

# Data Wrangling: From Raw Data to Networks

CSCI 5352: Network Analysis and Modeling

Notes written by Dr. Leto Peel  
Université Catholique de Louvain

October 28, 2019

## 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

# Datasets

## Two datasets

### Enron

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

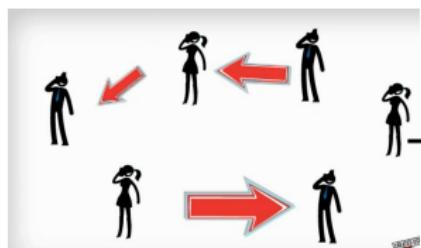
### MIT Reality Mining

- ▶ 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



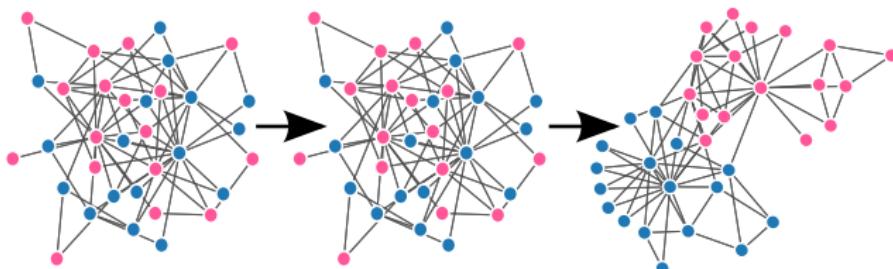
Real Interactions



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 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.

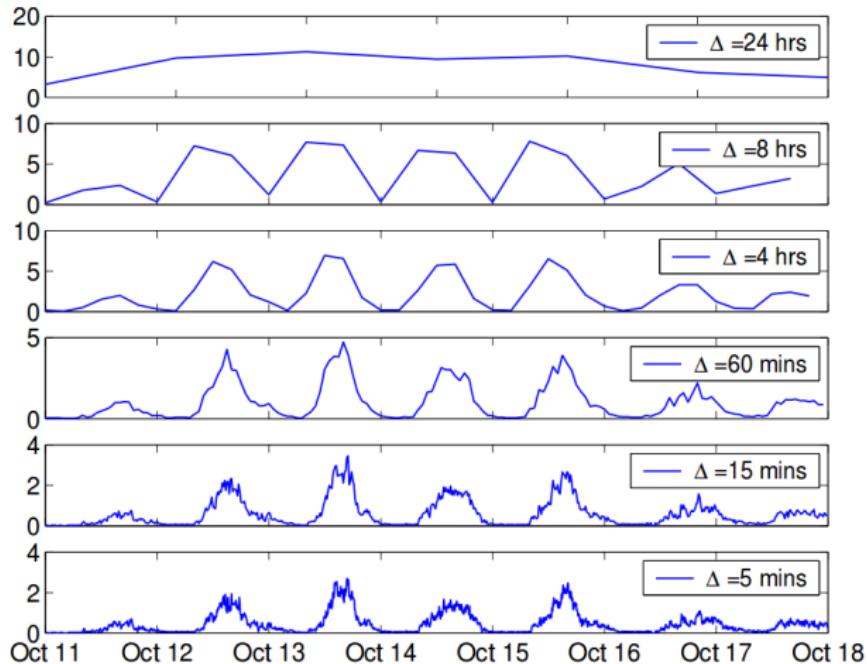
## Link reciprocity

- ▶ 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

- ▶ 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



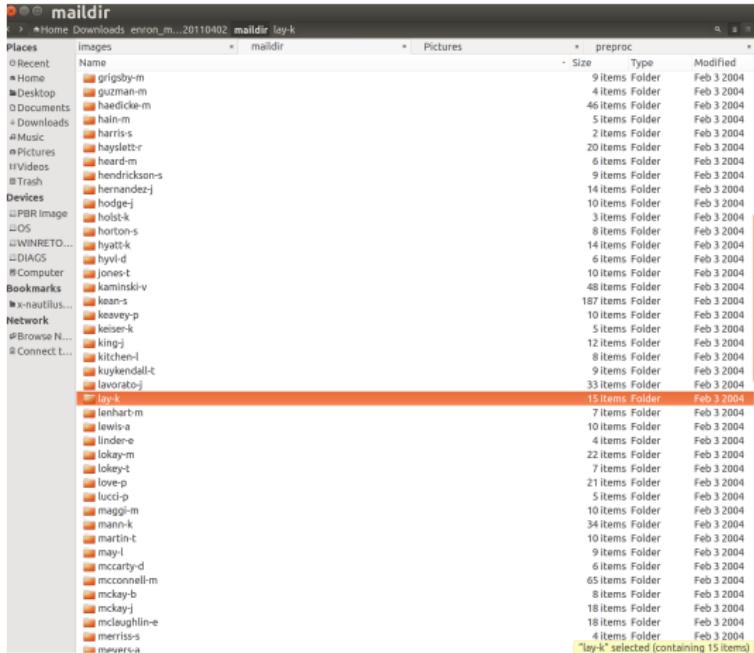
- ▶ Largest supplier of natural gas to North America
- ▶ “America’s Most Innovative Company” by the magazine Fortune from 1996 to 2001
- ▶ Misrepresentation of earnings and unethical practises
- ▶ End of 2001: One of the largest bankruptcies in American history

