UCS1722- Social Network Analysis

Analysis of COVID-19 Outbreak Using Graph Theory

Swetha Saseendran (185001183)

Swetha P (185001182)

Tarun V (185001184)

Vishal T (185001199)

Table of Content

Abstract

This report explains the epidemic spread using social network analysis, based on the given data. A network is defined and visualization is used to understand the spread of coronavirus among countries and the impact of other countries on the spread of coronavirus. The purpose of this project is to represent the COVID-19 spreading around the World using graph / networks theory. Then, we will study the properties of such a graph. We can use graph / network theory to study the propagation of the virus. It may be useful to understand how the countries should act in case of a new pandemic. We also used hierarchical clustering model to find transmission clusters from the network dataset by clustering based on the dendrogram created using the adjacent matrix of the network.

Keywords: Social Network Analysis, COVID-19, Community Detection, Clustering

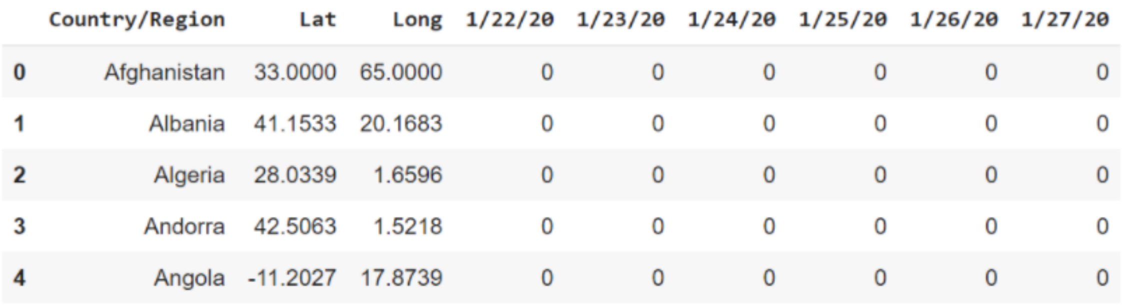
Materials and Tools

Dataset and Pre-Processing

From the World Health Organization - On 31 December 2019, WHO was alerted to several cases of pneumonia in Wuhan City, Hubei Province of China. The virus did not match any other known virus. This raised concern because when a virus is new, we do not know how it affects people. So daily level information on the affected people can give some interesting insights when it is made available to the broader data science community. This dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. The data is available from 22 Jan, 2020.

SNo	Serial number
Observation Date	Date of the observation in MM/DD/YYYY and death toll on that day
Lat	Latitude Value
Long	Longitude Value
Country/Region	Country Names

Table 1: Column Description



	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20
0	Afghanistan	33.0000	65.0000	0	0	0	0	0	0
1	Albania	41.1533	20.1683	0	0	0	0	0	0
2	Algeria	28.0339	1.6596	0	0	0	0	0	0
3	Andorra	42.5063	1.5218	0	0	0	0	0	0
4	Angola	-11.2027	17.8739	0	0	0	0	0	0

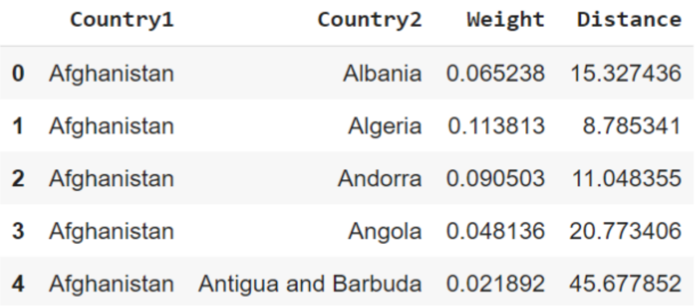
Table 2: Original Dataset

We first created a date vector which consists of all the deaths on a particular date.That is: $x_i = (d_1, d_2,d_3,..........,d_n)$ where x_i is each country and d_i is the death on date i. For each pair combination of the countries, we define weight to be:

$$W_{ij} = (\left|x_i - x_j\right| + \epsilon)^{-1}$$

We take the inverse of the difference between the mean times. The epsilon factor corrects the fact that the difference may be zero. In that case, the nodes will be strongly linked since the epsilon factor is a small number (0.001). We can consider that the inverse of the edge weight is a kind of distance between the nodes. The distance (for each edge) between each pair of country was defined to be:

$$D_{ij} = W_{ij}^{-1} = \left|x_i - x_j\right| + \epsilon$$



	Country1	Country2	Weight	Distance
0	Afghanistan	Albania	0.065238	15.327436
1	Afghanistan	Algeria	0.113813	8.785341
2	Afghanistan	Andorra	0.090503	11.048355
3	Afghanistan	Angola	0.048136	20.773406
4	Afghanistan	Antigua and Barbuda	0.021892	45.677852

Table 3: Graph Network Dataset

Graph Visualisation and Tools Used

The graph was generated using the Gephi editor. Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free. It's true that we could create the graph using the NetworkX python library but it's easier to use the Gephi software and, also, it's more complete and with it's customization we were able to get good insights about the structure of the COVID-19 propagation network.

