Analysing Social Network Reactions to 2016 Republican Primaries

Danny Thomson, Eric Ehizokhale

Introduction

2016 is a big year for American politics. President Barack Obama's second term will end, and a new president will take his seat in the White House. Both parties are ramping up to find the candidate with the best chance to win the presidency. The Republican Primary has been in the center of the spotlight with 14 different candidates all vying for the attention of the American public. Some have theorized that the way to stand out in this big crowd is to garner media attention, and many Republican candidates (particularly Donald Trump) have done exactly that by saying controversial statements and potentially seeing poll numbers rise. We want to analyze the truth to this hypothesis by studying the reaction on Twitter after media events, such as a national debate, as well as seeing the values voters care about by analyzing a graph of Reddit comment data.

We drew data from Twitter to use as our metric to measure the response after a presidential debate. While we initially wanted to analyze when the media hounded on a candidate's controversial statements, Twitter doesn't support a data stream for previous tweets. Instead, we've collected datasets of Tweets from November 9th-12th (around the time of the 4th debate) as well as datasets of Reddit comments for the month of September (2nd debate) to analyze how politicians are doing on social media.

Overview

We're using public Twitter data to attempt to answer these questions:

- Why are some Republican candidates polling better than others? If we trivially ranked Republican candidates by degree, or number of followers, then Rand Paul should be sitting at a cozy 4th place with 730,000 Twitter followers. However, he is doing significantly worse than his counterparts Ted Cruz and Carly Fiorina, both with fewer followers. Are there qualities of the Twitter or Reddit networks that might shed light to this question?
- How do people truly react to statements, news, or debate claims during the presidential campaign? We will be performing sentiment analysis on our Twitter dataset, as well as analyzing aspects of the Twitter network, to see if we can predict who did the best in a debate by just looking at a subgraph of Twitter sentiment.

Reviewing Prior Work

In [1], Lu, Shah, and Kulshrestha investigate the effects of social media in the 2014 India election. In order to analyse the effects of the twitter on the 2014 Indian elections, the authors downloaded around a year's worth of Twitter data about the Indian election. The resulting

dataset was around 7 gigabytes in size and contained over 10.6 million tweets. With a dataset that large, the authors needed to categorize the data into different classes in order to break the data down. To organize the data, the authors used keywords to assigned tweets to the political party along with a sentiment analysis to evaluate the opinion of the tweet. With their categorized. All these different examples showed that the BJP used Twitter in a much more effective matter.

In [2], Baskhy, Messing, and Adamic use network analysis tools to measure Facebook users' exposure to ideologically diverse news. They constructed a deidentified data set of 10.1 million active Facebook users and 7 million web links shared over a 6 month period, and measured the amount of cross-cutting content in the network by the amount of potential exposures (3.8 billion), true exposures (903 millions), and clicks (59 million). In any given cluster of friends, they found that liberals tend to be connected to fewer friends who share information through the other side and are less likely to click on ideologically different news than their conservative counterparts.

In [3], Kwak, Lee, Park, and Moon use a wealth of network analysis tools to examine whether Twitter trends closer to a social media platform or a news network. The Twitter directed network of followers to tweets is a non-power-law distribution, unlike most other social networks that have power law distributions with exponents between 2-3. Since only 22% of Twitter relationships are reciprocal (both people follow each other), the shortest path lengths tend to be larger and will have a higher degree of separation. One especially interesting area of analysis this article attacks is the impact of a retweet. They created information diffusion trees of every tweet that is retweeted and analyze the subgraphs of these retweet trees. This distribution of retweet trees also follows a power law distribution. Once retweeted, the tweet gets retweeted almost immediately in the 2nd through 4th hops.

Our project was motivated by many of the positive aspects of these papers. For example, we liked the sentiment analysis work done by the India paper, as well as noting how important retweets are as diffusion of information in the Twitter paper. We hope to combine both by measuring the reaction to retweeted information through sentiment analysis. In the context where presidential candidates need a comparative advantage to gain votes, understanding the what motivates followers to vote for one candidate over another by analyzing followers' preferences will be helpful, as we later show with our analysis of Reddit comment data.

