Project 1 Report

Christopher Hong A20400014

Deepika Padmanabhan A20456289

Abstract

In this project, we collected 1,000 tweets in English with hashtags of “#Trump” or “#Biden” or both from Twitter. We created and visualized retweet and quote directed graphs (networks). Furthermore, we analyzed the aforementioned networks by calculating the in-degree, betweenness-degree, out-degree, eigenvector centrality, katz centrality and hashtag frequency. The analysis showed that, among these 1,000 Twitter users, the majority retweeted or quoted others’ tweets rather than tweeted their own tweets. The hashtag frequency showed that far more Twitter users mentioned Trump than those who mentioned Biden.

1 Introduction

Twitter, created in 2006, San Francisco, CA, US, is a social network platform that enables users to share short text messages (tweets) to the world. Twitter users talk about a wide range of topics of the day such as celebrities, technology, politics, and more. They interact with others through following, liking, retweeting, quoting, etc. These connections and interactions form a huge social network.

In this project, to understand and extract patterns from Twitter users who mentioned one of the recent hottest political figures in the world, we collected, visualized and analyzed 1,000 tweets sampled from Twitter using the free and open-source projects Networkx and Jupyter Notebook. We started by collecting the tweets data using tweepy API, wrangled the collected tweets using Pandas, visualized and analyzed the retweeted and quoted tweets using Networkx.

2 Data Collection

We downloaded 1,000 tweets in English with hashtags of “#Trump” or “#Biden” or both using tweepy API. The downloaded data set has attributes such as user, retweeted status, quoted status, etc., which would be used for later social network visualizations and measurements.

We wrangled the data set by extracting attributes such as twitter user screen name, the screen names of users who retweeted tweets and the screen names of those who quoted tweets. These attributes would be used to create a retweet and quote tweet directed graphs.

3 Data Visualization

3.1 Retweet Network

We created a retweet directed graph (1275 nodes and 831 edges) from the data set using the networkx API (**Figure 1**)[3]. The source nodes were the screen names of users who retweeted someone else's tweets. The target nodes were those who tweeted tweets on Twitter. The larger the size of the target nodes, the higher the number of those tweets being retweeted.

