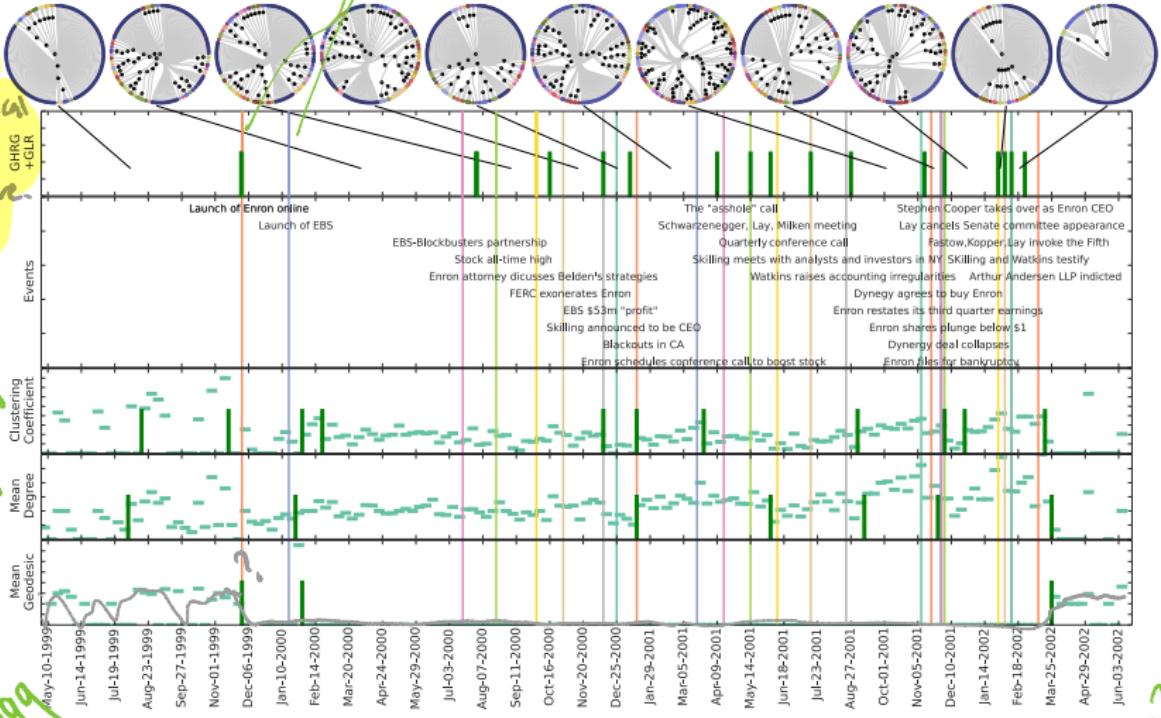


Source: time.com

Detecting change points

model
of
hierarchical
comm.
structure

points
potential charge points



OS-1999
May-

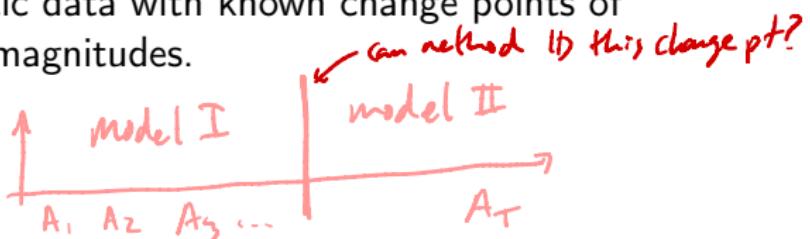
Jun-03-2000

How does the method work?

a tree.
hierarchical SBM.

- ▶ 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.

Synthetic testbed.



How do we know methods work if we only have f. Truth in synthetic data

Resources

Clustering: Art or Science?



v. funny.

Further reading:

Ulrike von Luxburg.
ICML.

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