

Data Wrangling: From Raw Data to Networks

Notes by Dr. Leto Peel
CSCI 5352: Network Analysis and Modeling

October 31, 2017

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:

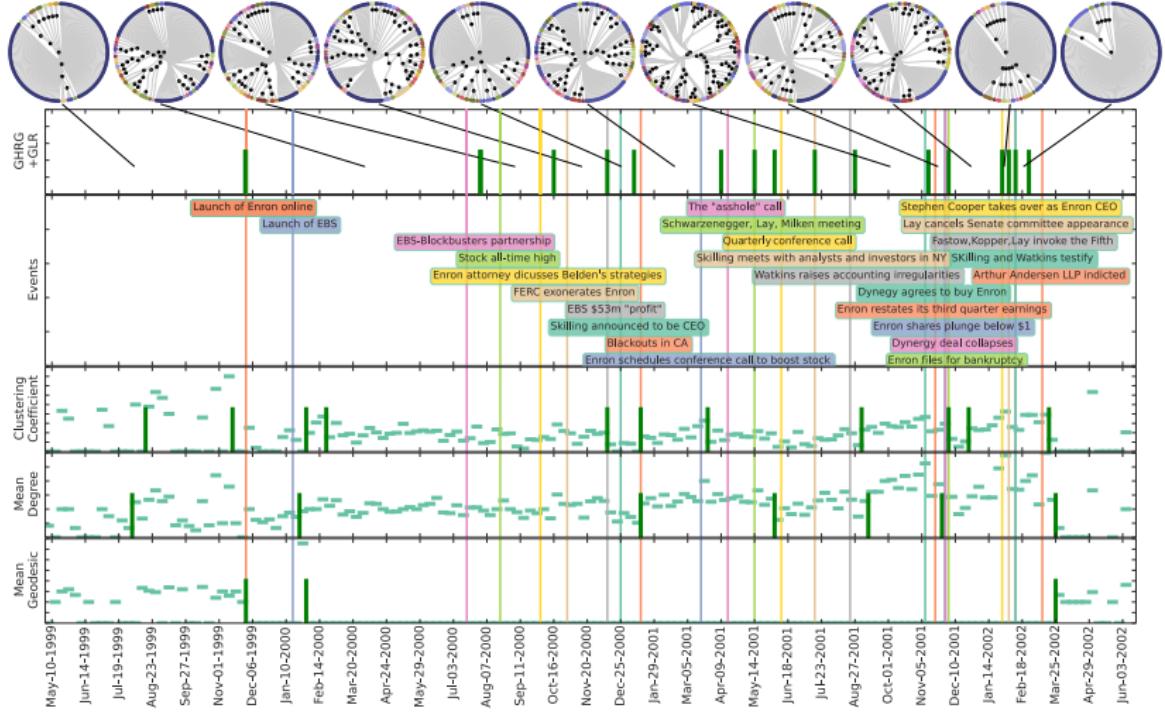
- ▶ the task we're trying to achieve
- ▶ the model and algorithm we are using

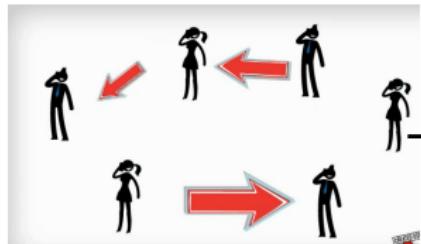
Motivating Example

Change-point detection

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

Detecting change points



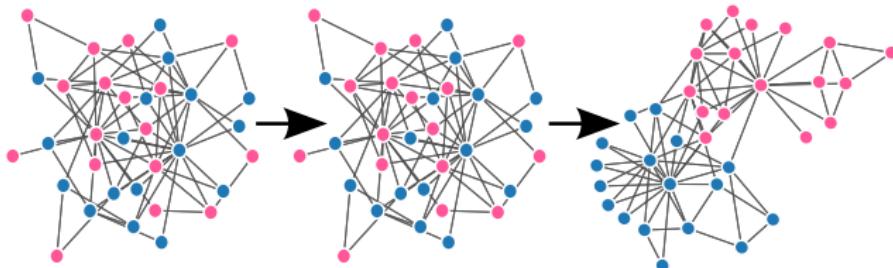


Real Interactions



Sensors

→ Data



Network Snapshots

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

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



SMS network



Voice call 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 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.