Twitter Analysis

1. Data Collection

We used the public Twitter API to collect data for this project. Our main dataset consists of about 200k tweets pulled between November 9th -12th to pull tweets about frontrunners of the Republican primaries (Donald Trump, Ben Carson, Marco Rubio, Ted Cruz, and Rand Paul) around the time of the fourth Republican presidential debate. Many of our challenges concerned collecting this dataset. The Twitter public streaming API returns only <1% of all

public Twitter traffic at any moment, so we assume that the tweets returned from the public API are independently and identically distributed across all potential tweets that could be returned.

2. Algorithms

We used a Naive Bayes classifier to perform sentiment analysis on these tweets after the debate. It uses Bayes' theorem and uses a strong assumption that features contribute independently to each classification and do not affect the probability of other features appearing [4]. To make a prediction on a new example with features x, we pick the classification that has the highest posterior probability after calculating

$$\begin{aligned} p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\ &= \frac{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1)}{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1) + \left(\prod_{i=1}^{n} p(x_i|y=0)\right) p(y=0)}, \end{aligned}$$

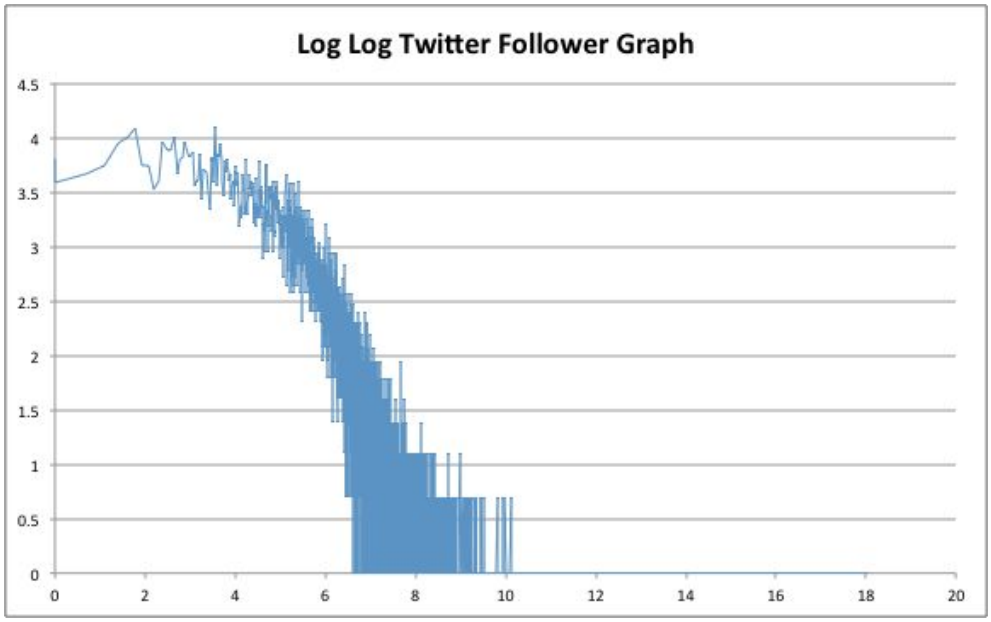
Naive Bayes is a popular classifier choice in text-based sentiment analysis, and we also prefered it because it has comparatively smaller generalization error than many other classifiers with a small training set. After optimizations, we got our classifier to around 83% generalization error (a good benchmark given that humans only agree on sentiment around 80-90% of the time).