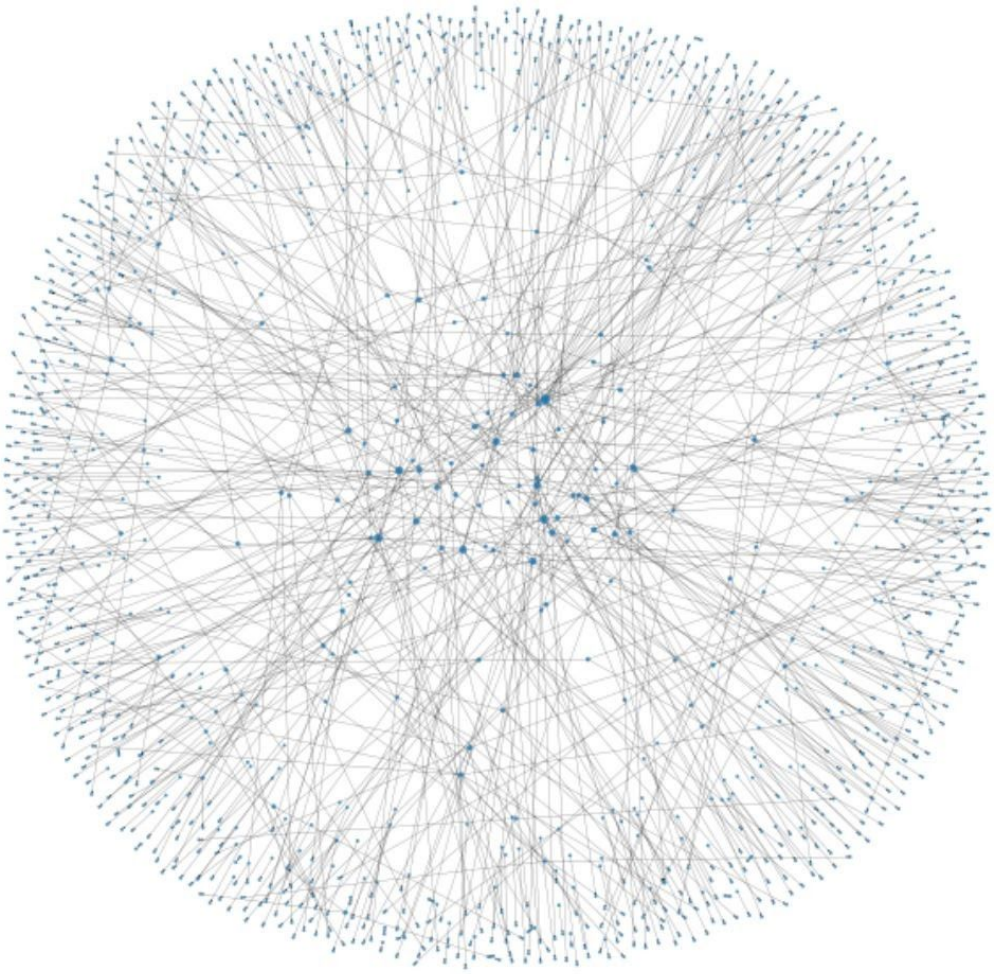

Figure 1

In order to get a better insight of the data, we filtered the data as follows:

- Retrieved the frequency count of a column ‘retweet_user_screen_name’. This would give us count of the number of users each retweeted user has tweeted to. Below is the sample of result:

```
tweets_processed['retweet_user_screen_name'].value_counts()
```

```
DonaldJTrumpJr    14
SenSanders        14
KamVTV            14
ewarren           14
JackPosobiec      13
                  ..
Qtah17             1
MysterySolvent     1
BoSnerdley         1
IPOT1776           1
whoisbenchang      1
Name: retweet_user_screen_name, Length: 471, dtype: int64
```

Figure 2

- The data was filtered with the rows matching retweet_user_screen_name with the usernames “DonaldJTrumpJr”, “SenSanders”, “KamVTV”
- A weighted digraph is plotted using the filtered data[2]. The weights of the edges are the number of retweets done by a particular retweet_user. This retweet count is normalized by dividing each value by the maximum count.

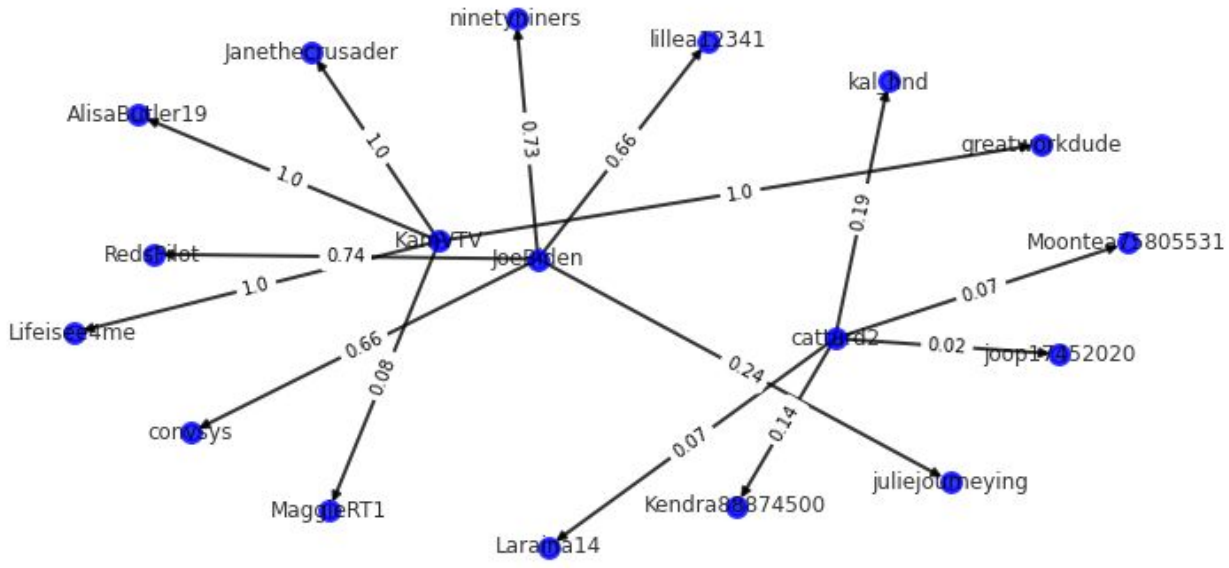

Figure 3

3.2 Quote Network

We also created a quote directed graph (314 nodes and 187edges) (**Figure 4**) the way as we did in creating the retweet one. But the source nodes were the screen names of users who quoted someone else's tweets, instead. The size of nodes located in the center were larger than those surrounding them, so those nodes (tweets) were being quoted more frequently than others. The quote graph included far less number of nodes than that of the retweet one. Thus, in this data set, less users quoted some else's tweets.

