DIG210 Data Culture
Data-Based Project
Professor Mundy
Brandon Liang

Twitter Sentiment and Network Analysis

Introduction

Coincidentally, this past month marked one of the most dramatic presidential elections in the history of the United States. Besides the roller-coasting election coverage, this was also an election heavily discussed on social media. With the rapid emergence of social media in the past decade, online users have easier and easier access to publicly display their opinion. On a subject matter like this year's presidential election, social media is second to none when it comes to examining public sentiments toward the two presidential candidates and the election outcome.

Thus, the objectives are to quantitatively measure and examine reactions on 2016 Election from social media and to compare such reactions before and after the election result. The social media of choice is Twitter, since it is the most commonly used social media platform for brief comments. My approach is to use Twitter API to scrape tweets on 3 different hash-tags: #Election, #Hillary and #Trump, before and after the election night of November 8th. This would produce a total of 6 different datasets and graphs.

Network Analysis

Since most twitter users retweet from other, I am interested in how the Twitter world connects as a social network. That being said, each twitter user is a source node; any twitter user whose tweet is retweeted is a destination node; a tweet is then an edge or a connection, connecting the source and the destination nodes. By defining this network relationship, we are able to see how connected we all are on social media. Which user is the center of gossips? Which users are not directly connected, but rather share a connection? What is the largest component of the social network? What does it show about us during the election online? All these interesting questions can be answered by force-directed graphs.

Sentiment Analysis

Sentiment analysis refers to using Natural Languaging Processing and related text analysis fields to identify and quantify subjective information such as attitude, affection and sentiment in textual source material. For example, "happy" may have a positive score while "sad" may have a negative one. In this project, sentiment analysis provides us with a quantified metric to measure the positivity or negativity of each tweet on the subject of 2016 Election. In the final graphs, color is incorporated to reflect the sentiment of each tweet (edge) of the tweeting user (source node). Macroscopically, it provides a straightforward platform to evaluate the overall public sentiment and specific sentiment distributions of each subcomponent.

Data & Methodology

Twitter provides public API for developers. I used its API to scrape relevant tweets information with the help of its hash-tag search function. As I mentioned earlier, I used 3 different hash-tags: #Election, #Hillary and #Trump, to get tweets both before and after the election night, yielding 6 total datasets. In each dataset, each data point includes the metadata of a tweet in addition to the text itself, such as user information, geographic location, number of followers, number of retweet, etc. I was most interested in the user of the account and the text of the tweet. From the text, I was able to identify and extract the original twitter user of the retweet. Moreover, I applied sentiment analysis to quantify each tweet text based on its sentiment. Final result included data points each containing the specific twitter user (source node), the original twitter user of the retweet (destination node), the text of the tweet (edge) and the sentiment score of the tweet (reflected by color).

In the final display, for simplicity reason, the graphs only show the name of a twitter user if his/her tweets were retweeted more than 5 times by other users.

Tools

Data scraping from Twitter and data processing were done in Python. More specifically, I used a Python module designed for using Twitter API called Tweepy. For sentiment analysis, I used the NLTK module in Python to tokenize each text and used NLTK's sentiment intensity analyzer to quantify each tweet's sentiment. For final visualization, I organized my data in a proper form for Force-Directed Layout Graph in D3.

Legend

Each source node (a twitter user who tweeted on one of these 3 subject matters) as well as each edge (a tweet) is categorized based on the corresponding tweet's sentiment score, the "compound score" of the tweet determined by NLTK's sentiment intensity analyzer. The compound score is in the range of [-1, 1], with "-1" meaning an extremely negative sentiment and "1" meaning an extremely positive sentiment. I broke down the compound score into 5 categories: [-1, -0.6), [-0.6, -0.2), [-0.2, 0.2), [0.2, 0.6) and [0.6, 1]. In terms of tweet sentiment, each range translates to "Very Negative", "Quite Negative", "Fairly Neutral", "Quite Positive" and "Very Positive". I also used 5 different colors to represent the sentiment categories in the final force-directed graphs and here is the legend:

Red: [-1, -0.6) --> Very Negative
Blue: [-0.6, -0.2) --> Quite Negative
Green: [-0.2, 0.2) --> Fairly Neutral
Purple: [0.2, 0.6) --> Quite Positive
Orange: [0.6, 1] --> Very Positive