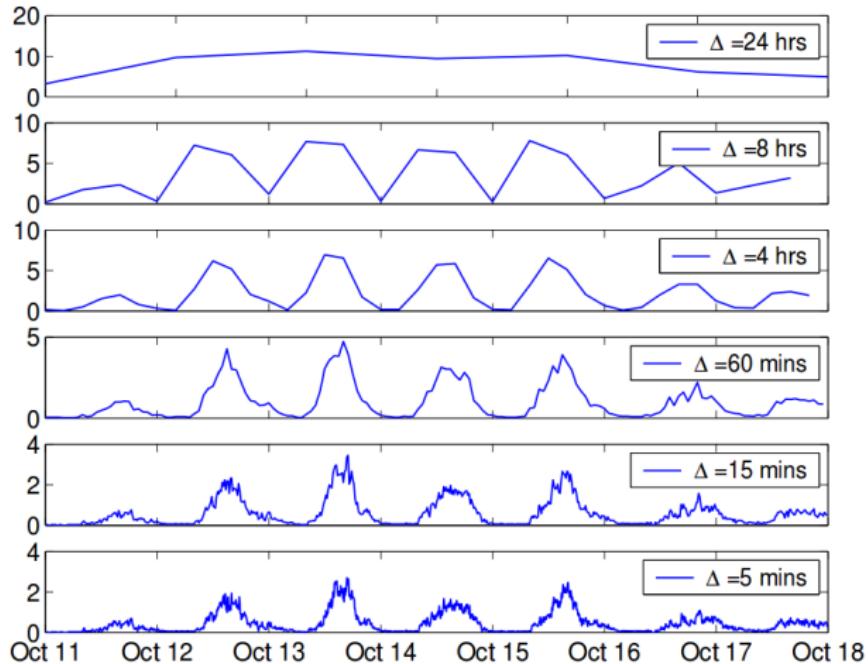
Link reciprocity

- ▶ Links that only exist in one direction indicate a mismatch between reality and sensor.
- ▶ Two choices: minimal or maximal set.

Temporal resolution

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

Temporal resolution



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



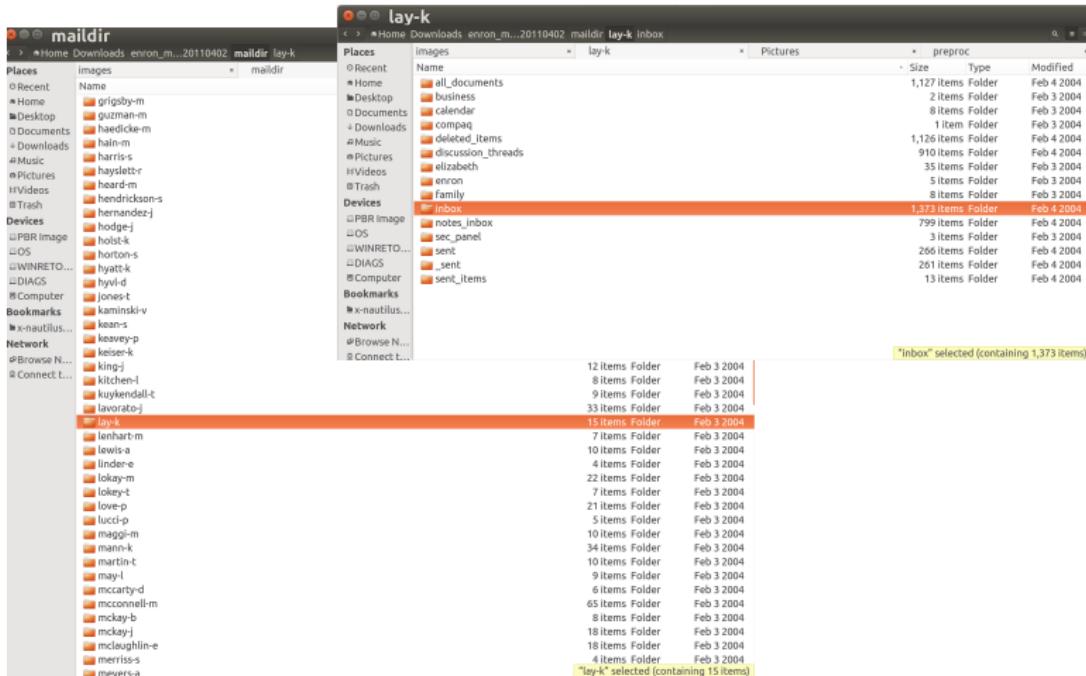
- ▶ Dataset publically released during the FERC investigation
- ▶ 151 Enron employee email accounts
- ▶ ~ 0.5 million messages, 1.4GB