Method	Accuracy
Single word, no optimizations	0.73
Removed general stopwords	0.71
Removed political-specific stopwords	0.74
Minqing Hu and Bing Liu manual lexicon features	0.78
Include bigrams	0.81
Positive/negative bigram lexicon	0.83

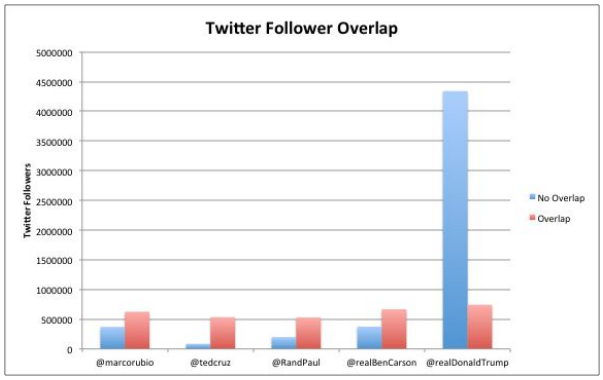
Table with Naive Bayes optimizations on manually classified political training set

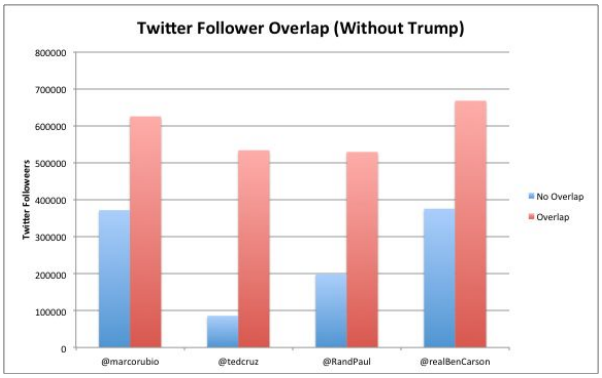
3. Analysis

We created a log-log graph of the follower numbers for each unique user in our dataset. It follows that Twitter followers do not follow a power-law distribution, but rather an exponential one.



We wanted to see the overlap of each candidate's Twitter followers to see if that could produce any insights of where followers of certain candidates sit in the overall Twitter graph. We found the below graphs.



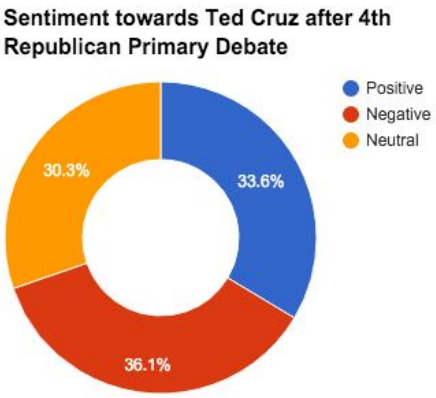


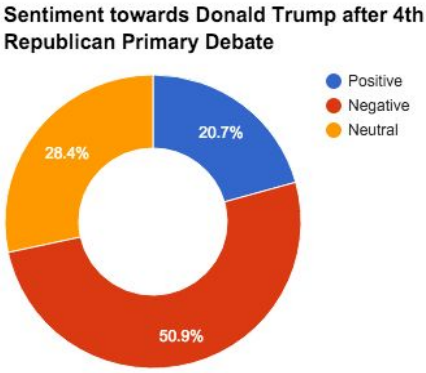
Many followers of Trump do not follow the other candidates, but this is to be expected — not only does Trump have a wealth more followers (over 4 million more than Ben Carson, with 1 million), but many of his supporters have expressed their disillusionment with modern politics and loyalty to Trump even if he changes to an Independent ticket. When looking at the follower overlap excluding Trump, we get a little more information. Both Ted Cruz and Rand Paul seem to not have as many followers who only follow them as candidates, which may signal fewer dedicated "fans" and lower polling numbers.