Dataset & Links

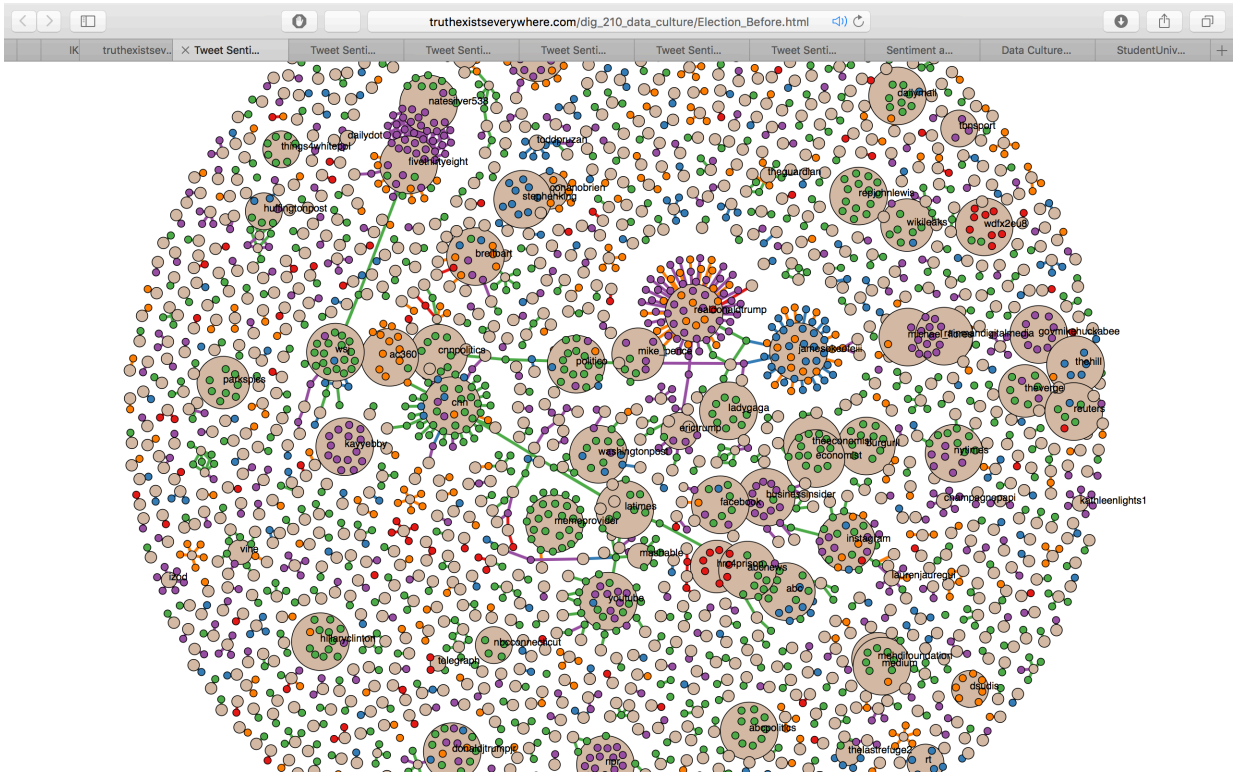
Scrapped & Processed Datasets:
https://github.com/BrandonLiang/Twitter_Sentiment_Network_Analysis_2016Election

Final Visualization and Display:
With the six datasets, I have produced six force-directed graph using my methodology. Each hash-tag topic has two datasets, one before the election and the other after the election. The link to the graphs is here:
http://truthexistseverywhere.com/dig_210_data_culture/final_project.html
(Perks: You can move your mouse to any node and zoom in/out by scrolling your mouse up/down.)

