The Search for Temporal Expansive Scale-Free Network Data

14th May, 2014

Introduction

This is a collection of various data sources that were investigated during the exploration for scale-free, temporal (/longitudinal/dynamic/time) network data.

Note: The ideal format during this exploration was GEXF (or a similar variant). Hence, to acquire appropriate data, it may have to be scrapped, accessed through Apis, and/or transformed between different formats.

Note: As of 23rd May. SFN is now a soft requirement. Data should still be locally sparse and locally dense though.

Data Sources

Arpil'14 early explorations.

Both KEDs and Terror Data were not evidently SFN.

KEDS

Terror Data (911)

Arpil'14 explorations.

Marvel Universe

Not as culturally relatable as possible. Program has been mostly set up though for it though.**Transcripts** (Tv & Movie)

A program was written to turn transcripts into temporal networks.
However, the weren't evident SFN structures in the transcripts tested and probably wouldn't be due to the ego-focused nature of shows.

Citations

Not as culturally relatable as possible. Requires network generation.

POK – protectors of Kronos

Appears like it would've been SFN, however not culturally relatable.

Machine Learning Sources exploration.

ML data sources were explored and many were turned down due to the time required to turn into a temporal network (let alone Sfn). Also, most were culturally unrelatable, or required additional study of domain knowledge for visualisations to hold significance. Most of the below could've been turned into temporal networks.

Sites

- https://www.hackerrank.com/
- https://www.kaggle.com
- http://www.infochimps.com/datasets

Kaggle had a range of potential datasources:
The top candidate (as of this writing) for the requirements was
> https://www.kaggle.com/c/facebook-ii/data <- bit massive, also not real labels (so not culturally relatable)

Second-tier candidates
- https://www.kaggle.com/c/dsg-hackathon/data?SiteLocations_with_more_sites.csv
- https://www.kaggle.com/c/stayalert/data
- https://www.kaggle.com/c/event-recommendation-engine-challenge/data (big data)

Third-tier candidates
~ https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data?testHistory.csv.gz
~ https://inclass.kaggle.com/c/aalto-music-listening-prediction/data
~ https://www.kaggle.com/c/yelp-recruiting/data
~ https://www.kaggle.com/c/WhatDoYouKnow/data
~ https://www.kaggle.com/c/dunnhumbychallenge/data

Movies

Movies in general are culturally relatable and can be formed into bipartite co-star temporal networks. They also appear to have various sources. Of all the datasets explored, movies best fit the criteria of potential scale-free networks, expansive sizes, culturally relevant, and temporal.

Movie / stars / tv-show data could be acquired from a variety of resources:

Wikipedia Tables – These could be copied directly (with minor manipulation), or accessed via the wikipedia media api.

Rotten Tomatoes – Provides a relatively straightforward api (this api however may only be permissible within the U.S.).

IMDB – A great resource. The one with the most sources of access of the sources explored in this list.

IMDB GD05 - A large (25mb) bipartite network of actors and movies
http://www3.ul.ie/gd2005/dataset.html

Freebase Dataset - A dataset of American Films (with actors)
http://graphlab.com/learn/notebooks/graph_analytics_movies.html

IMDB Lists – Such as 'Genre: Sci-Fi, 1500 Titles: 1970-2013' - (http://www.imdb.com/list/ls051626175/) can be filtered as desired and one needs only to run a scrapper to get the desired data (which can then be transformed into the appropriate temporal network format)

May'14 general exploration.

UCI network data repository

RFID data at a conference – http://networkdata.ics.uci.edu/data.php?id=110
(This was the sole temporal/dynamic/longitudinal data on the site)

UCI, Linton Freeman's datasets
http://moreno.ss.uci.edu/data.html

The following were found through searches for temporal / dynamic / longitudinal / time data. These datasets require transformation.

NEWCOMB,NORDLIE--FRATERNITY
Has potential.

SAMPSON--MONASTERY
Data may not have been that interesting / revealing.

VAN DE BUNT--DUTCH COLLEGE FRESHMEN
This one may have been interesting, it has been cited a few times now. However, it was ruled out due to the relatively small size of the network.

Nicholas Christakis' Data

http://www.nicholaschristakis.net/pages/research/r-interests.html
Nicholas Christakis' smoking, emotion, health etc. networks seem to be tied and hidden behind industry partners.

Stanford; SoNIA (Social Network Image Animator); Some Longitudinal Network Data Sources
http://www.stanford.edu/group/sonia/dataSources/

Data sets here require more digging, transforming, and filtering. Some quite old. Probability of scale free was minimal. Took an educated guess and skipped further exploration.

Cell Phone Data – http://realitycommons.media.mit.edu/index.html
Promising Cell phone data, does require quite a bit of transforming. It is fairly large data.

UCI KDD (Knol Discovery Db) Archive – http://kdd.ics.uci.edu/summary.data.type.html (promising)

EEG Data – http://kdd.ics.uci.edu/databases/eeg/eeg.html

Data may not have been that culturally interesting.

Other data sets were framed moreso as experimental results (sign language) rather than social / culturally relatable data (which we were searching for at the time).

May'14 deco data.

BOSCAR - NSW crime statistics - proxemics network possible, may be scale free, further investigation required to make sure we have latitude longitude details (lgas are certainly available). Temporal granularity may be available (yearly is certainly available).

House prices - NSW - this data can be fairly easily mined.

Twitter

It is also possible to get some interesting Twitter temporal data (that can be turned into a network of words, similar to the terror dataset). However, the probability of a SFN structure is just as likely as the IMDB and terror datasets (potentially nonexistent). Certain topics over time may be particularly interesting, e.g. #auspol .

Note

Some early data sources were also explored but not mentioned here as they were predominantly not temporal network data.

Scrapping tools

Below is a list of various scraping tools. They offer varying levels of ease, some have GUIs, the largest differentiator is if it can support pagination (scraping multiple pages).

- https://import.io
  - - can scrape multiple urls
- https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn
  - - can apparently scrape mutliple pages
- beautiful soup
  - - python
- http://open.dapper.net/ ~ pagination missing
- http://www.kimonolabs.com/welcome.html ~ pagiation missing
- Style sheet browser extensions could also be applied to pages and visited and copied.

Conclusions

In searching for network data that was expansive, temporal, and potentially scale-free. Of the datasets explored, Movie data appeared to fit the criteria best, with IMDB lists being the most flexible.