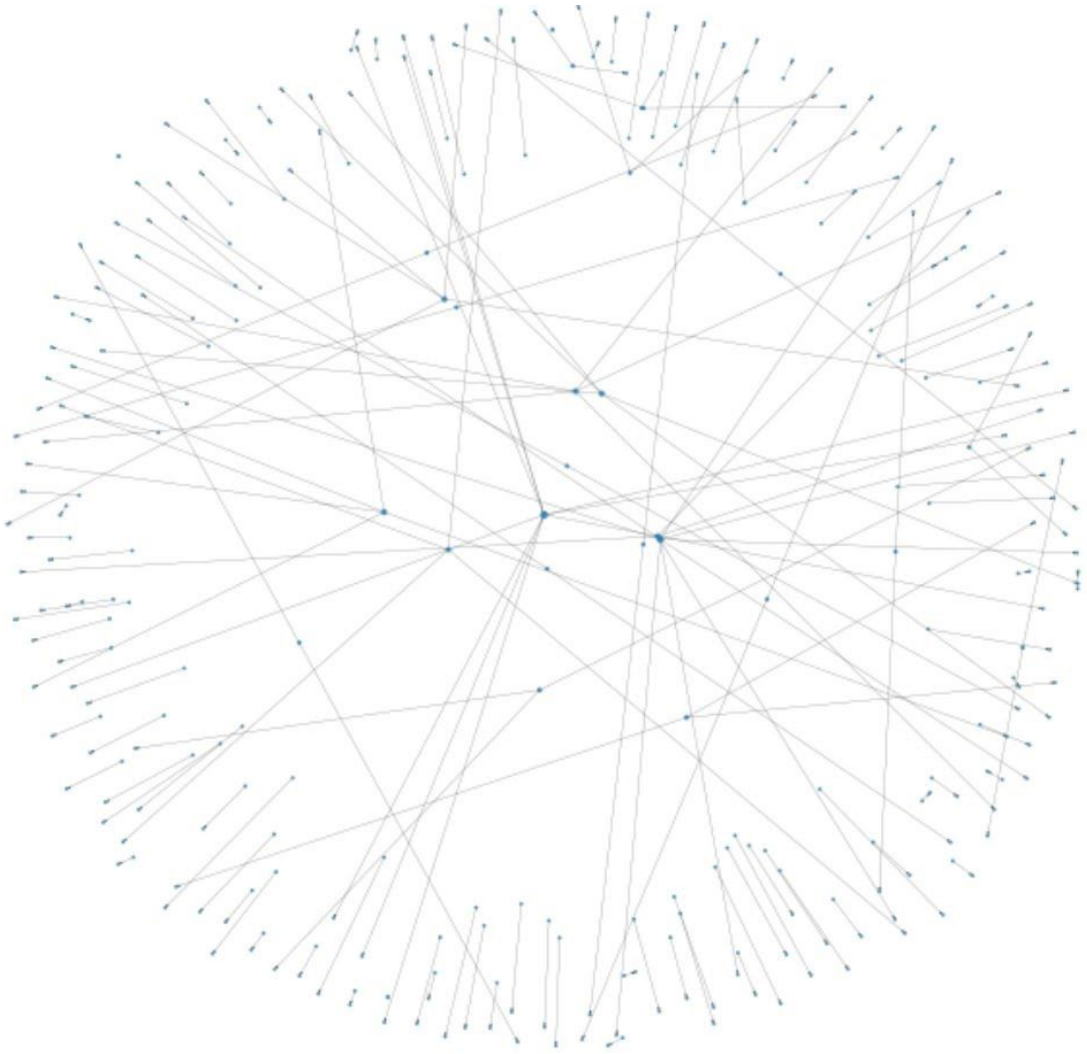

Figure 4

4 Network Measures Calculation

4.1 In-degree and Out-degree Centrality

We calculated the in-degree centrality and out-degree centrality for both the retweet directed graph and the quote one. For the retweet graph (**Figure 5**), the in-degree distribution was skewed to the right. The in-degree centralities of the majority range from 0.0000 to 0.0009, while a handful of those were above 0.0020. For the quote graph (**Figure 6)**, the in-degree distribution was pretty similar to that of the retweet one. From **Figure 7** and **Figure 8 ,** we observed similar distribution for out-degree centrality of retweet and quote network as well. Thus, the majority of tweets were retweeted or quoted far less often than a small proportion of them and it indicated that only a handful of tweets were dragging most users' attention.

