Covid-19 pandemic on Twitter 2020

Network Science Final Project

Ngoc Diem Le
ngocdiem.le@studenti.unipd.it
2009466

1. Introduction

No one can deny the importance of the network in our daily life due to its application in many fields. In general, a network represents systems as sets of interactions or relations between various entities. In particular, when talking about social networks, the relationship between people is modeled. In terms of biology, the interactions among biological components are studied using network theory, and the Internet is considered the largest computer network in existence which has been and will continue to make positive contributions to the development of society. Network science is the study of the collection, management, analysis, interpretation, and representation of relational data[1]. Moreover, it can help to understand the complex systems that, at least regarding some specific phenomena, possess a well-defined, unique network structure or topology which can be a realization relation between a system and a network[2].

In 2020, the world faced an outbreak of the Covid-19 pandemic for the first time which has been causing heavy damage in many fields and taking many people's lives. Moreover, the development of technology and social media not only has brought many benefits in terms of keeping people safe, informed, and connected during the pandemic but also has facilitated the spread of misinformation.

To understand the issues related to the Covid-19 pandemic were concerned by people on social media, especially on Twitter, as well as which account has the most influence on the perception of the community, this project will focus on analyzing the networks by approaching network theory and natural language processing technique to analyze the tweets from Twitter. The intention is to evaluate the public concern related to the Covid-19 pandemic on Twitter for two months from March 2020 to May 2020.

The dataset was generated by using Twitter Application Programming Interface (API) with the hashtag #covid19. Due to the limitation of the number of collected tweets and also the research period is not long enough, only two months, the results obtained are not meaningful and not good enough to represent the whole community in general. There are three networks considered in this project which are words network for words extracted from tweets, hashtags network for hashtags used in tweets, and mentioned accounts network for the accounts were mentioned in tweets. For words used in tweets, the main topic people are concerned with is related to the Covid-19 pandemic which was the most important issue during that time and these words are "covid", "people", "new", "cases", "deaths", and so on. In terms of hashtags network, since the hashtag #covid19 was used to generate the dataset so that hashtags related to the Covid-19 pandemic were used most such as #coronavirus, #Covid_19, or #CoronavirusOutbreak together with hashtags like #China, #StayHome, or #SocialDistancing and so on. For the third network, the account that plays a key role to influence the entire network is @realDonaldTrump which is the official Twitter account of the former President of the United State Donald Trump.

Throughout this project, Python libraries like Numpy, Seaborn, Matplotlib, and WordCloud are used to process and visualize the data. Moreover, Python's Natural Language Toolkit (nltk) is used to extract meaningful words using in tweets and NetworkX is used to analyze the networks by calculating the values related to the properties of networks. Finally, to have a better view in terms of the structure of networks, Gephi is used to generate graphs.

2. Dataset

As mentioned in the previous part, this project used the dataset generated by the Twitter API with the hashtag #covid19 to analyze tweets related to the Covid19 pandemic. These tweets were collected using the Python script introduced during the second laboratory of the Network Science course. In addition, to better understand and explore the pattern of the network, the default language was set as English. However, due to the limitations of the Twitter developer account, only 100 tweets per day were generated from 15th March to 14th May in 2020 so that there are 6000 tweets in total. Moreover, the reason why choosing 15th March 2020 as the first day of collecting tweets is that due to the analysis on Google Trends, the word covid19 started rising dramatically on that day. Figure 1 shows trends of searching term covid19 and figure 2 below shows the information of the first 5 tweets.

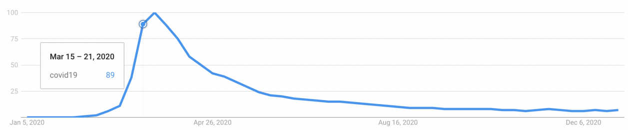

Figure 1. Google Trends of the word Covid19

	created_at	text	user	retweeted_status	entities
0	Sun Mar 15 23:59:59 +0000 2020	RT @danielsgoldman: FINAL UPDATE (thread): My #COVID19 test came back positive. \n\nI want to th...	{'id': 64090063, 'id_str': '64090063', 'name': 'Dan Maurer', 'screen_name': 'reruam', 'location'...	{'created_at': 'Sun Mar 15 17:26:13 +0000 2020', 'id': 1239241504226508802, 'id_str': '123924150...	{'hashtags': [{'text': 'COVID19', 'indices': [46, 54]}], 'urls': [], 'user_mentions': [{'screen_...
1	Sun Mar 15 23:59:59 +0000 2020	7) It's just a matter of time. The growth rate of #covid19 is really very high. \nLOCKDOWN #Indi...	{'id': 140772348, 'id_str': '140772348', 'name': 'Kranthi Thej Kandula', 'screen_name': 'kthej',...	NaN	{'hashtags': [{'text': 'covid19', 'indices': [50, 58]}, {'text': 'India', 'indices': [90, 96]}, ...
2	Sun Mar 15 23:59:59 +0000 2020	Do your patriotic duty\n- Stay home\n- Clean hands frequently w alcohol-based hand rub or soap &...	{'id': 17325149, 'id_str': '17325149', 'name': 'Saad Abedine 🤬😷🤟🏼', 'screen_name': 'SaadAbedine'...	NaN	{'hashtags': [], 'urls': [{'url': 'https://t.co/LDxqe2UJV1', 'expanded_url': 'https://twitter.co...
3	Sun Mar 15 23:59:59 +0000 2020	Even as supply chains face disruption in the near term from the #coronavirus disease (#COVID19),...	{'id': 2950739564, 'id_str': '2950739564', 'name': 'Connect Commercial Real Estate', 'screen_nam...	NaN	{'hashtags': [{'text': 'coronavirus', 'indices': [64, 76]}, {'text': 'COVID19', 'indices': [86, ...
4	Sun Mar 15 23:59:59 +0000 2020	RT @micheljessica_: Wasn't even scared until all the grocery stores started clearing out. I'm no...	{'id': 20729264, 'id_str': '20729264', 'name': 'Molly Hazelrig', 'screen_name': 'mollymorgana', ...	{'created_at': 'Sat Mar 14 07:21:35 +0000 2020', 'id': 1238726956180373504, 'id_str': '123872695...	{'hashtags': [], 'urls': [], 'user_mentions': [{'screen_name': 'micheljessica_', 'name': 'Jessic...

Figure 2. Overview of the tweets in 2020

For the preprocessing part, hashtags are directly extracted from the text column and mentions are extracted in user_mentions in entities column. In addition, tweets were cleaned by removing hashtags, mentions, hyperlinks, and punctuation and kept only meaningful words without stop words for the preparation of building words network. Moreover, the location of users was extracted for the creation of choropleth maps.

3. Exploratory Data Analysis

To discover the trends and patterns of the dataset, some visual techniques are used to create different types of graphs such as choropleth maps, pie charts, bar charts, and word cloud plots.

3.1. Number of tweets per country

The choropleth map 3 visualizes the number of tweets per country and 4 shows the top 5 countries with the highest number of tweets. As shown in figure 4, the US has a remarkably high number of tweets with 1901 tweets, followed by Canada and the UK with 378 and 276 respectively. It can be explained that these 3 countries use English as their primary language and the US is recognized as the country that has the most people on Twitter in recent years. Moreover, Nigeria appeared in the top 5 countries which is a surprise for an African country, whether this country plays an important role in the network, the answer may be found after analyzing the network.