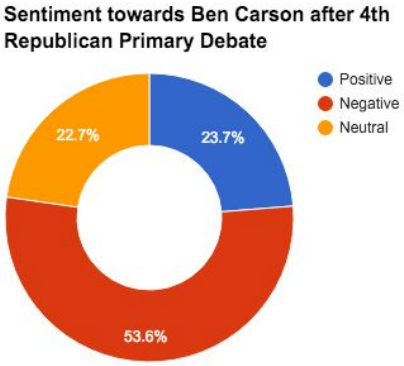
Determining influence through potential retweets seems very pertinent when looking at a political subgraph of Twitter. Analysis of tweets from researchers at Northeastern and Brown showed that approximately 25% of all tweets are retweets. After writing a script, we discovered that depending on the data sample we look at, approximately 45-55% of our

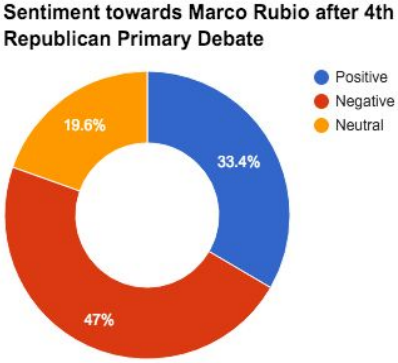
tweets are retweets (we didn't run the script on all of our files, but we've found that tweets collected on Nov 9, before the primary debate, were around 45% retweets, while tweets collected during the debate were around 55% retweets). [5]

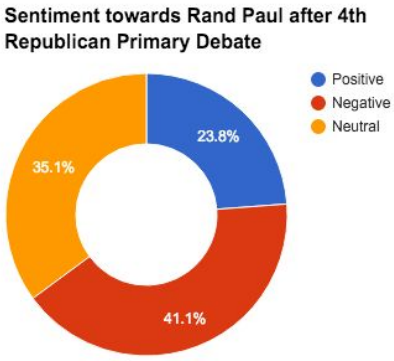
We got the following figures after running the sentiment analysis on our data.











From the Twitter public API, it implies that the tweets returned from their tracker are IID. For this reason, we can assume that our sentiment analysis performed on a subgraph of all tweets generalizes well to the sentiment of the full Twitter graph during this time. From this, we see that Marco Rubio and Ted Cruz have the highest positive sentiment after the debate,

and according to major news outlets and political analysts, both of them performed very well during the debate. In contrast, analysts said Donald Trump underperformed, and both Ben Carson and Rand Paul did okay.

Interestingly, Rand Paul has a high amount of neutral sentiment. A visual glance at the classified tweets shows that many tweets collected on Rand Paul were simple news headlines of his statements, and he didn't seem to have as many users independently reacting to anything he claimed. Limitations of the public Twitter API kept us from exploring further network properties of Rand Paul's followers further, but this suggested that there may be something unique about Paul's followers that doesn't contribute as much lively discussion to Twitter. In order to potentially answer our questions, we turned to a new dataset — Reddit.

Reddit Analysis

1. Overview

Reddit is an online community in which users can submit links to specific subreddits (groups on a specific topic) that get upvoted or downvoted by users in the community. Users can then comment on links posted and reply to each other's comments like a standard social network. By subscribing to specific subreddits, users can make their front page only display information immediately pertinent to them. Because Reddit allows for clearer grouping of categories, we wanted to see if we could find any trends behind the topics people talk about who discuss the major Republican candidates.

2. Data Collection

We created a graph using a publically available dataset consisting of all 55.5 million comments published in the month of September [6]. We chose September because it was the most recent month available to us, and the second Republican primary debate happened in the middle of September, which would spark Reddit discussion about the debates. For each candidate, nodes in our graph are subreddits, and edges are created in this way: we get users who have commented about that candidate at all during the month. We then filter comments based on political keywords such as economy, immigration, guns, etc. Finally, we create an edge between all of these filtered subreddits a user has commented in. In other words, for a Reddit graph of candidate X, an edge exists between two subreddits if a user who has commented on X has posted in both subreddits. When this happens more than once, we increment the edge weight by 1.

3. Figures

Using Gephi, we created visualizations of these subreddit graphs.

