Data Wrangling: From Raw Data to Networks

2014.10.14  Dr. Leto Peel
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

Physical proximity network

SMS network
Noise

The dataset is very noisy.

Sources of noise / missing data:
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 a 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 publicly released during the FERC investigation
- 151 Enron employee email accounts
- $\sim 0.5$ million messages, 1.4GB
Email data
Email data

Custom email subfolders (name and depth)!
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.
Dear 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.
To build a network we need to identify the nodes (i.e. the email addresses)
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
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

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