3.2. Number of verified users

Figure 5 shows the proportion of verified users, there is only 4% of users are verified since verified users mostly are accounts of public interest, and given the fact that in

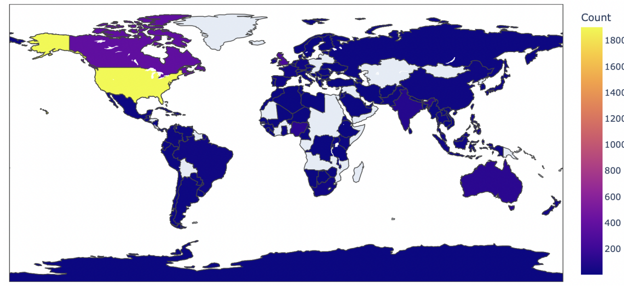

Figure 3. Number of tweets per country in 2020

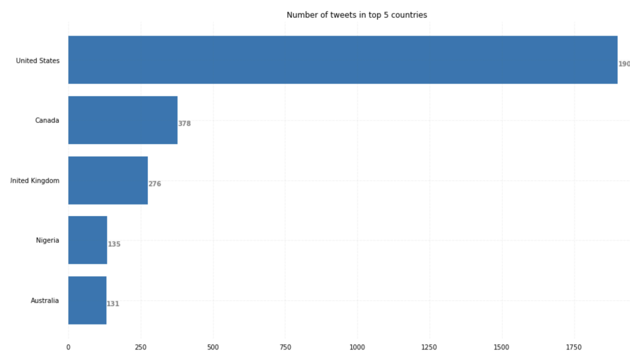

Figure 4. Top 5 countries with the highest number of tweets in 2020

order to be verified, an application is needed but it does not guarantee that everyone will get approved.

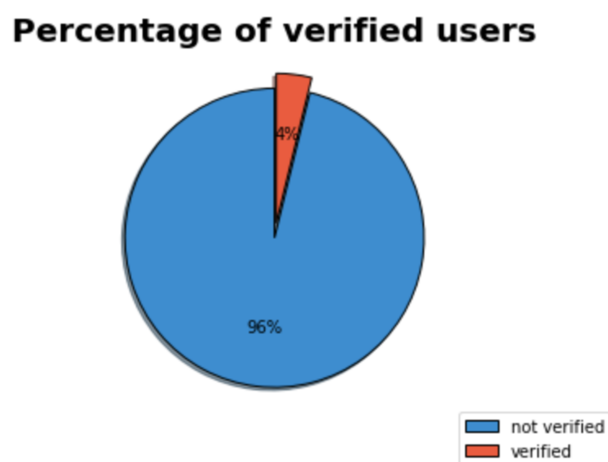

Figure 5.

3.3. Mentioned accounts

Based on the figure 6, the most mentioned account is the official account of the former president of the United States of America, Donald Trump. Next, the World Health Organization (WHO) and the Nigeria Centre for Disease Control (NCDC) hold the second and third place, followed by the account of Dr. Dena Grayson and the Centers for Disease

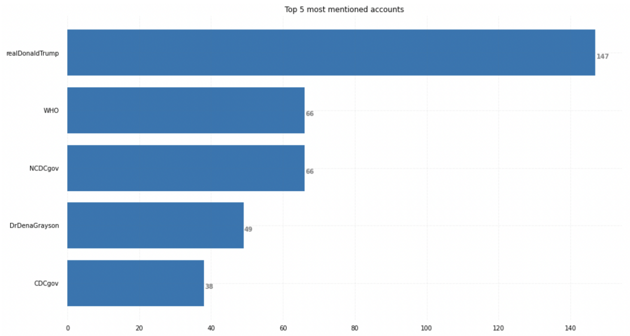

Figure 6. Top 5 Twitter's accounts were mentioned most

Control & Prevention (CDC).

3.4. WordCloud plots

Figure 7 and 8 visualize the frequency of words using in tweets. In figure 7, the five words that appear the most are "covid", "people", "cases", "new", and "us", meanwhile, in figure 8 the word "today" appears together with the words above without the word "covid".

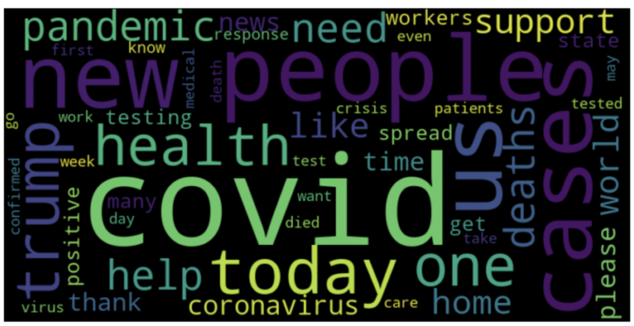

Figure 7. WordCloud for words frequency

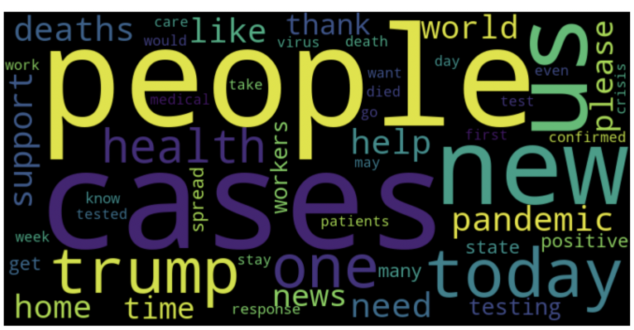

Figure 8. WordCloud for words frequency without "Covid"

Moving to the hashtags analysis, figure 9 shows that the hashtag #COVID19 is significantly used followed by #coronavirus. Since the hashtag #COVID19 accounted for the most used hashtags in tweets, by removing all of the hashtags related to the words "covid" and "coronavirus", the figure 10 visualizes other hashtags used most such as #China, #StayHome, #pandemic, #SocialDistancing, and #Trump.

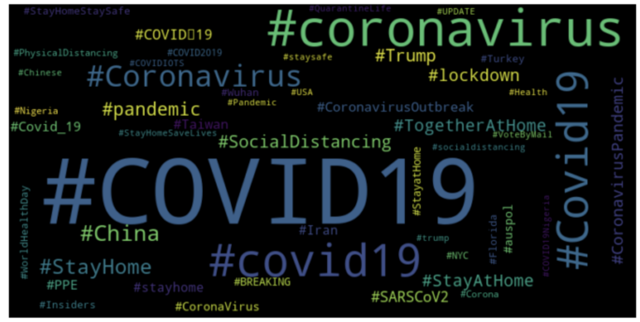

Figure 9. WordCloud for hashtags frequency

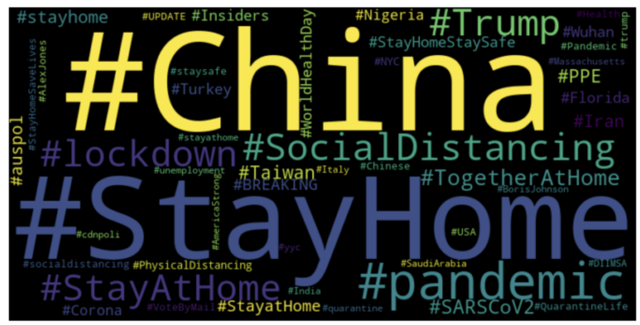

Figure 10. WordCloud for hashtags frequency without "Covid"

4. Network Analysis

After having the overview of the dataset from the previous part, three networks are created based on the collected tweets. In particular, they are words, hashtags, and mentioned accounts. Due to the number of hashtags related to Covid-19 was used such as #COVID19, #coronavirus, and other versions of these hashtags, if only analyzing these most used hashtags, it cannot be able to explore other topics people are also concerned about. Therefore, another network was created which is the hashtags network without #Covid19 and related hashtags, and for short, it is called the hashtags network without #Covid19. In the end, four networks are generated.