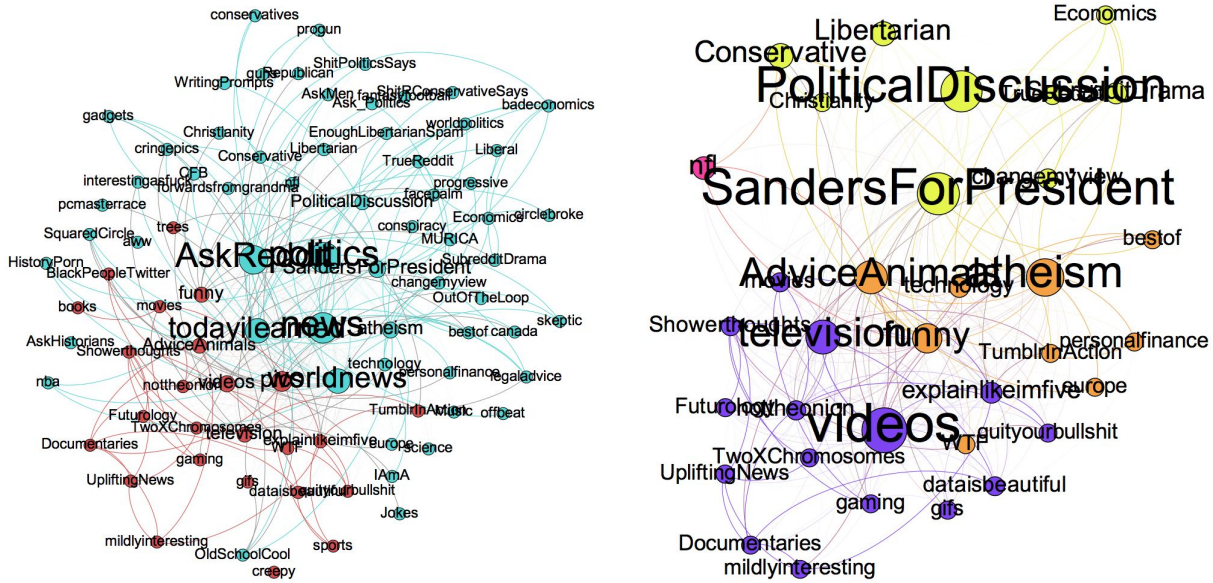


Figure 1: Overall Ben Carson subreddit graph (left), Ben Carson graph with five top weighted nodes removed (right)

Nodes were ranked on betweenness centrality, where larger nodes have larger betweenness. Betweenness centrality is a measure of the amount of times a given node appears in the shortest paths between any two other nodes, and we used this as a main centrality measure for our visualization because it shows the subreddits that one often would have to “cross through” to get between any two general interests a user might have.

Nodes are clustered using the Gephi Modularity clustering functionality that uses the Louvain algorithm to optimize clusters. Visually, we notice these clusters give some semantic meaning. Before removing the top nodes (left graph), we see the blue nodes tend to be more politically focused, whereas the red nodes are an assortment of generally popular subreddits (funny, adviceanimals, videos, explainlikeimfive). Because AskReddit, news, worldnews, and politics are commonly the most central nodes for all of the candidates, We reran the graph removing these nodes to see any of the unique subreddits for each candidate.