# Enron email dataset

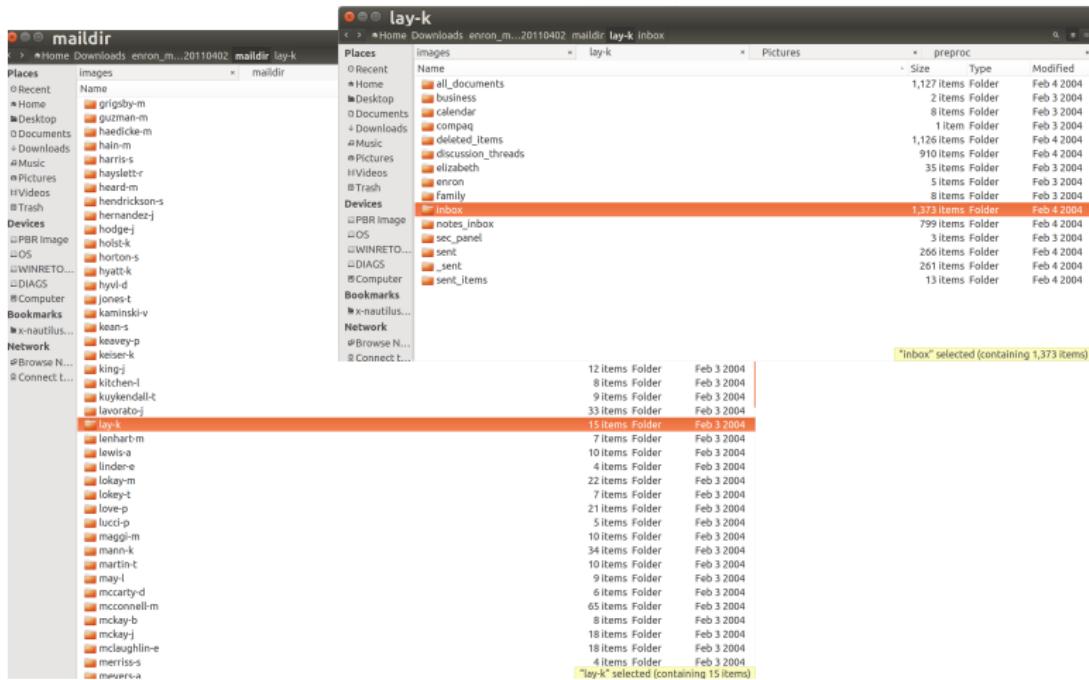


- ▶ Dataset publically released during the FERC investigation
- ▶ 151 Enron employee email accounts
- ▶ ~ 0.5 million messages, 1.4GB