4.1. Words Network

4.1.1 Network Overview

For the first network, Gephi with layout ForceAtlas2 was used to visualizes the initial network with words extracted from tweets, shown in figure 11 and figure 12 are the degree distribution of the network. Based on the degree distribution figure, it can be seen that words network is a scale free network because its degree distribution follows the power law $p_k = Ck^{-\gamma}$. In addition, the figure 13 shows the distribution in log-log scale and also the complementary cumulative distribution with estimated $\gamma = 2.21869$ and $k_{min} = 50$. Moreover, the table 1 shows some basic characteristics of the network.

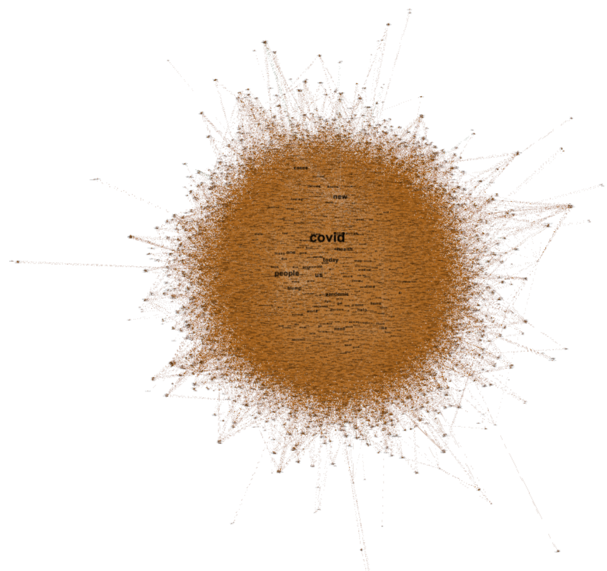

Figure 11. Initial network

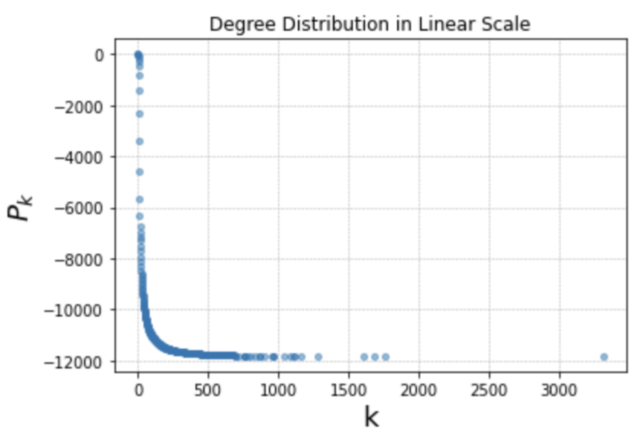

Figure 12. Degree Distribution in Linear scale

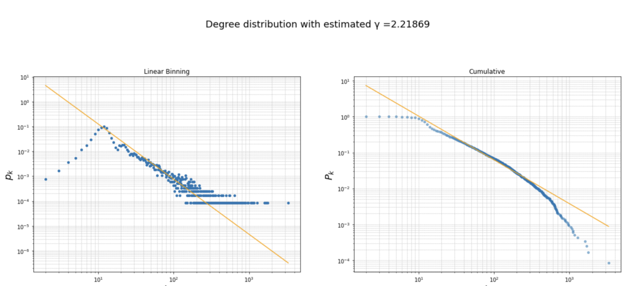

Figure 13. Degree Distribution

#Nodes	11823
#Edges	212882
Network Density	0.00304
Range of degree	2 - 3310
#Components	13
#Nodes in largest component	11808 (99.873%)
#Edges in largest component	212864 (99.992%)
Network Diameter of largest component	5
Average clustering coefficient	0.73437
Transitivity	0.097

Table 1. Words network properties

Network density is defined as the proportion of potential ties that are present in the network compared to how connected it might be. Based on the values in the table 1, the network density is 0.00304. In addition, the number of components in the network is 13 and there are 99.873% of nodes belong to the largest component. In terms of the largest component, the diameter of this component is 5 means that the longest path of the shortest paths between any two nodes is 5.

Figure 14 shows the frequency count of the occurrence of each clustering coefficient. A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together, the value is equal to 0 means that there are hardly any connections in the neighborhood of a node and if the value is equal to 1 means that the neighborhood of that node is fully connected. It can be seen in the figure 14 that a huge number of coefficients belong to the range of 0.9 and 1.0 and there is 9.98% of nodes have the clustering coefficients equal to 1 which means that most of the neighbors of nodes are well connected with the average is 0.73437. Based on the table 1, there is a contrast between the average clustering coefficient and transitivity. In particular, because there are a lot of nodes have 1 as the clustering coefficient so that the average clustering coefficient is pretty high, meanwhile, transitivity is a measure of the percentage of open triads that are triangles, and since there are a lot of open triads which do not form triangles so the transitivity is low, only 0.097.

Moreover, figure 15 visualizes the distribution of clustering coefficients according to the degree centrality. It can be easily seen that the higher the clustering coefficient, the less degree centrality is. Node with the highest degree centrality may have the lowest clustering coefficient. Based on figure 15, node with highest degree centrality is the word ”covid”, this node has clustering coefficient is 0.015871, followed by ”people”, ”new”, ”us”, and ”today”. Furthermore, communities may be connected by large degree nodes, as seen in figure 16, ”covid” and ”people” which are two nodes with the highest degree centrality that belong to their communities and being in the same community means that they

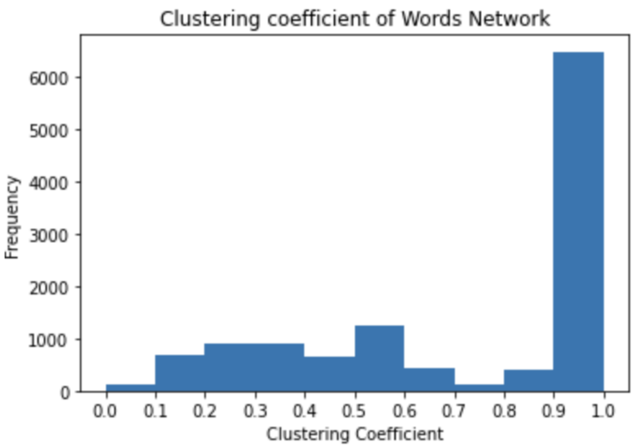

Figure 14. Distribution of Clustering coefficient

are sharing mostly neighbors within the same community and the tendency to form a triangle with other neighbors is small, therefore these clustering coefficients are quite small despite the degree of them are very high.