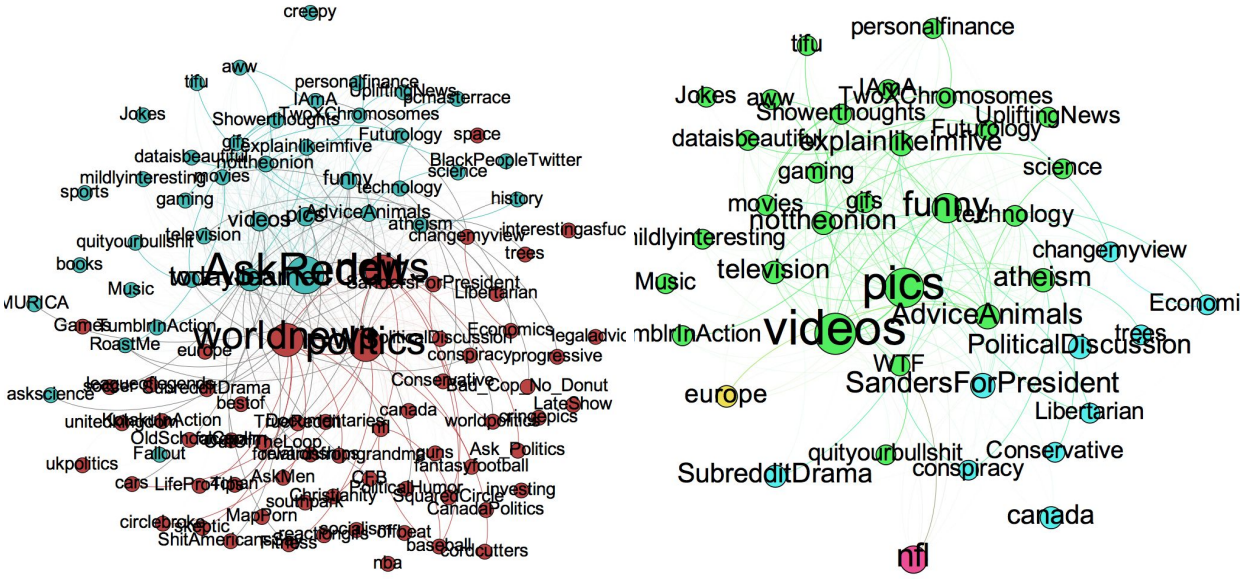


Figure 2: Overall Donald Trump subreddit graph (left), Donald Trump graph with five top weighted nodes removed (right)

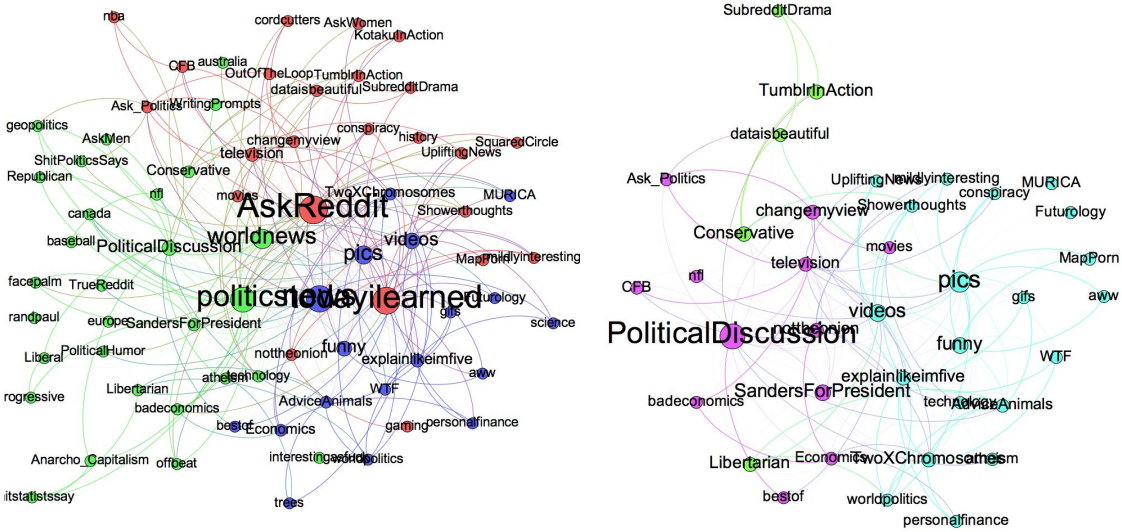


Figure 3: Overall Marco Rubio subreddit graph (left), Marco Rubio graph with five top weighted nodes removed (right)