## Email data

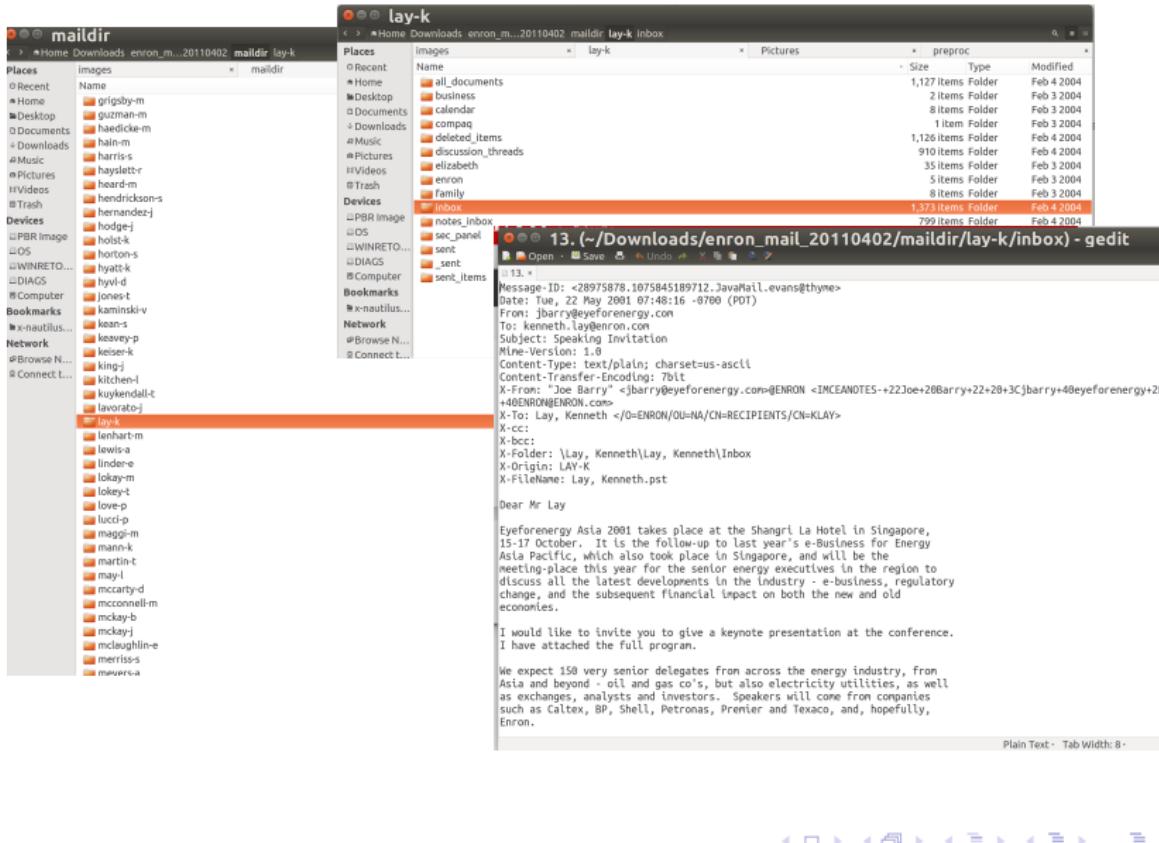


# Email data

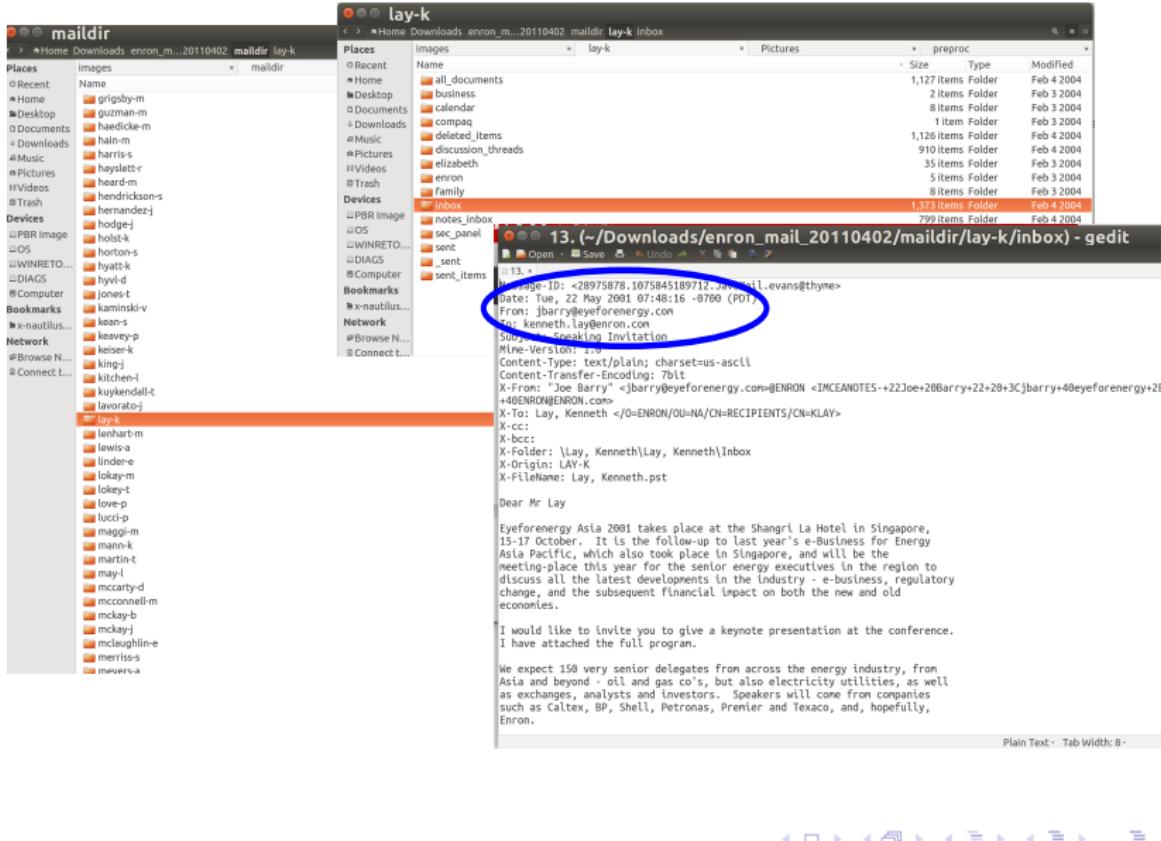


Custom email subfolders (name and depth)!