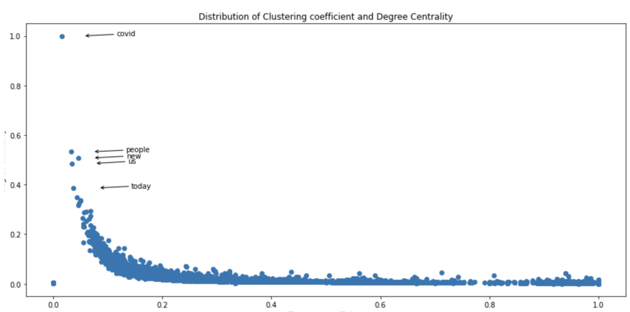

Figure 15. Distribution of Clustering coefficient vs Degree centrality

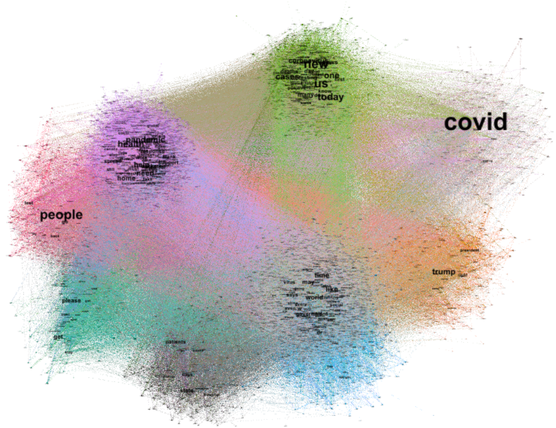

Figure 16. Communities of words network

4.1.2 Centrality

In network analytics, centrality plays an important role in identifying crucial nodes in a graph. Each node could be important from an angle depending on how ”importance” is defined. Figure 17 shows the comparison between four centrality measures: Degree centrality, Betweenness centrality, Closeness centrality, and Eigenvector centrality. In a non-directed graph, the degree of a node is defined as the number of direct connections a node has with other nodes and the degree centrality metric defines the importance of a node in a graph as being measured based on its degree which means that the higher the degree of a node, the more important it is in a graph. Based on figure 17, the most five important words in terms of degree centrality are ”covid”, ”people”, ”new”, ”us”, and ”today”. Secondly, talking about betweenness centrality, this metric defines and measures the importance of a node in a network based on how many times it occurs in the shortest path between all pairs of nodes in a graph. Not surprisingly, the five words that have the highest betweenness centrality are the same with degree centrality. Next, the closeness centrality metric defines the importance of a node in a graph as being measured by how close it is to all other nodes in the graph. Particularly, it is defined as the sum of the geodesic distance between that node to all other nodes in the network. Based on figure 17, the most five important words in terms of closeness centrality are ”pa7”, ”anime”, ”emojis”, ”describe”, and ”necker”. Figure 18 visualizes the words network based on closeness centrality values, it can be seen that all words in the top 5 belong to their components so that the distance of these nodes to all other nodes in the same graph is smallest. After counting all nodes which have 1 as the value of closeness centrality, there are only 6 nodes accounted for 0.05%, these nodes are colored by green in figure 18. Finally, eigenvector centrality measures the importance of a node in a graph as a function of the importance of its neighbors. In particular, if a node is connected to highly important nodes, it will have a higher eigenvector centrality score as compared to a node that is connected to less important nodes. Once again, figure 17 shows that the most five important words in terms of eigenvector centrality are ”covid”, ”people”, ”new”, ”us”, and ”today”, the same with degree centrality and betweenness centrality measures.

Figure 19 visualizes the relationship between degree centrality and PageRank of the words network. It can be seen that they are almost linearly correlated therefore using PageRank to analyze the network does not show much meaning. That is the reason why PageRank is not considered in the figure 17.

4.2. Hashtags Network

Table 2 shows some basic information related to hashtags network. In this network, nodes represent the hashtags

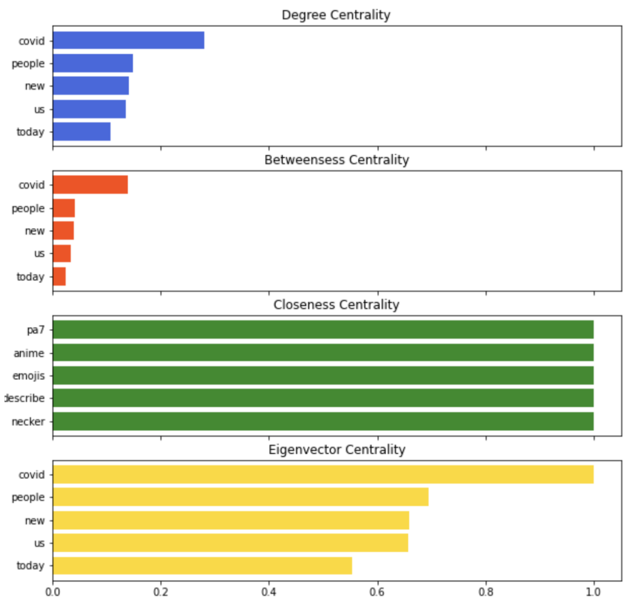

Figure 17. Centrality Comparison in Words network

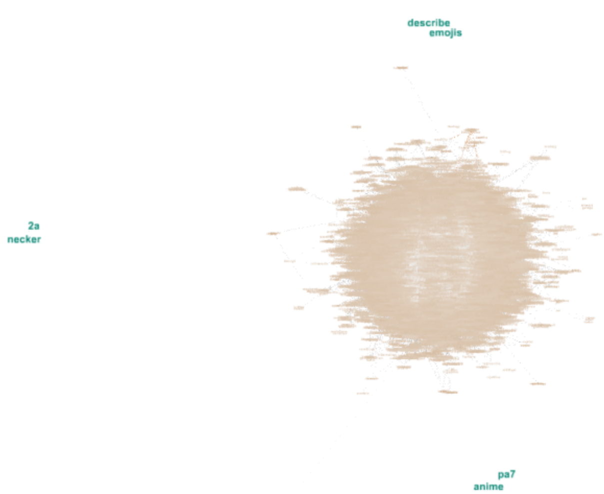

Figure 18. Closeness Centrality of Words Network

used in collected tweets. In terms of network density, the hashtags network has the same value as the words network which is 0.0034. Figure 20 visualizes the initial network using Gephi with layout Yifan Hu and the size of the node label is the degree of that node. It can be seen that the network has a very dense subnetwork in the center and a lot of sparse subnetworks outside.

Moreover, there are 329 connected components in this network and the figure 21 visualizes the largest components which accounted for 73.547% of the total nodes of the network with the size of the labels representing the degree of the nodes.