Email data

The screenshot shows a Mac OS X Finder window with a sidebar on the left and a main content area on the right. The sidebar lists various locations such as Recent, Home, Desktop, Documents, Downloads, Music, Pictures, Videos, Trash, Devices, and Bookmarks. The main content area displays a hierarchical list of folders under 'maildir' and 'lay-k'. The 'maildir' folder contains subfolders for names like grigsby-m, guzman-m, haedick-m, hain-m, harris-s, haylett-r, heard-m, hendrickson-s, hernandez-j, hodgej, hoist-k, horton-s, hyatt-k, hyvi-d, jones-t, kaminski-v, keans-s, keavy-p, keiser-k, kingj, kitchen-l, kuykendall-t, lavorato-j, and lay-k. The 'lay-k' folder is highlighted with a red border and contains subfolders for lenhart-m, lewis-a, linden-e, lokay-m, lokay-t, love-p, lucid-p, maggi-m, mann-k, martin-t, may-l, mccarty-d, mcconnell-m, mckay-b, mckay-j, mclaughlin-e, merriis-s, and meyerc-a. A tooltip at the bottom right of the 'lay-k' folder indicates it contains 15 items.

Places	Images	maildir	Pictures	* prepoc
Recent	Name			
Home	grigsby-m			9 items Folder Feb 3 2004
Desktop	guzman-m			4 items Folder Feb 3 2004
Documents	haedick-m			46 items Folder Feb 3 2004
Downloads	hain-m			5 items Folder Feb 3 2004
Music	harris-s			2 items Folder Feb 3 2004
Pictures	haylett-r			20 items Folder Feb 3 2004
Videos	heard-m			6 items Folder Feb 3 2004
Trash	hendrickson-s			9 items Folder Feb 3 2004
Devices	hernandez-j			14 items Folder Feb 3 2004
Image	hodgej			10 items Folder Feb 3 2004
OS	hoist-k			3 items Folder Feb 3 2004
WINRETO...	horton-s			8 items Folder Feb 3 2004
DIGITS	hyatt-k			14 items Folder Feb 3 2004
Computer	hyvi-d			6 items Folder Feb 3 2004
Bookmarks	jones-t			10 items Folder Feb 3 2004
x-nautilus...	kaminski-v			48 items Folder Feb 3 2004
Network	keans-s			187 items Folder Feb 3 2004
Browse N...	keavy-p			10 items Folder Feb 3 2004
Connect L...	keiser-k			5 items Folder Feb 3 2004
lay-k	kingj			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
	linden-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
	lucid-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
	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
	merriis-s			4 items Folder Feb 3 2004
	meyerc-a			"lay-k" selected (containing 15 items)

Email data



Custom email subfolders (name and depth)!