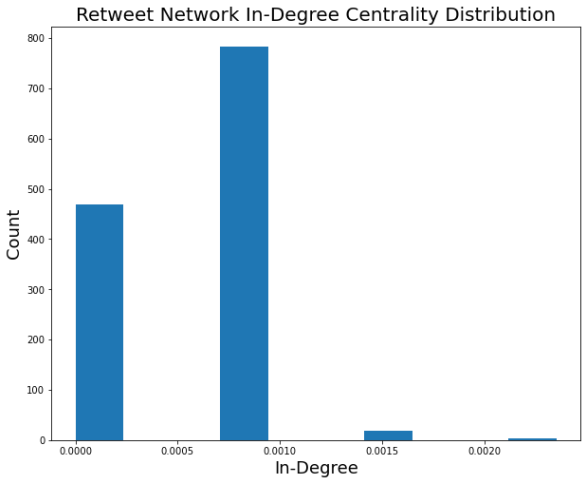

Figure 5

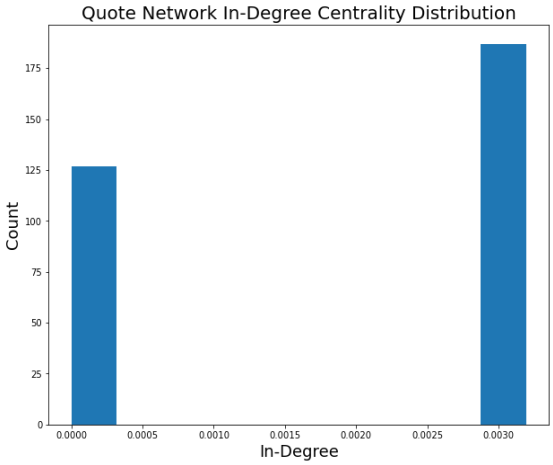

Figure 6

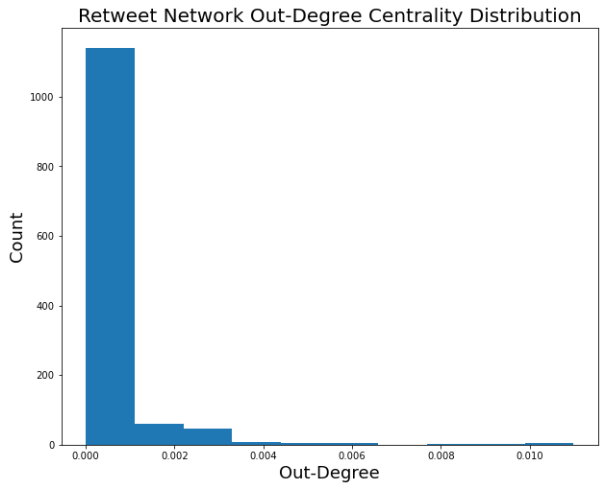

Figure 7

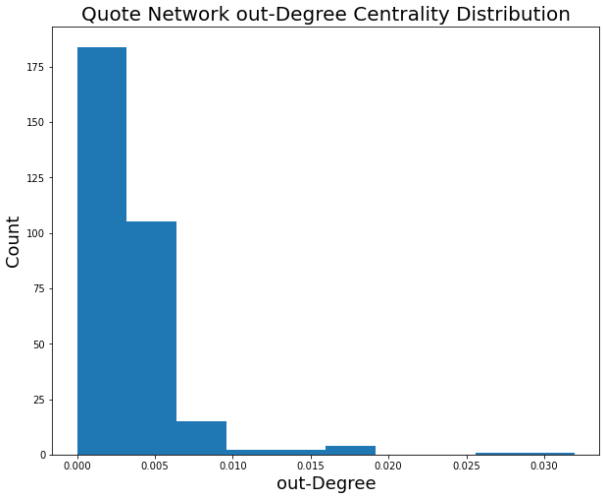

Figure 8

4.2 Betweenness-degree Centrality

We calculated the betweenness-degree centrality for the retweet and quote graphs. Based on the measurements shown (**Figure 9** and **Figure 10)**, there was only one node whose screen user name was “Chicago1Ray” connecting other nodes in the retweet graph and there was not such node in the quote graph. Thus, in this data set, there were almost noTwitter users retweeted or quoted other users’ tweets, meanwhile their tweets were not retweeted or quoted by others as well.