## Email data



## Email data



## 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

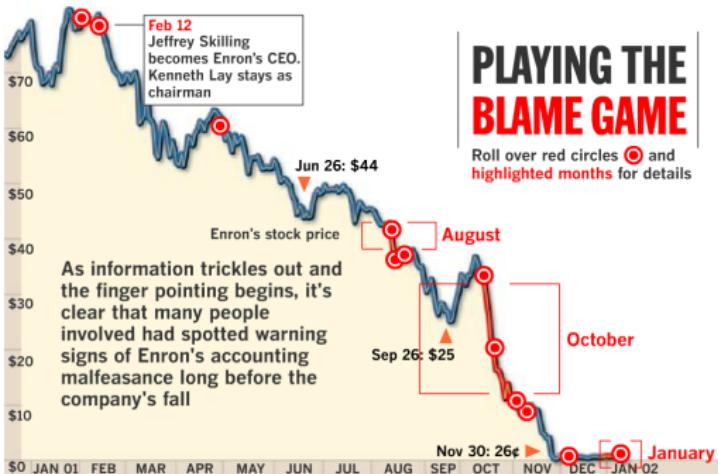
- ▶ Often we use metadata (non-network data) as part of network analysis
- ▶ e.g. Comparing large-scale structure to node level information
- ▶ For change-point detection we are interested in how changes relate to external events

# Events

## Enron's Collapse

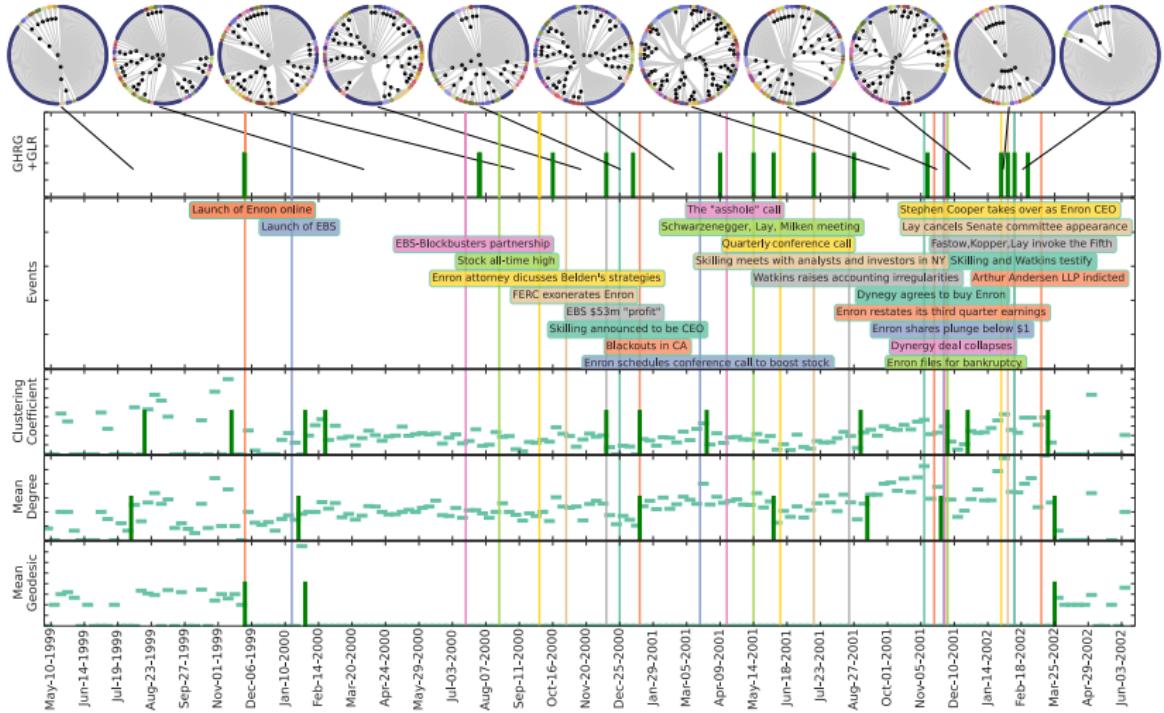
[Digg](#)[Facebook](#)[Twitter](#)[Linkedin](#)[Email](#)

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