**Week 2 Part 2, Data 620 Assignment**

Group 6

Alice Friedman, Stephen Jones, Jeffrey Littlejohn, Jun Pan

Date: 06/10/2019

version 1.0

Requirement

Centrality measures can be used to predict (positive or negative) outcomes for a node.

Your task in this week’s assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes.  As an additional constraint, there should be at least one categorical variable available for each node (such as “Male” or “Female”; “Republican”, “Democrat,” or “Undecided”, etc.)

In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

For this week’s assignment, you are not required to actually load or analyze the data.  Please see also Project 1 below.

You may work in a small group on the assignment.   You should post your document to GitHub by end of day on Wednesday 6/12.

Title: Analysis of Twitter of Ex-Senator to evaluate the impact of teenagers in political campaign

Introduction

Twitter has become a pervasive tool in election campaigns. Candidates, parties, journalists, and a steadily increasing share of the public are using Twitter to comment on, interact around, and research public reactions to politics (Jungheer 2016). Social media networks form in Twitter around a wide range of terms. As each person uses Twitter they form networks as they follow, reply and mention one another. These connections are visible in the text of each tweet or by requesting lists of the users that follow the author of each tweet from Twitter.  Recently, an essay by Jamie Lauren Keiles on New York Times, “Are These Teenagers Really Running a Presidential Campaign? Yes. (Maybe.)”, causes our attention. Social media is playing more and more important roles on president campaign science President Obama. Teenagers are more active than other age groups by accessing social media tools such as Twitter, Facebook and Instagram. It is important to investigate the impact of teenager group on president campaign. Former Alaska Senator, Mike Gravel, allowed two high school kids accessed his Twitter account for president campaign candidacy. It has caused lots of attention. We are going to scrape Twitter API for those Twitter accounts linked to Mr. Mike Gravel for further analysis.

### Hypothesis

The verified users and un-verified will form two groups. The two groups will form two subgroups surrounding with the person with the highest centrality in each group. We will see who are the most influential person in the subgroup. By social network graph, we will be able to see the communities like different age groups, democrat groups/republication groups, women’s groups/men’s groups and who are closest friends. We will also take a look at whether the community bridge is undecide group? We also anticipate the nodes that belong to more than one group will have higher degree centrality on average than the nodes that belong to the other groups.

Methods

A web scraping through Mike Gravel’s accounts will be conducted. After Mike Gravel announced for President of 2020, all returned tweets recently will be collected. There are two scripts designed to used together to create dynamic social networks from Twitter.com. (1) Gathers tweets containing the same hashtag; (2) make social networks from csv of tweets.  The interaction network collector will form a network between Twitter users based on the mentions in the set of monitored tweets (using the Streaming API). The network of interactions will be adaptively partitioned into snapshot graphs based on the frequency of interactions. Then, each graph snapshot will be partitioned into communities using the Louvain method. Dynamic communities will be extracted by matching the communities of the current graph snapshot to communities of previous snapshot. Finally, these dynamic communities will be ranked and presented to the user in accordance to three factors; stability, persistence and community centrality. The users will browse through these communities in which the users are also ranked in accordance to their own specific snapshot centrality, browse through wordclouds containing a summary of the most frequently used terms in each dynamic community and the most frequent urls. The PageRank algorithm will be used to measure the feature of centrality (Community evolution analysis; Twitter network analyzer).

Discussion

As Kozloski mentioned in his campaign twitter study, The Twitter API on a free developer account will only allow you to search back 9 days from when you execute the function. In order to go back further, we will need to purchase an expensive twitter API subscription. It is unlike that we can afford such expense. A cheaper method for larger projects is simply running the function for the maximum amount of tweets you we collect (just increase the n parameter until we get the same number) and re-running it every 9 days and then merging our data frames into a larger file. Because we have limited time frame, the longitudinal research is difficult to conduct in our case.

References

Community evolution analysis. <https://github.com/socialsensor/community-evolution-analysis>

Jungherr A, Twitter use in election campaigns: A systematic literature review. Journal of information technology & politics. 2016:13(1) 72-91.

Kelies, New York Times. 2019-06-06. <https://www.nytimes.com/2019/06/06/magazine/mike-gravel-teens-twitter-presidential-campaign.html?action=click&module=Editors%20Picks&pgtype=Homepage>

# Kozloski AA. Twitter Analysis in the 2018 Senate Runoff Election

<https://rpubs.com/gladdar/Miss_Election_Twitter_2018>

Twitter networker analyzer: <https://github.com/rajathvsm/TwitterNetworkAnalyzer/blob/master/submission.py>