Betweenness Centrality in Retweet Network

	screen_name	betweenness_centrality
0	neal_katyal	0.000000
853	JuneMSanders	0.000000
852	ninetyniners	0.000000
851	karinacocoo	0.000000
850	superpenguin17	0.000000
...	...	...
420	fbatista_nyc	0.000000
419	tomselliott	0.000000
426	Parker1Erin	0.000000
1274	seamj11	0.000000
233	Chicago1Ray	0.000002

Figure 9

Betweenness Centrality in Quote Network

	screen_name	betweenness_centrality
0	CarlyELehwald	0.0
212	Adam_Van14	0.0
211	BradleyWhitford	0.0
210	sallyagale	0.0
209	ProjectLincoln	0.0
..	...	...
102	tatereeves	0.0
101	Deckard61S	0.0
100	DailyCaller	0.0
107	printingsharon	0.0
313	srsundevil	0.0

Figure 10

4.3 Eigenvector Centrality

Eigenvector centrality evaluates the importance of the neighbors (or incoming neighbors in directed graphs)[1]. A sample of measurements are as follows:

```
#calculate eigenvector centrality of the retweet network
centrality = nx.eigenvector_centrality(retweet_network)
values=[(node,round(centrality[node],4)) for node in centrality]
#display first 20 centrality values
values.sort(key=lambda x: x[1],reverse=True)
values[:20]
```

```
[('fireman452a', 0.7071),
 ('ASparklyWTF', 0.7071),
 ('neal_katyal', 0.0),
 ('JimmyA_Shook1s', 0.0),
 ('T_S_P_O_O_K_Y', 0.0),
 ('BAMAPERRY', 0.0),
 ('SenWarren', 0.0),
 ('tkdcoach', 0.0),
 ('donwinslow', 0.0),
 ('JamesQBulkhead', 0.0),
 ('maggiemilantoni', 0.0),
 ('vuocolo_lauren', 0.0),
 ('TrumpitC', 0.0),
 ('128JC3', 0.0),
 ('CREWcrew', 0.0),
 ('SuzanneSpsjess', 0.0),
 ('roper_93', 0.0),
 ('lap346', 0.0),
 ('TheRealHoarse', 0.0),
 ('Peterson99La', 0.0)]
```

Figure 11

From measurements shown in **Figure 11** , we can see that nodes in retweet_network have minimal or no incoming edges (pointing-in) edges. And hence the eigen centrality is 0 for most of the nodes in the network.

4.4 Katz Centrality

Katz centrality helps in correcting the problem of eigenvector centrality by adding a bias term[1]. The bias term β is added to the centrality values for all nodes no matter how they are situated in the network (irrespective of the network topology).

```
#calculate katz centrality of retweet network
centrality = nx.katz_centrality(retweet_network)
values=[(node,round(centrality[node],4)) for node in centrality]
#display first 20 centrality values
values.sort(key=lambda x: x[1],reverse=True)
values[:20]
```

```
[('mineisC', 0.0341),
 ('GREED_90', 0.0341),
 ('joespiv75259283', 0.0341),
 ('KamZenolay', 0.0315),
 ('DeeinMaryland', 0.0315),
 ('TheChoiceIsYou4', 0.0315),
 ('Lifeisee4me', 0.0315),
 ('drhug', 0.0315),
 ('JPC07', 0.0315),
 ('jilstereo', 0.0315),
 ('snoozjunkie', 0.0315),
 ('RandyTedford', 0.0315),
 ('sellersj17', 0.0315),
 ('TrollAxer', 0.0315),
 ('klizmiz1', 0.0315),
 ('Cindy_52s', 0.0315),
 ('thinkingthru2', 0.0315),
 ('MrE65034844', 0.0315),
 ('SuziAnnRyan', 0.0315),
 ('elvislver56', 0.0315)]
```

Figure 12

4.5 Hashtag Frequency

Out of the 1000 tweets, 80.2% of those had hashtag “#Trump” and only 25.2% had “Biden”. Thus, in those tweets, more twitter users mentioned Trump than those mentioned Biden.

5 Conclusion

The results of the analysis showed the pattern from Twitter users who tweeted or retweeted about one of the popular topics. This kind of analysis will also help in looking into how the tweets are connected to each other and find out the popular topics. As Twitter has been widely used by researchers in various fields, knowing what people are talking about online and getting insights from the tweets will help authorities understand what is happening and how people are reacting to a particular situation.

References:

[1]http://www.cs.iit.edu/~kshu/smm.pdf
[2]https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.drawing.nx_pylab.draw_networkx_edge_labels.html
[3]https://networkx.github.io/documentation/stable/reference/generated/networkx.convert_matrix.from_pandas_edgelist.html