To analyze how connected the network is, clustering co-

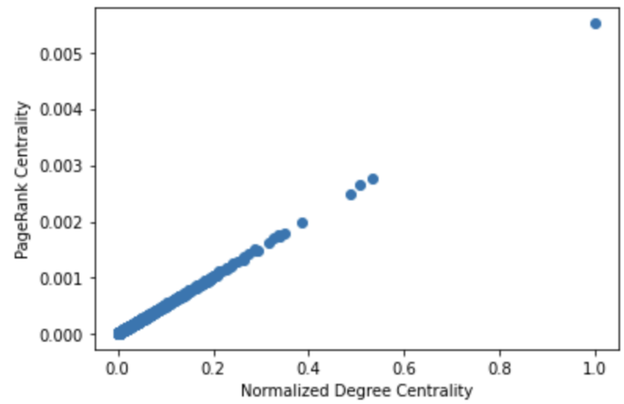

Figure 19. Relationship between Degree Centrality and PageRank in Words Network

#Nodes	1652
#Edges	4634
Network Density	0.0034
Range of degree	2 - 812
#Components	329
#Nodes in largest component	1215 (73.547%)
#Edges in largest component	4040 (87.182%)
Network Diameter of largest component	7
Average clustering coefficient	0.5465
Transitivity	0.0978

Table 2. Hashtags network properties

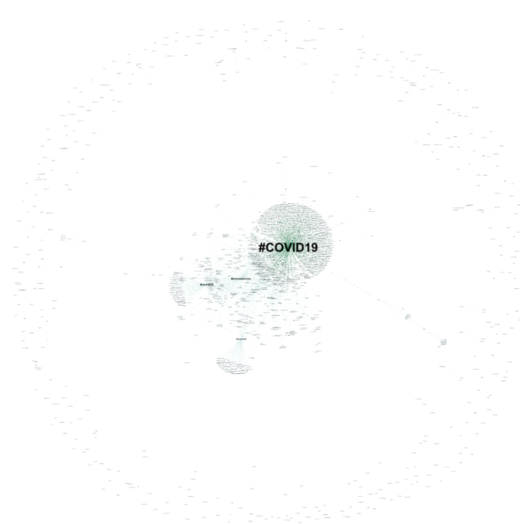

Figure 20. Hashtags network

efficients are calculated and the figure 22 shows the frequency of the clustering coefficient in this network. The proportion of clustering coefficients equal to 0 is 41.16% and clustering coefficients equal to 1 is 50.48%. These results show that despite the network having many neighbors

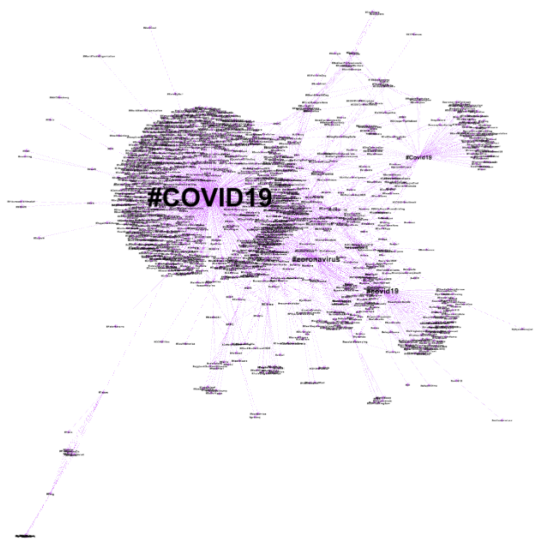

Figure 21. Largest component of Hashtags network

which are close together but there are still a lot of neighbors which are not connected at all as can be easily seen in figure 20.

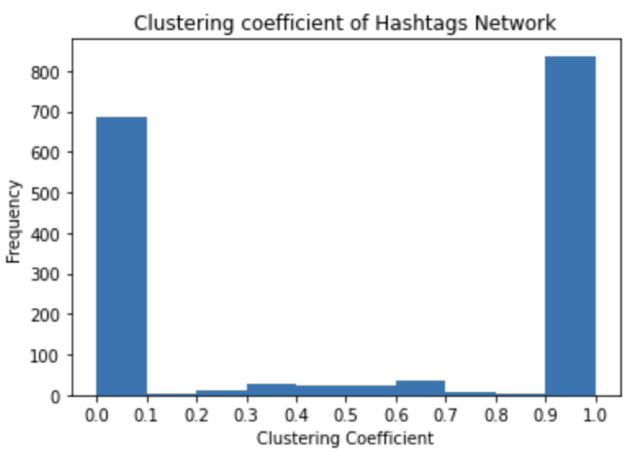

Figure 22. Distribution of Clustering coefficient

In terms of centrality, figure 23 shows the comparison between 4 types of considered centrality, degree, betweenness, closeness, and eigenvector centrality. In this case, there is no significant difference between these types.

Finally, figure 24 visualizes the network communities by using modularity. The algorithm turned out that there are 351 communities which are indicated by different colors in the graph.

4.3. Hashtags Network without #Covid19

In this section, the list of initial hashtags collected from tweets decreased by removing all of the hashtags related to the words "covid" and "coronavirus" to have a more

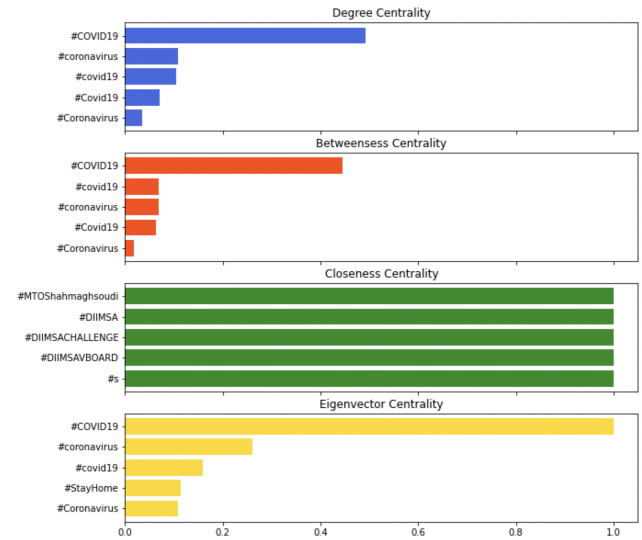

Figure 23. Centrality Comparison in Hashtags network

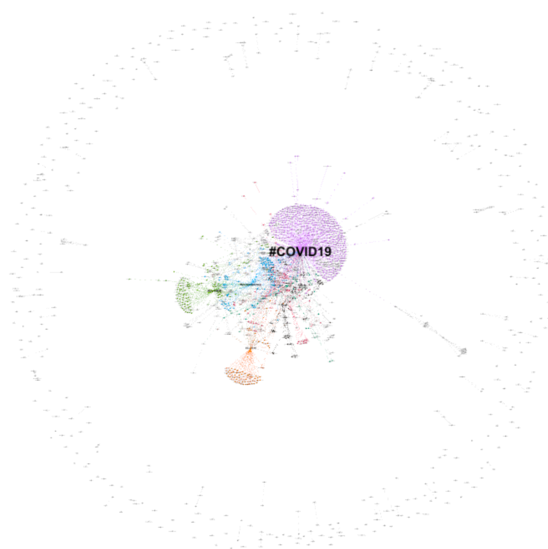

Figure 24. Community detection

objective view of this network. Table 3 shows some basic properties of this network. The number of nodes and edges declined and the network density is slightly dropped to 0.0024. Furthermore, the highest degree dramatically reduced to 45 instead of 812 as the initial network which shows the importance of the hashtags related to "covid" and "coronavirus" in the network when connecting other hashtags. By contrast, the number of components remarkably increase to 813. Since the network now is more sparse so that the number of nodes in the largest component fell to 15.879% compared to 73.547% of the initial network. The figure 25 visualizes the largest component of the network with label size representing the degree of the node.

