Social Network Analysis – Winter 2022
Lab 1b: Descriptive Network Analysis – Local and Global Properties

Deadline: Friday, February 11th at 11:59 pm

The purpose of this lab is to learn how to conduct descriptive network analysis using the statistical software package R. Using the "vosonSML" package, this assignment will make use of a data set you collect from one of three social media platforms (i.e., Twitter, YouTube, and Reddit) by defining either hashtags (e.g., #MeTooMovement) or urls (e.g., a metoo movement thread on Reddit and YouTube). A network is generated from the interactions (e.g., @mention and co-commenting) between users/actors included in the same hashtags or discussion thread. For example, Twitter user A replies to/retweets a tweet by B. It creates a link from A to B. On Reddit or YouTube, user X starts a thread and Y comments on the thread. Then, the interaction creates a link from Y to X. You will be visualizing and interpreting individual and global network properties of this type of network.

We will cover this material in lab section on **January 21st**. However, feel free to work ahead, and reach out to Brennan if you have any other questions. Thanks!

General Instructions:

1. This lab has two parts indicated with Roman numerals (**III,IV**) in the outline below.
2. Prepare a report that includes your responses to **all the questions** for all four parts outlined below. Label your responses with the instruction and prompt number (for example, "9"). Incorrectly labeled responses may receive a lower grade. For each response in your report, you should report your results and interpret them as specified in the prompt. Insert network images into your report in the appropriate places. In RStudio, you can click "Export / Copy to Clipboard" and paste directly into the Word document. You will be graded primarily on the completeness and accuracy of your responses, but the clarity of the prepared report will also affect your grade. While students may work together to perform the analysis, each student must execute his or her own code, and is responsible for writing the narrative in the report and submitting it.
3. Upload your report as a PDF, R code script and RData file to the Lab 1 Assignment in Canvas by **Friday October 15th, at 11:59 pm..**
4. Please delete the instructions from your final hand-in.

PART III: Individual Network Properties (25 points)

In this part, **you will continue using the same dataset that you gathered in lab 1a.** You should be able to reload the exact same data using the RData file you generated in lab 1a. You will compute individual-level network measures and identify some key users in your network. Further, you will conclude this lab exercise with discussion in your main findings based on the visualizations, measures and your own analysis.

Individual Network Properties Instructions:

1. Generate the giant component graph (only the single largest component) for your network from Lab 1a. **You will use the giant component graph, rather than the full network, for all analysis in this lab**.
2. For each node in your network, calculate different centrality measures: (a) in-degree, (b) out-degree, (c) betweenness, (d) in-closeness, (e) out-closeness, (f) eigenvector, (g) Burt's network constraint, (h) hub score, and (i) authority score.

Individual network properties questions to answer in your assignment:

1. **(10 points)** Provide a table ranking the top 5 nodes in your network on each centrality measure. Each centrality means (a) in-degree, (b) out-degree, (c) betweenness, (d) in-closeness, (e) out-closeness, (f) eigenvector, (g) Burt's network constraint, (h) hub score, and (i) authority score.
2. **(10 points)** Briefly describe each centrality measure. How is each computed and what does its number mean in your network (e.g., a high centrality score means…)?
3. **(5 points)** How does the centrality of nodes vary with different types of centrality metrics? Why is this the case? Please offer some potential explanations using certain nodes as examples.

PART IV: Global Network Properties (40 points)

In this part, you will identify global network structures of your network such as subgroups within a network provides much information to social network researchers, and a variety of algorithms have been developed to identify and measure subgroups. You will use some of igraph's built-in tools to identify subgroups and central nodes for visual inspection.

Global Network Properties Instructions:

1. Calculate the coreness of each node in the giant component graph. Plot the graph and color nodes based on their coreness.
2. Run a community detection algorithm for the graph. If you want to use another algorithm, replace with your choice. Check how many communities are created. Calculate the modularity score using these communities. Plot the graph using the community detection results. Make sure that your nodes are colored based on the communities.
3. Create a plot for the in-degree distribution of the graph. Create a log-log plot based on the in-degree distribution. Calculate a power law fit to the in-degree distribution. Hint: use 'power.law.fit()' in 'igraph'.
4. Create a plot for the out-degree distribution of the graph. Create a log-log plot based on the out-degree distribution. Calculate a power law fit to the out-degree distribution.
5. Compute the clustering coefficient and the average path length for the graph. Also, compute the clustering coefficient and average path length for 1,000 randomly reshuffled networks based on the graph. Plot the distribution of 1,000 simulated clustering coefficient values from the reshuffled networks and add the vertical line on the plot indicating the value of average path length from the graph. Create the same plot for average path length.
6. Run a one-tail t-test to examine whether the value of clustering coefficient from the graph is different from the simulated distribution. Run the same t-test for average path length.
7. For the final question (part IV question 10), look at "twitterData," "youtubeData," or "redditData" depending on your choice of the social media platforms. These data contain more detail information than you've so far been using in this lab. See if any of the information (e.g. usernames, comments) or additional descriptive analysis might be helpful in interpreting your network.

Global network properties questions to answer in your assignment:

1. **(3 points)** Briefly describe (a) what k-core is, (b) what insight this k-core decomposition method provides, and (c) what is the highest/maximum level, k, of cores present in your network (e.g., Do any 3-cores exist in your network? Do any 4-cores? 5-cores? etc.)?
2. **(3 points)** Visualize your network using k-core decomposition and include the visualization in your report. In a paragraph, discuss your interpretation of the visualization and whether the results of k-core decomposition make sense based on your expectations of the network.
3. **(3 points)** Pick one of community detection algorithms to run on your network. Which

community detection algorithm did you choose and why?

4. **(3 points)** How many communities have been created? For your network, what might a community of nodes potentially have in common?
5. **(3 points)** What is a modularity score? Interpret the modularity score of your results of community detection?
6. **(6 points)** Plot the communities and include the plot image in your report. What information does this layout convey? Are the communities well-separated, or is there a great deal of overlap? Describe the actors between any components and cliques (i.e., brokers). What are common features of these actors?
7. **(3 points)** Present and interpret the in- and out-degree distribution based on your network as well as a log-log plot. Compute and interpret the estimate of the c slope (i.e., alpha value). Note that a *p* value (KS.p) less than 0.05 indicates the empirical data doesn't fit with the power-law distribution.
8. **(3 points)** Present in a plot the observed and simulated values for each average path length and clustering coefficient based on the original network and 1,000 randomly shuffled networks.
9. **(3 points)** Based on these data would you conclude that the observed network demonstrates small world properties? If so, why? If not, why not?
10. **(10 points)** In two or three paragraphs, discuss your major findings of your network based on all the analyses you've done in this exercise and also your own additional analysis if necessary. Your answer here will be evaluated based on depth and comprehensiveness. Thus, you're encouraged to utilize extra information to answer this question. For instance, you can take a look at your original data (i.e., "twitterData," "youtubeData," or "redditData" if you work with the provided R code) in R. These data frames include additional user, text, and time information for your network. Similarly, if you need more insights from your network, feel free to run correlation and regression analysis based on your data collection.