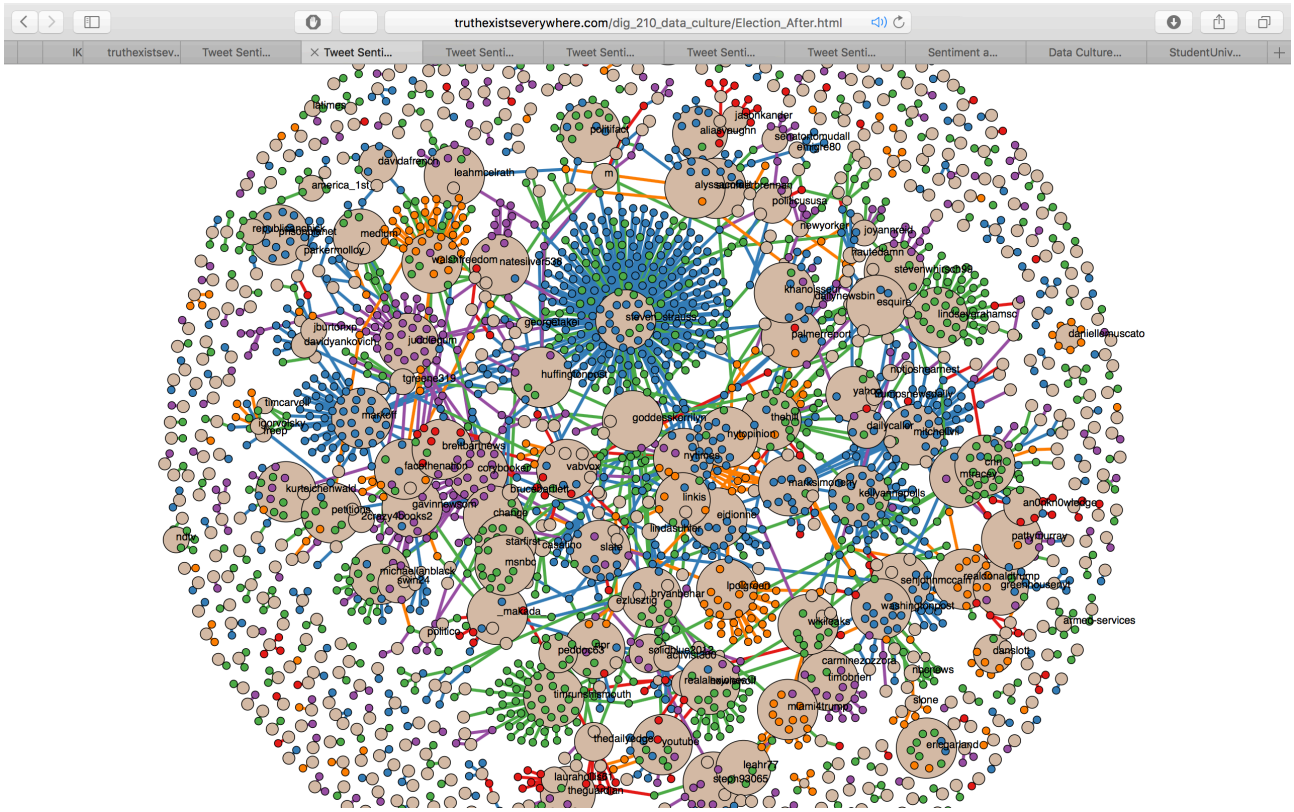
Visualization

Here are the six screenshots for each graph:

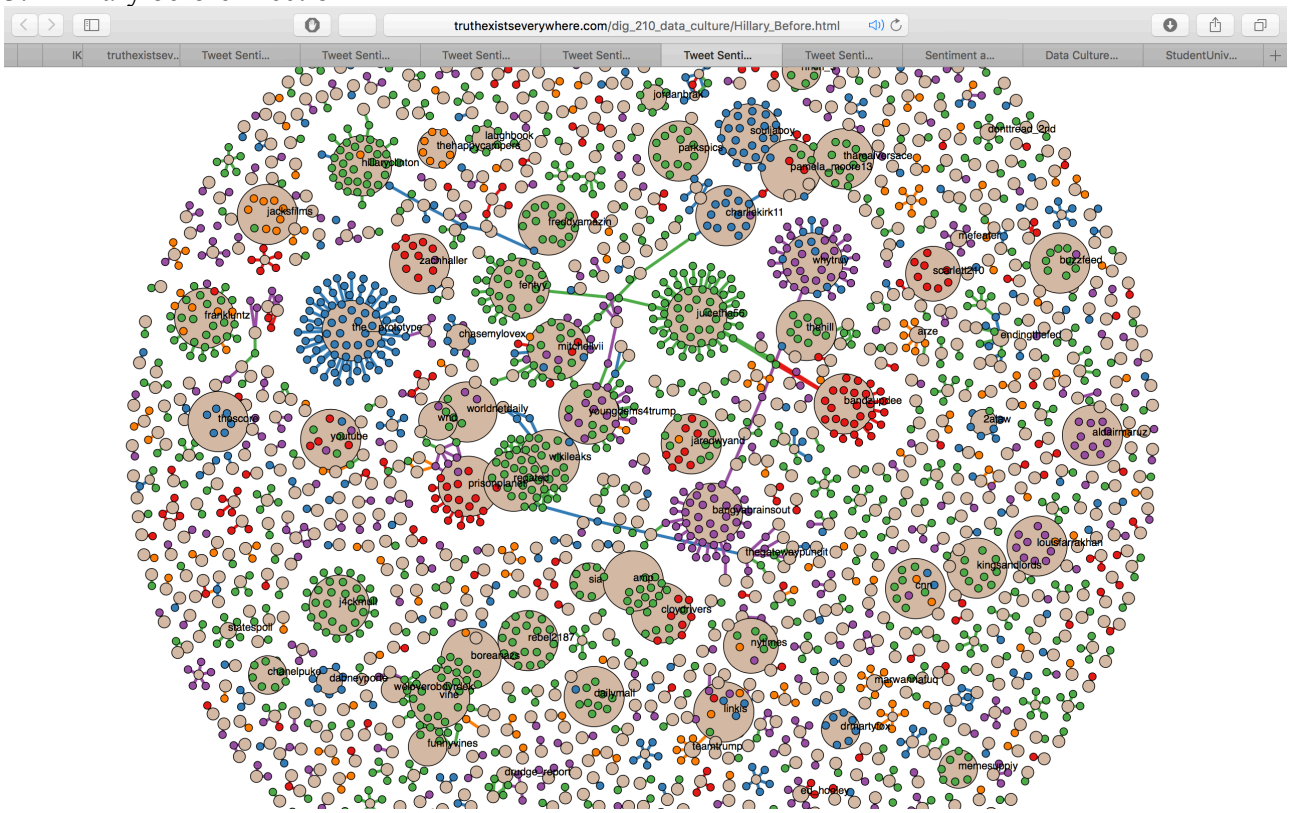
1. #Election before Election



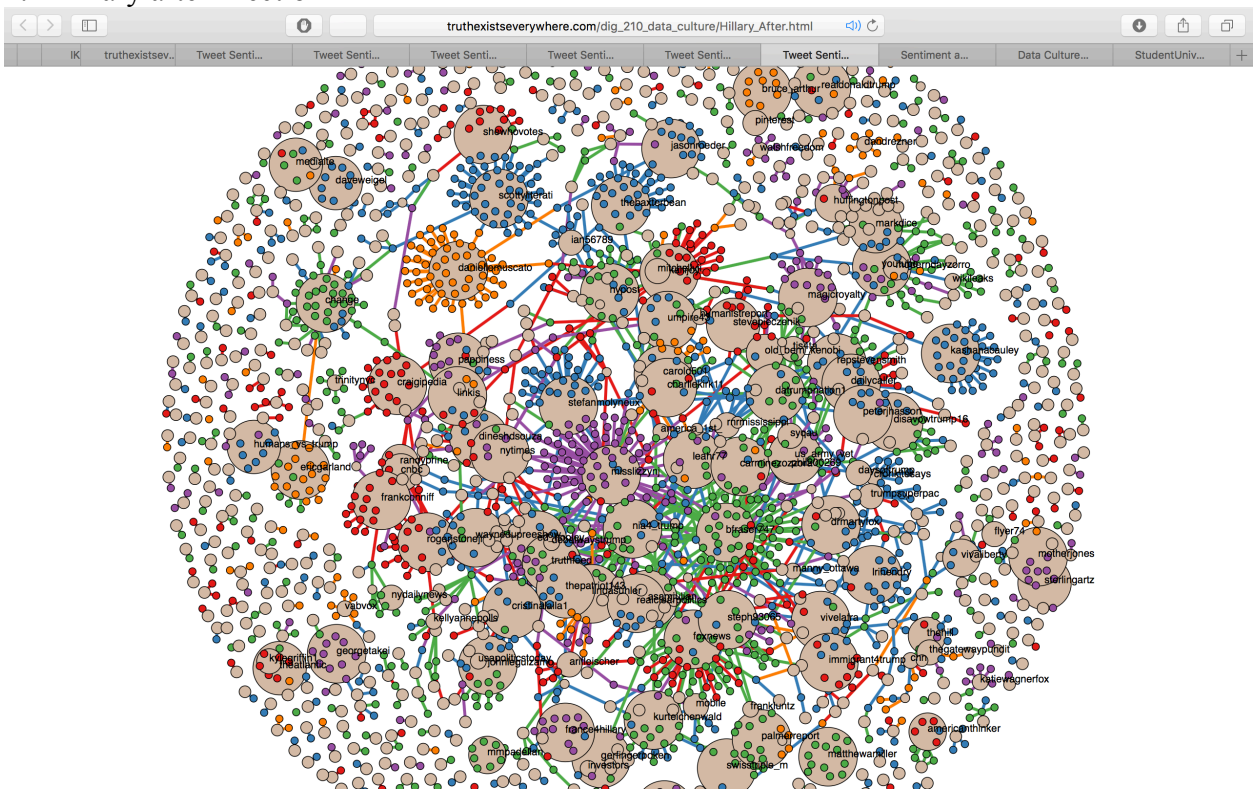