#Nodes	1560
#Edges	2992
Network Density	0.0024
Range of degree	2 - 45
#Components	813
#Nodes in largest component	248 (15.897%)
#Edges in largest component	774 (25.869%)
Network Diameter of largest component	12
Average clustering coefficient	0.364
Transitivity	0.5822

Table 3. Hashtags network without #Covid19 properties

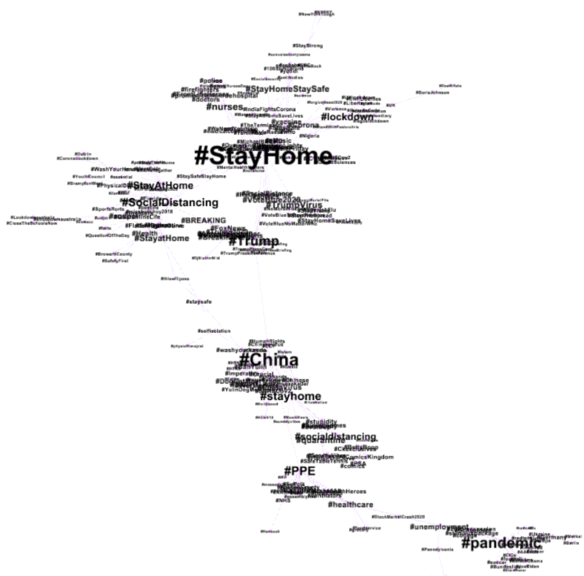

Figure 25. Largest component of Hashtags network without #Covid

Figure 26 below shows the frequency of clustering coefficient of the hashtags network without having hashtags related to ”covid” and ”coronavirus”. It can be seen that the distribution is different from the graph of the initial network. The frequency of coefficients equal to 0 is significantly higher than those coefficients equal to 1, 60.77% and 34.17% respectively.

Figure 27 compares between four types of centrality. Excluded closeness centrality, the three other values perform quite similarly with #StayHome playing an important role in this network, followed by #China and #Trump. In terms of eigenvector centrality, #Dog appears for the first time this is also an important node in this network. Considering closeness centrality, it can be seen that there is 35.17% of nodes have 1 as the closeness centrality. The figures 28, 29, and 30 visualize the network based on degree centrality, betweenness centrality, and eigenvector centrality respectively with the colors of label indicate the communities they belong and the sizes of label indicate the centrality measure

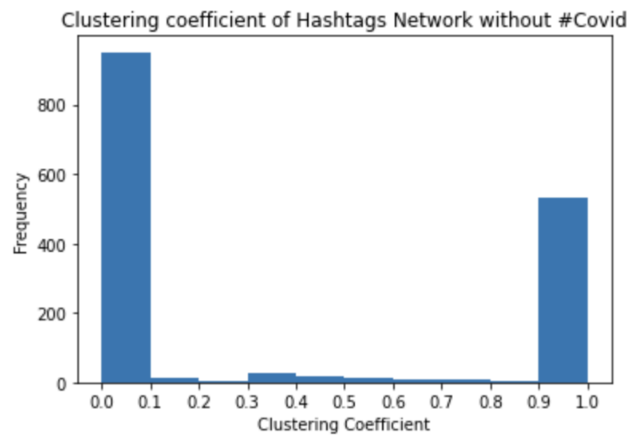

Figure 26. Distribution of Clustering coefficient

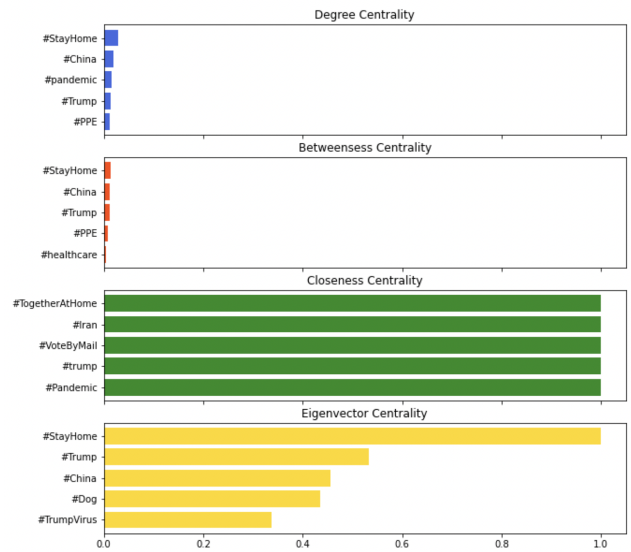

Figure 27. Centrality Comparison in Hashtags network without #Covid

considered.

4.4. Mentioned accounts network

In this section, the final network is introduced, this network is related to Twitter’s accounts which are mentioned the most in collected tweets. First of all, basic properties are shown in table 4 and figure 31 visualizes the network with the label of a node representing the degree of that node. Once again, this network is sparse with a significantly huge number of components, 2762. In particular, figure 32 gives a better view of the largest component of this network.

For the clustering coefficient, figure 33 visualizes the frequency of clustering coefficients in this network. In particular, the percentage of clustering coefficients equal to 0 accounted for the most with 77.15% while there is only 19.04% for those coefficients equal to 1.

