SYSM 6302 Project

A major component of 6302 is a semester-long project in which you will create and analyze a real world network. You will work in groups of up to 2 of your choosing. You should build a *new* network - this can be an update of a network someone has already made (e.g., movie-actor network, but for different years or different focus - Bollywood?). I suggest you pick something you are interested in - whether it is biology, transportation, social, sports, music, etc.

Requirements:

- Networks may be undirected or directed, but not a bipartite graph.
- Network size (number of nodes) must be greater than 200 (exceptions can be made, but if you make a smaller network, it has to be really well done and have nice analysis). Ideally, although depending on the context, even bigger is better. The idea is not to collect these by hand, but instead use scripts to parse the data.
- Networks should be unique. Although two groups may choose to build networks of a particular type (e.g., musicians that have gone on concert with each other), they should sample a different part of this network type (e.g., focus on different music genre or decade). There are a handful of publicly available datasets - try not to exactly reproduce these either. Check here:
  http://snap.stanford.edu/
  http://pajek.imfm.si/doku.php?id=data:pajek:vlado
  http://networkdata.ics.uci.edu/index.html
- Networks are almost always incompletely sampled, so be clear with what and how you are sampling them. For example, do not construct the network of *rock* musicians who had a concert together. Instead focus on the network of *rock* musicians who had a concert together *and were featured on the top 100 billboard during 2000 and 2010* - this limits the scope and makes it clear that you are just finding connections between the musicians who fall under this description.

Deliverables:

1. **Proposal** (1 page maximum) that contains a description of your planned network including clearly defining what a node and what an edge are, where you will get the data, and how you will collect and compile the data. Be specific! If you change your network idea after you make your proposal, then you need to make and submit another proposal.
2. **Zip file** containing a gml network data file; a text file with a description of your network (see below) and how it was obtained; and any original source files and scripts you used to curate the data. Where relevant, node and edge link names should be included in the gml representation.
3. **Presentation** (10-15 minutes) in the last week of class and corresponding PDF of your slides and any other media you use. You may also submit a supplementary document if you want to include background analysis for your slides/presentation. This is limited to 2 pages of text and any number of figures.

Grading:

- **Proposal: 5%** Your written proposal is submitted on time and contains all the relevant information.
- **Network: 45%** You have a network that adheres to the requirements. The gml file works and the network is correct. The summary text file is descriptive and includes all the necessary information to understand how the network was constructed. Source code used to generate the data is included.
- **Presentation: 40%** The presentation is clear and details are explained well. Media (slides/videos/images) augment the presentation. Presentation is no longer than the allotted time. Attendance to both presentation sessions is mandatory, otherwise you will lose points.
- **Awesomeness: 10%** Going above and beyond, looking at something creative, doing a thought provoking analysis.

Timeline:

- By Class 10 (Feb 19): Find a partner (optional).
- By Class 12 (Feb 26): Communicate a preliminary topic using the Google sign-up form.
- By Class 14 (Mar 5): Submit proposal via eLearning.
- By Class 20 (Apr 2): Show that you have built your network (you have a fully functional GML file)
- Class 27 & 28 (May 3-7): Presentations.
- May 9: Slides (PDF), optional supplementary material (PDF) and zip-file due (through eLearning).

When describing and analyzing your network consider some of the following concepts (essentially everything we studied):

- why did you chose the type of network you did (e.g., weighted, directed)?
- centrality measures
- diameter, paths
- components, clustering coefficients, modularity and community detection
- degree distribution, power law
- comparison to random network models
- dynamics
- **really important:** focus on the interpretation of your analysis not just the statistics!

Examples:

- co-appearance of characters in a movie/book/play (e.g., Star Wars, Hamlet)
- transportation network (e.g. bus, road, train networks)
- social networks
- artist-music networks
- concept map of a course

Example Summary File:

Authors:
Justin Ruths

Description:
Lord of the Rings co-appearance (undirected, weighted) network, in which characters are connected if they are mentioned in the same chapter. A full list of characters was downloaded from http://lotr.wikia.com/wiki/List_of_characters. A bipartite graph was created for characters mapping to chapters (of the books, not the movie), then a one-mode projection was taken to get the undirected network of characters, weighted by the number of chapters they co-appeared in. Isolated nodes (characters) were dropped.

Nodes: 163
Edges: 3636
Clustering: 0.57
Diameter: 4

Technical Details:
For each character we used the name that was likely to appear the most often for that character (e.g., bilbo for Bilbo Baggins). They may have led to a few false positives in finding the characters throughout the book.

Data Source:
Character list: http://lotr.wikia.com/wiki/List_of_characters
Book text as PDF.

Data Source Author/Organization:
None