2. #Election after Election



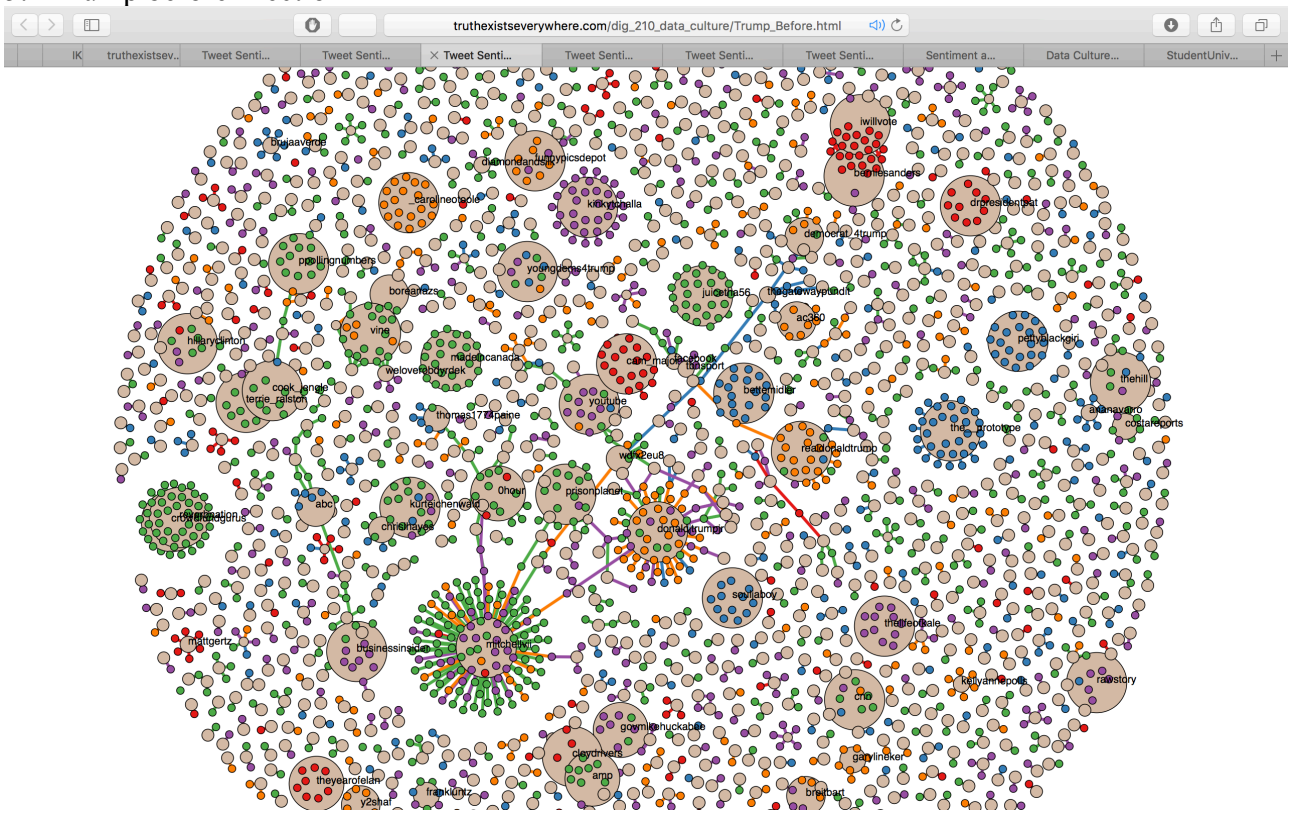
3. #Hillary before Election



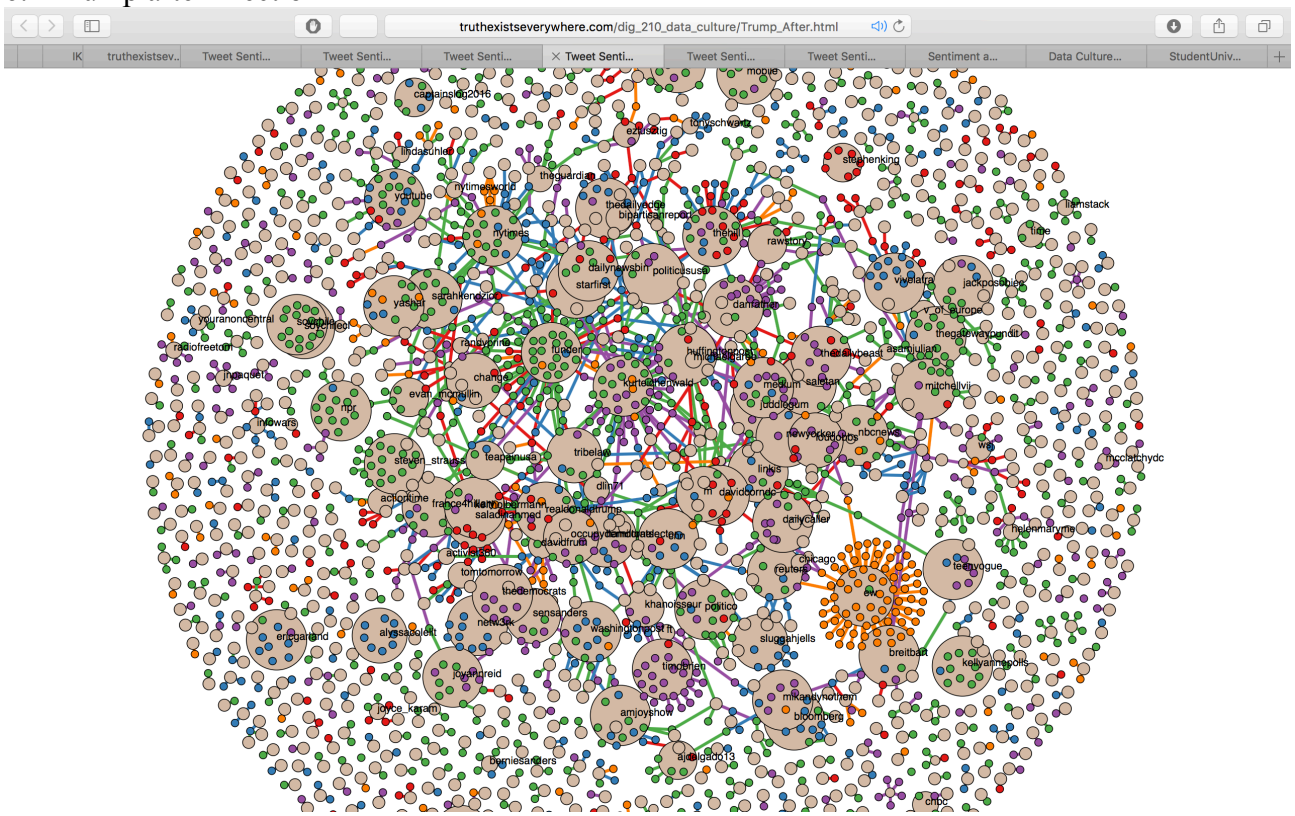
4. #Hillary after Election



5. #Trump before Election



6. #Trump after Election



Results & Arguments: Are We Panicking?

