BIA-658A

It is the time for you to start working on your project. On Oct 24 please form groups with 3-4 students, start with a group meeting, and make sure that you mark the calendar with the timeline of deliverables (at the end of this assignment).

Meeting Guidelines

Have a meeting (in person or online) with your group members, and discuss:

1. The skillsets of each member, and divide the task of data cleaning / analysis / writing / poster making. for example:
   a. Proficiency of R or other programming/analytics languages
   b. Taking/Have taken Web Analytics and knows how to work with APIs provided by Social Media companies

   Note: Do not underestimate the time of data cleaning. If you have a messy dataset, at least 50% of the time will be spent on data cleaning. The workload can be and should be adjusted dynamically as the project progresses. If a member is spending too much time on a certain task, it is the group's responsibility to share the task.
2. Set-up a shared cloud drive (Dropbox/Google Drive/OneDrive) folder and/or a Github repository for the group.
3. The types of project you wish to do. Refer to the slides for the first class. You do not need to have a consensus now (you have until 11/7 to decide), but a brainstorming session would help.
4. If you decide to work on an empirical project:
   Do you decide to crawl/collect your own data? - If yes, start writing crawler now
   Do you wish to use public available data? - If yes, you could get the data from

   Select one or more datasets from this list (I will post an updated list on Canvas)
   http://snap.stanford.edu/data/
   https://networkdata.ics.uci.edu/index.html
   http://webscope.sandbox.yahoo.com/catalog.php?datatype=g
   http://www-personal.umich.edu/~mejn/netdata/
   http://socialcomputing.asu.edu/pages/datasets
   https://github.com/gephi/gephi/wiki/Datasets
   http://toreopsahl.com/datasets/
   http://www.casos.cs.cmu.edu/computational_tools/data2.php
   http://deim.urv.cat/~alexandre.arenas/data/welcome.htm
   http://www3.nd.edu/~networks/resources.htm
   http://vlado.fmf.uni-lj.si/pub/networks/pajek/data/gphs.htm
   http://www.crowdflower.com/data-for-everyone
   NYC/Hoboken/Jersey City Open Data https://nycopendata.socrata.com
   http://www.hobokennj.org/open/
   http://data.jerseycitynj.gov/
   European Open Data
   http://open-data.europa.eu/en/
   US Open Data
   http://www.data.gov/
   Yahoo Webscope Program
   https://webscope.sandbox.yahoo.com/
   Yelp Dataset Challenge
   https://www.yelp.com/dataset_challenge
   Open data challenges
   http://opendatachallenges.org/

GitHub Archive
http://www.githubarchive.org
This Github page contains an updating directory of publicly available datasets:
https://github.com/caesar0301/awesome-public-datasets
Active and completed Kaggle competition. https://www.kaggle.com/competitions (Some of the datasets are not network by nature, but can be analyzed if you construct a meaningful network.)
Search google for ___ data dump. For example, reddit data dump, stackoverflow data dump

Sample project frameworks include:

i. Different nodes with different network properties will have different outcomes (for example: Clusters, Networks, and Firm Innovativeness, SMJ 2005.pdf, and Predicting Individual Behavior with Social Networks, Mkt Sci 2014.pdf, Community Size and Network Closure, AER 2007). For this type of study make sure that your dataset contains node attributes/outcome variables, or you are able to merge the network dataset with other publicly available datasets.
ii. A through descriptive analysis of one or more networks, and the implications of the findings (for example: The structure of scientific collaboration networks)

5. If you decide to work on an algorithm project, it is challenging for you to develop a novel algorithm given the time. Rather, you can study in detail and compare several algorithms that are not discussed in class using an empirical dataset. For example, comparison of community detection/structural equivalence algorithms. Many authors provide source code on Github or R packages. Questions can be asked include: What are the pros/cons of the algorithms according to the authors? Do they work as intended? What are the differences in terms of outcome and speed?
6. Class/work schedule for the rest of the semester, and other commitments that may hinder progress or regular group meetings (traveling etc.).

Don't be too ambitious, work on a simple project and try to find something interesting out of it. Do it carefully, present it clearly (oral and writing) -- it will take longer than you think.

Timeline of Deliverables
10/24: As a HW assignment, submit a statement on Canvas to indicate that you had the meeting.
11/7: A brief, 1-2 paragraph proposal on what you want to analyze.
11/7: in class: Representatives of each group (does not have to be group leader) give a 1 to 2-minute "elevator pitch" to the class on why they select the project and what might be the implications of the project
11/28, 12/5: Project presentation.
12/15: Project report (5-12 pages 1.5 line space) due.