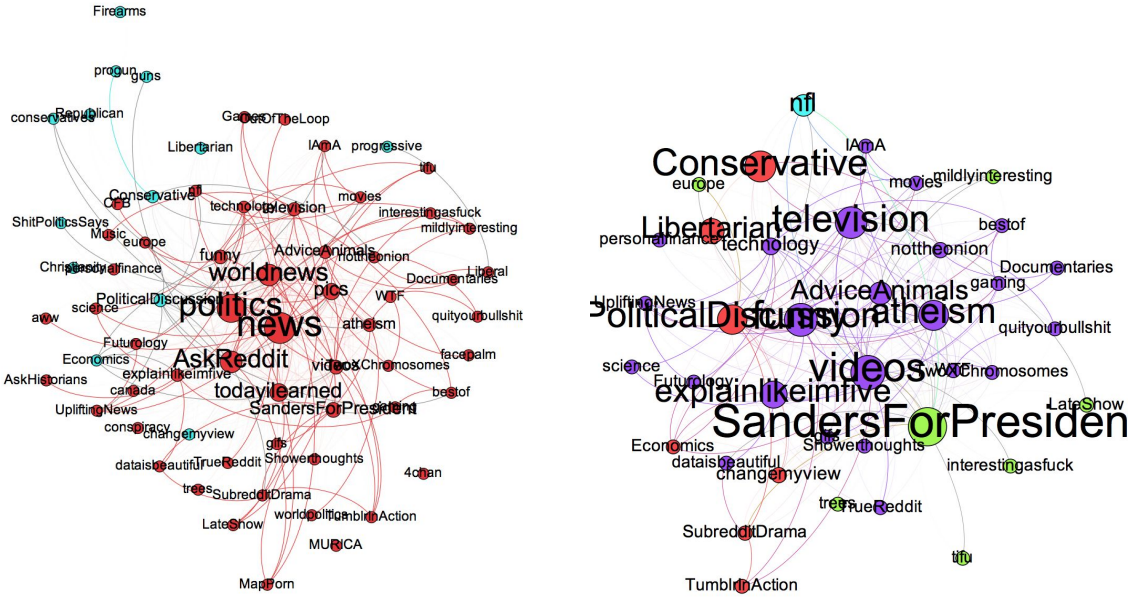


Figure 4: Overall Ted Cruz subreddit graph (left), Ted Cruz graph with five top weighted nodes removed (right)