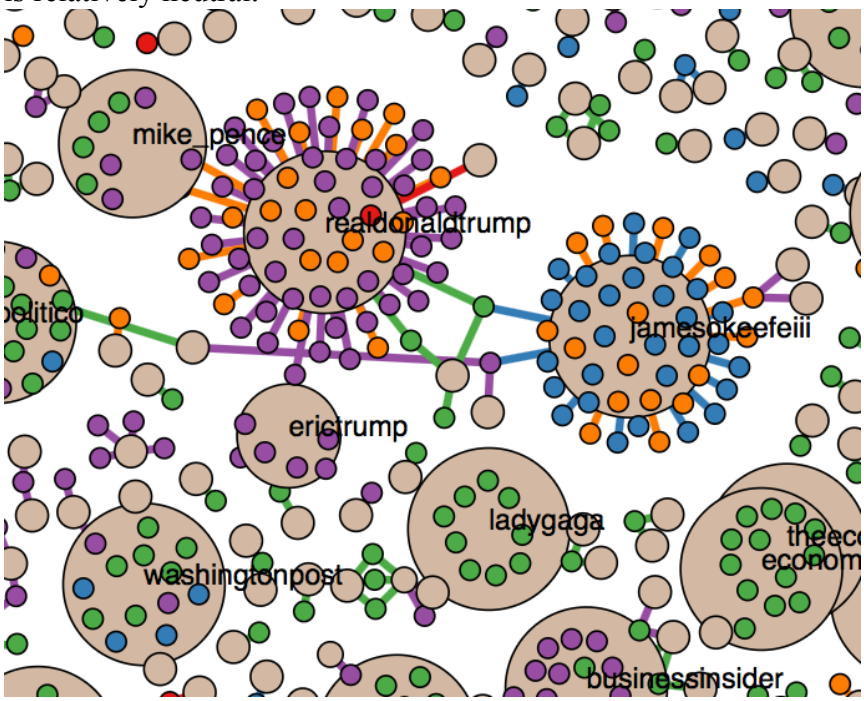
From the first glimpse of the 6 graphs, I found a common theme with regard to network connection: for each hash-tag, the graph based on data after the election is a lot tighter than that based on data before the election. A key reason is that in the D3 parameter tuning, I passed in "1" as the gravity parameter for all 6 graphs, which reflects the gravitational force that holds all nodes toward a centroid. The more connections a node has, the stronger force it is applied to. With the same gravity parameter through all 6 graphs, a tighter connection of nodes implies that after the election, twitter users tended to retweet more frequently from other twitter accounts and these "author" twitter accounts tended to be retweeted more frequently; in other words, after the election, it was shown on Twitter that the source of words became narrower and the impact of those words became heavier.

Moreover, when I compared the color distribution before and after the election, here are some interesting results.

1. #Election

Before the election, other than the graph being much looser, the overall sentiment seemed quite neutral, reflected by a dominant spread of green nodes and edges. However, there are some outlier components. A large node under the name "realdonaldtrump" has only

purple and orange connections, meaning its tweets were pretty positive and popular. The node "realdonaldtrump" also share quite a few positive connections with "mike_pence" and "erictrump". The account "fivethirtyeight" also has a lot of purple connections, meaning its tweets were quite positive as well; notice that the node "fivethirtyeight" shares a lot of connections with "natesilver538", which is expected because Nate Silver is the editor-in-chief for 538. However, an account called "jamesokeefeii" had quite polarizing tweets, as it has dominantly orange and purple connections. Moreover, there is one twitter user who connects with both "jamesokeefeii" and "realdonaldtrump"; the connection to "jamesokeefeii" is quite negative and the connection to "realdonaldtrump" is relatively neutral.



After the election, the overall sentiment had a different dynamics. We see a rising number of blue connections in the center of the graph, dominated by an account under the name of "steve_strauss", whose tweets were retweeted 238 times. The node also shares many common connections with "huffingtonpost", as Steven Strauss is a leading small business author for Huffington Post. Moreover, accounts like "steve_strauss" share connections with many big player nodes with a wider variety of colors, a reason why this graph is a lot tighter than the previous one. There are many more implications to explore in this graph, but the main takeaway is that, general public's sentiment toward #Election reached a lower point after the result of the election.

2. #Hillary

In the midst of FBI re-investigation and controversies, Hillary Clinton's public reception before the election is reflected quite well in the graph. Most accounts had fairly neutral words to say, including "hillaryclinton"; however, some were very aggressive, such as "the_prototype", "bangyabrainsout", "whytruy", "prisonplanet" and "wikileaks", which all have negative connections.