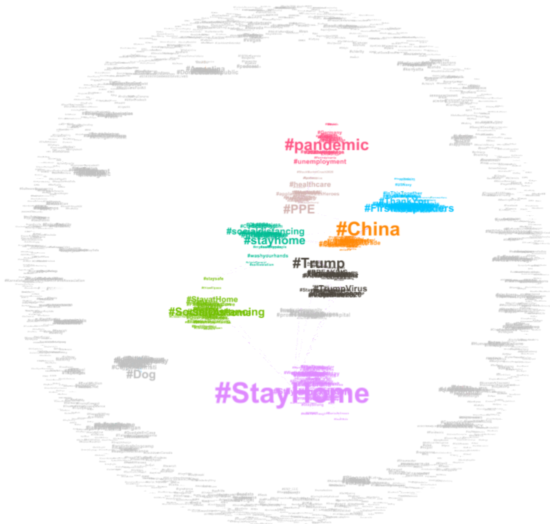

Figure 28. Degree Centrality of Hashtags network without #Covid

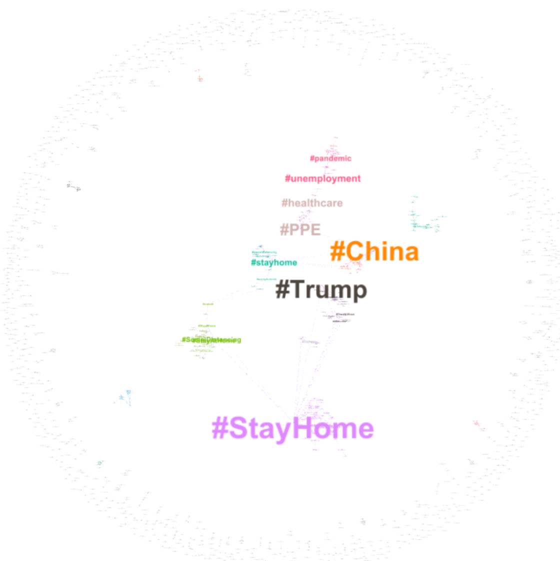

Figure 29. Betweenness Centrality of Hashtags network without #Covid

For centrality comparison, the figure 34 visualizes the top 5 accounts that have the highest values in terms of centrality measure. In general, the figure shows that the values that measure the centrality of mentioned accounts network are not very high. It can be easily seen that @realDonaldTrump is the most important, the most influential account because it is accounted for the first place in 3 out of 4 criteria, followed by @CDCgov which always appears in the top 5 excluded closeness centrality. Moreover, compared to the top 5 accounts with the highest mentioned frequency which are reported in the EDA part, @NCDCgov

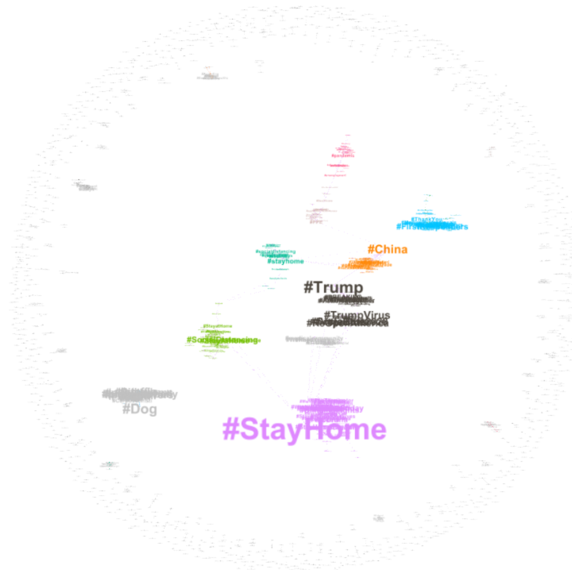

Figure 30. Eigenvector Centrality of Hashtags network without #Covid

#Nodes	4249
#Edges	6690
Network Density	0.0007
Range of degree	2 - 110
#Components	2762
#Nodes in largest component	457 (10.755%)
#Edges in largest component	1246 (18.625%)
Network Diameter of largest component	12
Average clustering coefficient	0.204
Transitivity	0.3664

Table 4. Mentioned accounts network properties

and @DrDenaGrayson disappear in the centrality comparison.

Considering degree centrality, in addition to @realDonaldTrump, @WHO, and @CDCgov, @NYGovCuomo (The official administration account of Governor Andrew M. Cuomo which is now archived) and @nytimes (The New York Times) also play an equally important role in terms of directed neighbors these nodes have in the graph.

The betweenness centrality is one of the most common measures used in network analysis. Since network density affects betweenness centrality in some cases, particularly, the network density is only 0.0007 which is very small so the values of betweenness centrality are also very small. Based on the comparison graph 34, the top 5 entities that have the most control over the flow and circulation of information are @realDonaldTrump, @WHO, @CDCgov, @POTUS (the official account of President Biden), and @AamerAnwar (A Scottish political activist and Lawyer to

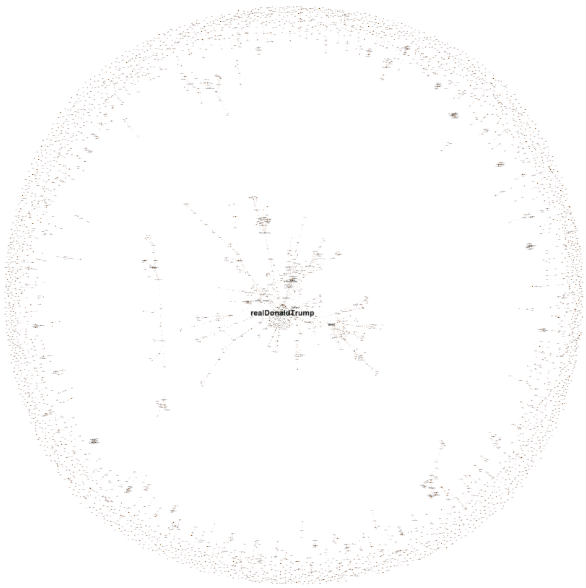

Figure 31. Mentioned accounts network

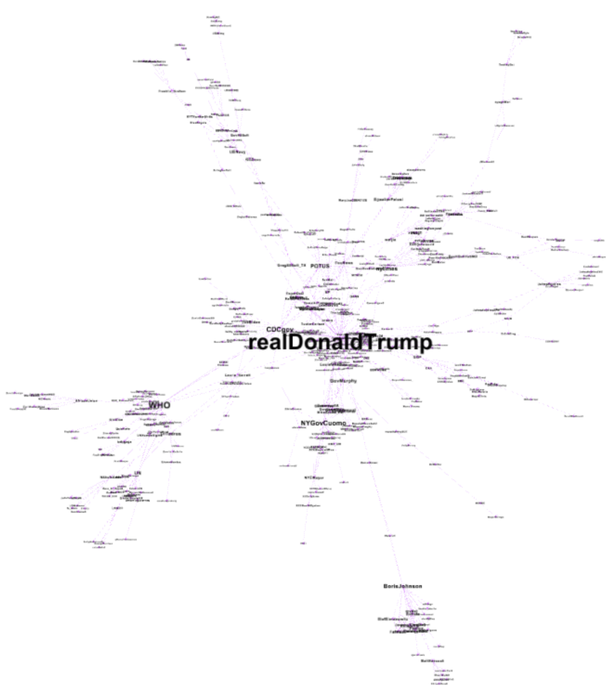

Figure 32. Largest component of mentioned accounts network

COVID19 Bereaved Family Group).

Furthermore, considering closeness centrality, there is 29.75% of the accounts have 1 as the value of closeness centrality. It can be seen in figure 31, the network is pretty sparse with lots of components so that it is easy to find many nodes which influence the network, in particular, their components, most quickly as in the words network and hashtags network without #Covid.

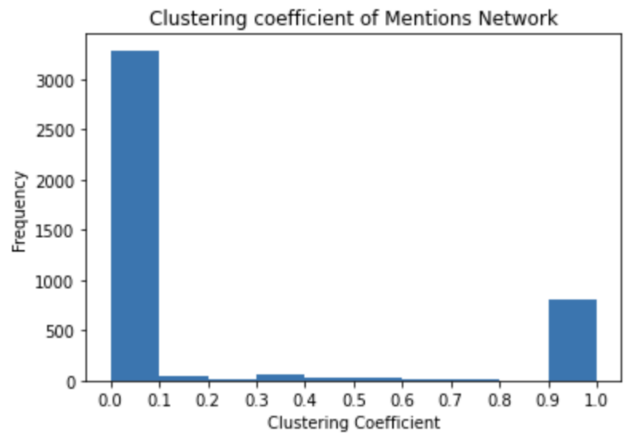

Figure 33. Distribution of Clustering coefficient

Finally, in terms of eigenvector centrality, the top 5 are @realDonaldTrump, @CDCgov, @POTUS, @WhiteHouse, and @NYGovCuomo. There is a tendency that nodes with a higher degree centrality may have also a higher eigenvector centrality, that is the reason why @realDonaldTrump is in the first place since this account has strongly influential ties.