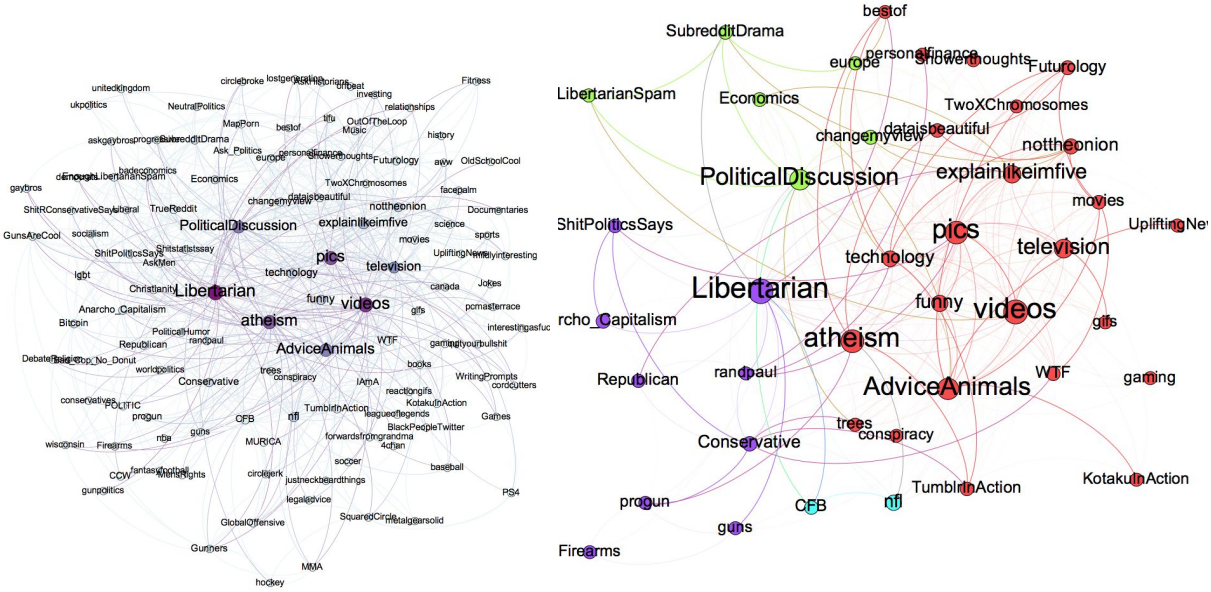


Figure 5: Overall Rand Paul subreddit graph (left), Rand Paul graph with five top weighted nodes removed (right)

4. Analysis

When the most central subreddits were removed from each candidate's graph, we see a few differences between central topics around each candidate. Many of Donald Trump's more central subreddits are generally popular subreddits without any specific political focus, such

as pics, videos, funny, nottheonion, etc. In contrast, of Rand Paul's central nodes focus on political issues, such as libertarian, economics, conservative, political discussion, etc. Rand Paul is the most libertarian candidate of the current running Republicans and has a more niche subset of followers, in that .

Of course, the above data is not a perfect reflection of what potential republican voters find most important. While some subreddits in the data make perfect sense (gun, progun, personal finance), others, namely SandersForPresident, are not typical interests of American conservatives. This likely comes from the fact that users on Reddit often engage in debates with those with opposing viewpoints, so that a user posting about a Republican candidate might be attacking Bernie Sanders frequently in that subreddit, or a Sanders supporter might be attacking republican candidates.

Overall Summary

The root of many of our difficulties for the project lies in the massive dataset required to make and analyze a meaningful Twitter network, and the limitations of the public Twitter API. Our project proposal had some very ambitious goals that are not possible without paying multiple thousands of dollars for data. Nevertheless, through the analysis of the Twitter network we could accomplish with the public API, as well as the further analysis sought through our subreddit network, we were able to come to interesting conclusions about the current Republican race. Namely:

- Sentiment in Twitter during the Republican debates changes in real time, and by doing sentiment analysis on any independently and identically distributed subset of the Twitter network, we are able to see with high confidence which candidates "won" the debate.
- Rand Paul's high amount of Twitter followers don't necessarily translate into spreading a more favorable influence over Twitter. His followers, like Cruz, tend to follow other candidates and not just him, so it's unlikely that strong pro-Paul messages are generated throughout the network. Additionally unlike Cruz and as shown by the Reddit data, his supporters are interested in a more niche range of political topics and do not necessarily have discussions in more generally popular subreddits as much.

References

[1] Lu, D., Shah, A., & Kulshrestha, A. (2014). India's #TwitterElection 2014. CS 224W Fall 2014.
[2] Bakshy, E., Messing, S., & Adamic, L. Exposure to ideologically diverse news and opinion on Facebook, Science 348(6239), 1130-1132, 2015.
[3] Kwak, H., Lee, C., Park, H., & Moon, S. What is Twitter, a social network or a news media? World Wide Web Conference, 2010.
[4] Ng, Andrew. "Generative Learning Algorithms." Web. 16 Nov. 2015.

[5] Liu, Yabing, Chloe Kliman-Silver, and Alan Mislove. *The Tweets They Are A-Changin': Evolution of Twitter Users and Behavior*. *Northeastern University*. N.p., n.d. Web. 16 Nov. 2015.

[6] https://www.reddit.com/r/datasets/comments/3oiv9z/reddit_september_comment_archive_is_now_available/