Email data

lay-k

Places Images lay-k Pictures prepoc

Recent Name all_documents 1,127 items Folder Feb 4 2004
griggsby-m business 2 items Folder Feb 3 2004
guzman-m calendar 8 items Folder Feb 3 2004
haedickle-m compaq 1 item Folder Feb 3 2004
hain-m deleted_items 1,126 items Folder Feb 4 2004
harris-s discussion_threads 910 items Folder Feb 4 2004
haylett-r elizabeth 35 items Folder Feb 3 2004
heard-m enron 5 items Folder Feb 3 2004
hendrickson-s family 8 items Folder Feb 3 2004
hernandez-j hodgej 1373 items Folder Feb 4 2004
host-k hortons 799 items Folder Feb 4 2004
horton-s hyatt-k inbox
hyvi-d jones-t kaminsky-v keans-s keavney-p keiser-k king-j kitchen-l kuykendall-t lavorato-j lay-k lenhart-m lewis-a linden-e lokay-m lokey-t love-p lucid-p maggi-m mann-k martin-t may-l mccarty-d mcconnell-m mckay-b mckay-j mclaughlin-e meriss-s meureca

Devices PBR Image OS WINRETO... DIAGS Computer Books sent_items

Bookmarks x-nautilus...

Network #Browse N... #Connect t...

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
X-From: "Barry J" Barry <jbarry@eyeforenergy.com@ENRON<INCEANOTES+223Joe+20Barry+22+20+3Cjbarry+40eyeforenergy+2E+46ENRON@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

lay-k

Places Images Name Pictures preproc

Recent Name all_documents business calendar compaq deleted_items discussion_threads elizabeth enron Family inbox notes_inbox sec_panel sent _sent sent_items

Devices PBR Image OS WINRETO... DIAGS Computer

Bookmarks nautilus...

Network Browse N... Connect t...

lay-k lenhart-m lewis-a linden-e lokay-m lokey-t love-p lucid-p maggi-m mann-k martin-t may-l mccarty-d mcconnell-m mckay-b mckay-j mclaughlin-e meriss-s meure-a

1,127 items Folder 2 items Folder 8 items Folder 1 item Folder 1,126 items Folder 910 items Folder 35 items Folder 5 items Folder 8 items Folder 1373 items Folder 799 items Folder

Modified Feb 4 2004 Feb 3 2004 Feb 3 2004 Feb 3 2004 Feb 4 2004 Feb 4 2004 Feb 3 2004 Feb 3 2004 Feb 4 2004 Feb 4 2004 Feb 3 2004 Feb 3 2004 Feb 3 2004 Feb 4 2004

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

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-Mailer: "Eve Barry <jbarry@eyeforenergy.com>@ENRON <IMCANEOTES+223Joe+20Barry+22+20+3Cjbarry+40eye@forenergy+2E+46ENRON@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

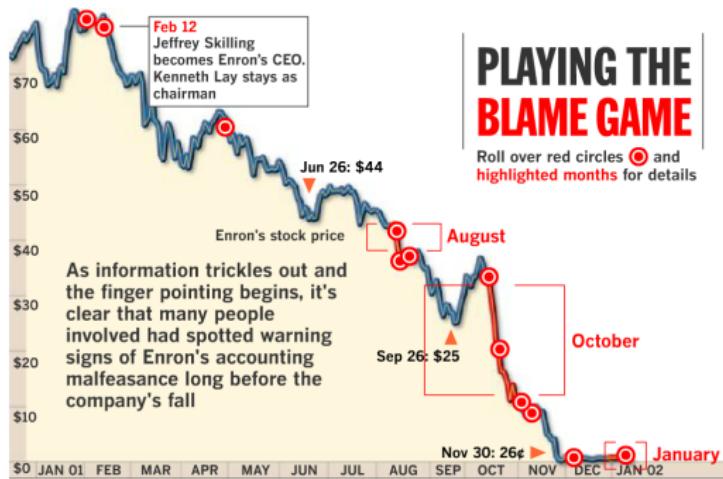
 Facebook

 Twitter

 LinkedIn

 Email

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



Source: time.com

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: <http://tinyurl.com/letopeel/datasets.html>

Software:

- ▶ Python
- ▶ Matlab
- ▶ Octave