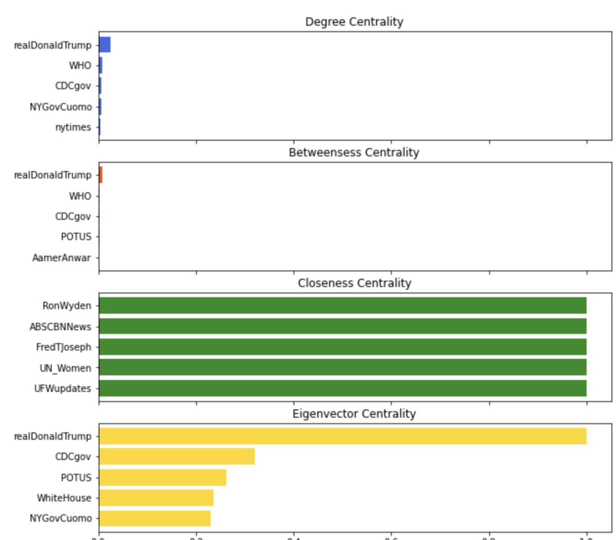

Figure 34. Centrality Comparison in mentioned accounts network

Figures 35, 36, and 37 visualize the network based on degree centrality, betweenness centrality, and eigenvector centrality respectively with the colors of label indicate the communities they belong and the sizes of the label indicate the centrality measure considered.

5. Conclusions

In conclusion, this project aimed to analyze tweets from Twitter accounts with a specific hashtag #covid19 to dis-

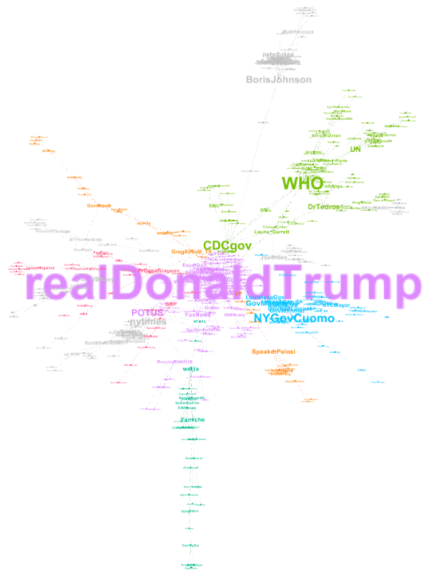

Figure 35. Degree centrality in mentioned accounts network

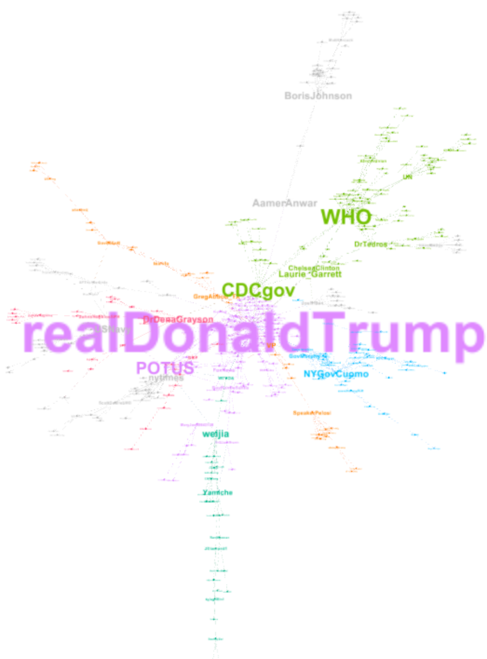

Figure 36. Betweenness centrality in mentioned accounts network

cover what topics people were concerned about and which account influenced most in two months in 2020 from 15th March to 14th May. Since most of the collected tweets are from the US, the official account of the former president Donald Trump plays a key role, and the topics related to the Covid-19 pandemic like #StayHome, #SocialDistancing, #China, and #lockdown were concerned the most due to

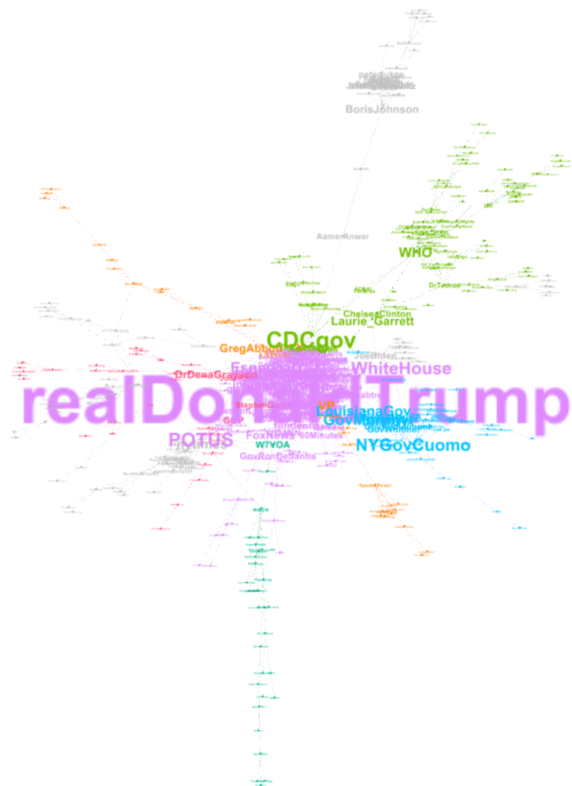

Figure 37. Eigenvector centrality in mentioned accounts network

the importance of these hashtags in the network. Moreover, comparing exploratory data analysis and network analysis, there is a slight difference between the two parts. In particular, 2 of the 5 accounts (@NCDCgov and @DrDenaGrayson) with the highest frequency do not appear in the top 5 accounts in the network analysis, which indicates that the node with higher frequency may have no effect in the network analysis. Furthermore, despite a significantly high occurrence of tweets in Nigeria and also the number of times @NCDCgov was mentioned, the topics related to this country did not influence the entire network in general.

For future work, I would love to collect more tweets in terms of the quantity and length of time since the information extracted from 6000 tweets in 2 months is not enough to generalize the issue of social interest. Furthermore, sentiment analysis can be performed in order to evaluate the public opinion related to the Covid-19 pandemic, and also compared with network analysis, we can have a broader and more detailed look at this issue. In addition, to see how topics and influence people are changed over time, collecting tweets from 2 different years can be a better option in terms of comparison.

References

[1] ULRIK BRANDES, GARRY ROBINS, ANN McCRANIE, and STANLEY WASSERMAN. What is network science? *Network Science*, 1(1):1–15, 2013.

[2] Gábor Elek and Eszter Babarczy. Taming vagueness: the philosophy of network science. *Synthese*, 200(2):1–31, 2022.