After the election, thing got wild. The graph witnesses an increasingly wider variety of sentiment toward Hillary. There are a lot more blue connections, indicating quite negative

sentiments; there is also a fair share of red connections, from nodes such as "immigrant_trump". Amid the disappointment of the result, at least two accounts stood for Hillary, which are "daniellemuscato" who had all orange connections and "bfraser747" who stood in the close to the centroid of the graph and had mostly green connections. Moreover, a fair share of green and purple connections close to the centroid of the graph are surrounded by dominantly red and blue edges; this shows that there must be some sort of twitter conversation in the center of the social network of which the sentiment oscillates between positive and negative.

3. #Trump
Before the election, the sentiment in general was at peace; however, there are still outliers, such as "berniesanders", who seemed to have only negative things to say about #Trump. On the other hand, "realdonaldtrump" and "donaldtrumpjr" had some quite nice things to say about #Trump. One other thing to notice that, there aren't many nodes that share connections; in other words, most accounts that tweeted on #Trump are quite isolated in the graph.
After the election, this isolation halted. Nodes became tighter through an enormous amount of mutual connections and the public sentiment rose despite some occasional red and blue connections in the center of the graph. It shows that while the public was quick to accept the result, there were still some heated debates going on in the center of the social network. An account called "ew" is quite isolated in the middle of the graph, with mostly orange connections; it seems that this twitter user was quite content about the election result.

There are still a lot more to study and extract from the 6 graphs. But the core is evident. We as the public have been reacting in a wild fashion regarding the result of the election. It shows how dramatically people's emotions and sentiments changed during the course of the election and how closely related we become after a key event like this. Potentially, it even shows how simple it is for the public to change its view and how easy it is for us to be affected by others' opinions.

Assumption, Limitation & Discussion

Exciting as it looks, this data analysis piece still has limitations based on its assumptions. The key assumption is based on this question: are these tweets wholesome and representative? The answer is "I am not sure", because there is no standard to follow to ensure my sample data are enough to represent the whole population, let alone that these are only from users online; the only alternative I could do is to keep gathering more and more samples to grow my database and make my arguments stronger. No matter how seemingly informative these graphs and analysis are, they are only a suggestive tool to help us understand the subject matter better, not a definitive one.

This ties into the general misconception and application of data analytics. Data science is an emerging technology that offers enormous power, but it is an advanced technology with tradeoffs that most tend to downplay or ignore. When applied properly, data analytics can offer powerful insights and hidden patterns behind numbers; but proper application requires careful examination and understanding of assumptions and

representativeness. For instance, data visualizations of the graphs in this project include oversimplification of turning complex tweet texts into decisive quantifications; while the sentiment scores offer succinct understanding of the tweets, what the tweets actually said should also not be forgotten. As one expert elaborates, "the danger lies in trusting data [and data analysis] too much without grasping its limitations and the potentially flawed assumptions of people who build the models."[1] Nowadays, more and more individuals and companies have realized the importance and power of data analysis but few truly understand the essence of what data provides, as the MIT Sloan School of Management professor Erik Brynjolfsson puts it, "the key thing to understand is that data science is a tool that is not necessarily going to give you answers, but probabilities."[2]

Moreover, in my opinion, data is a subset of information, but not necessarily information; in other words, data itself is just another form of numbers. It doesn't become richly useful unless combined with proper tools of algorithms and human domain knowledge. Just like fuel doesn't become productive until pumped into vehicle, data isn't lively without narrative. In this project, the sentiment scores and tweets connections would have been meaningless without the background of this unprecedented presidential election and the understanding of social network.

Another anecdotal fact is that while I am using all these twitter users' tweets for analysis, they are most likely unaware of that. This is due to a voluntary permission they gave away by annoyingly skipping through lines of "terms of use and privacy" when they registered their accounts.

Technical Difficulty

1. It may take a while for the graph to load properly after clicking each link. This is due to the large volume and complex intercrossing connections of the nodes.
2. It is hard to fit all nodes and edges into the web browser window without breaking the balance of the graph. Thus, I enabled the drag and zoom functions for users to closely examine any part of the graph they desire to.
3. Some edges that are overlapped by other nodes are not displaying their links since they are covered. But it is not hard to identify such edges' connections based on their positions.

Reference

1. Steve Lohr & Natasha Singer. *How Data Failed us in Calling an Election*. http://www.nytimes.com/2016/11/10/technology/the-data-said-clinton-would-win-why-you-shouldnt-have-believed-it.html?_r=0

[1] *How Data Failed Us in Calling an Election*
[2] *How Data Failed Us in Calling an Election*