However, we also used the NetworkX library for analysing some of the Graph metrics like Edge Weight Distribution and Distance Distribution. We also used Hierarchical Clustering using scipy library to form several clusters and find who had a stronger influence in spreading COVID among the cluster. We used agglomerative clustering which is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.

Results

The following are the results obtained from the network dataset we generated. We visualised the network, and analysed several graph metrics to draw inferences on the nature of the network to analyse the community spread. We also performed hierarchical clustering using the dentogram formed using the adjacency matrix of the network to detect different community spread and different transmission groups.

Network Visualisation

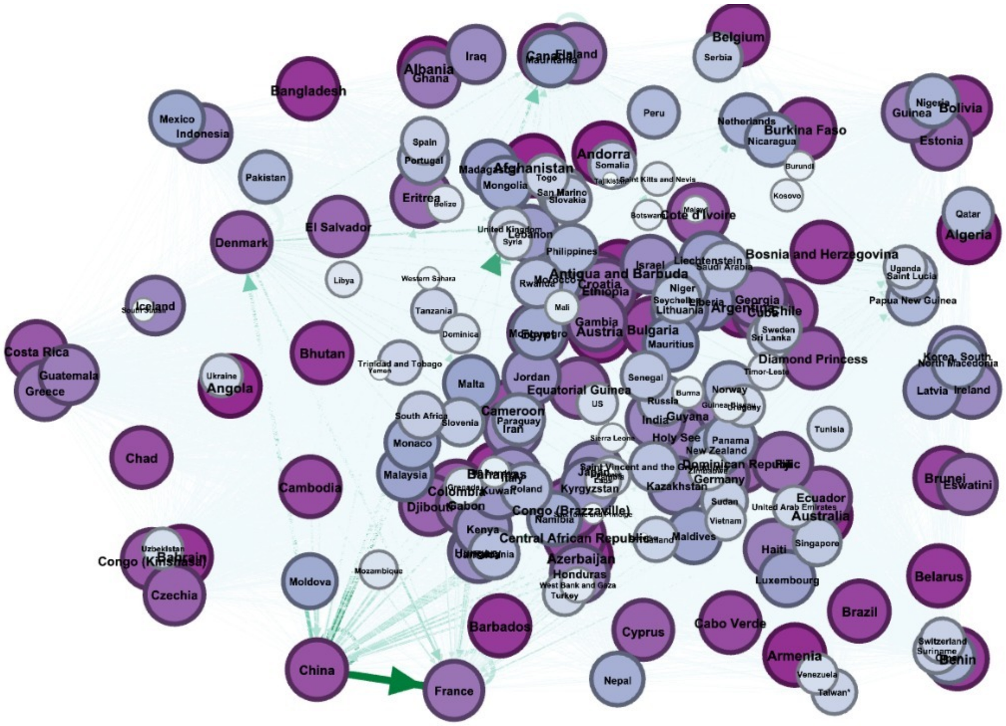


Figure 1: Network Visualised

The outward edge of China is thicker than other edges as it has more weight since it's impact on spreading COVID was more which is justified since, it's known that COVID originated in China.

Outdegree is the number of outward directed graph edges from a given graph vertex in a directed graph. The size of the nodes, their colors and the colors of the edges are directly related to the outdegree of the node, where the weight is bigger if the mean time of occurrences are similar. So, we have the following figure:

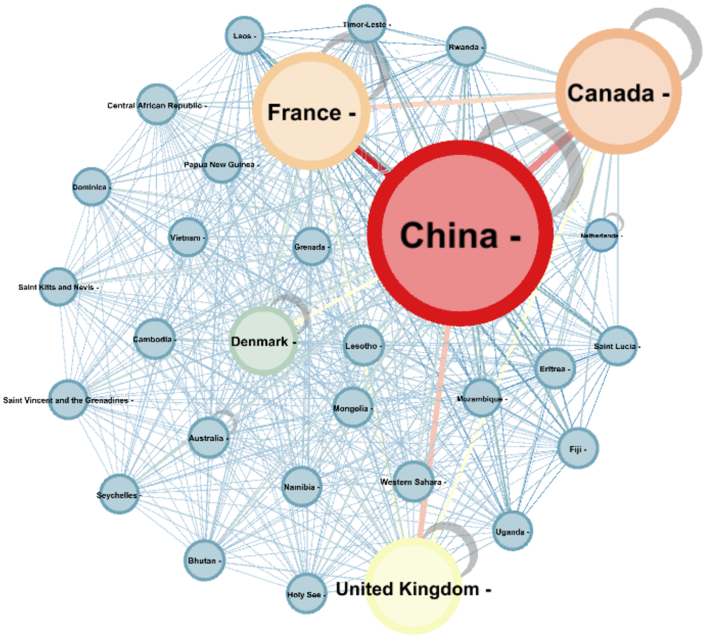


Figure 2: Network Visualisation based on the outdegree

The graph also shows us that it seems that we have a propagation following the path "China - Europe - Africa - America ". Those countries which have higher out-degree consider as the main countries where many patients came from. Patients of these countries have traveled to many countries. So those countries can be considered as the main hubs of COVID-19.

There is a strong link between Canada and France. They are connected, which makes sense when we think that there is a great cultural proximity between Quebec and France, which may interfere in the flights and in the personal contacts among different kinds of people.

Graph Metrics- Degree

The degree distribution is the probability distribution of these degrees over the whole network.

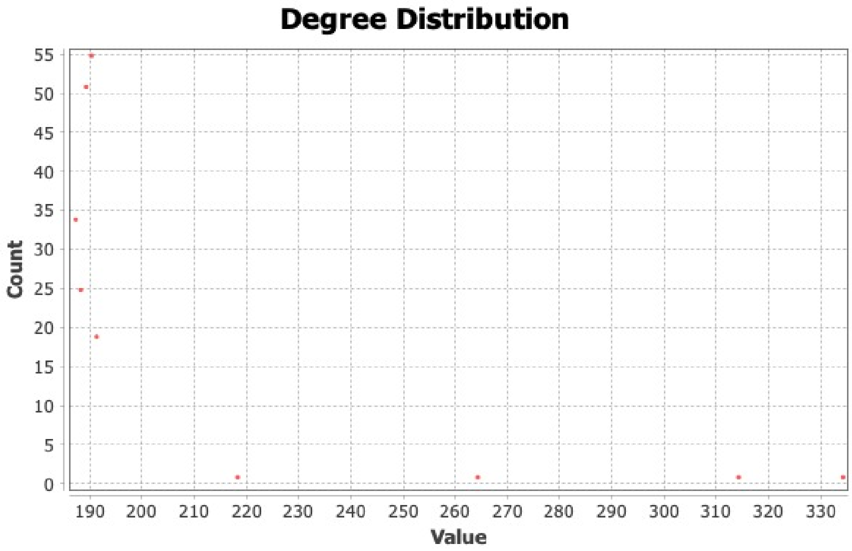


Figure 3: Degree Distribution

Total number of leaving vertices is known as **outdegree** and the total number of entering vertices is known as **indegree**.

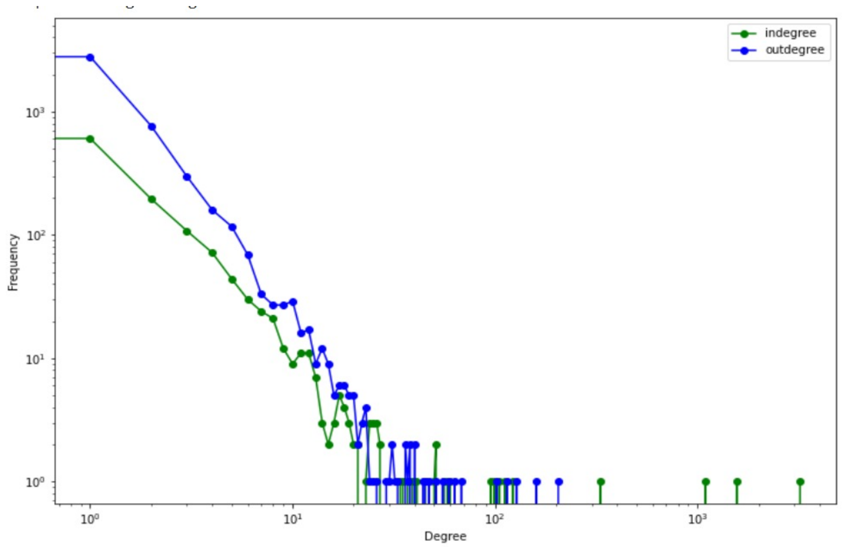


Figure 4: In-Degree and Out-Degree Distribution

Graph Distance Report

Betweenness centrality measures the number of times a node lies on the shortest path between other nodes. A high betweenness count could indicate someone holds authority over disparate clusters in a network, or just that they are on the periphery of both clusters

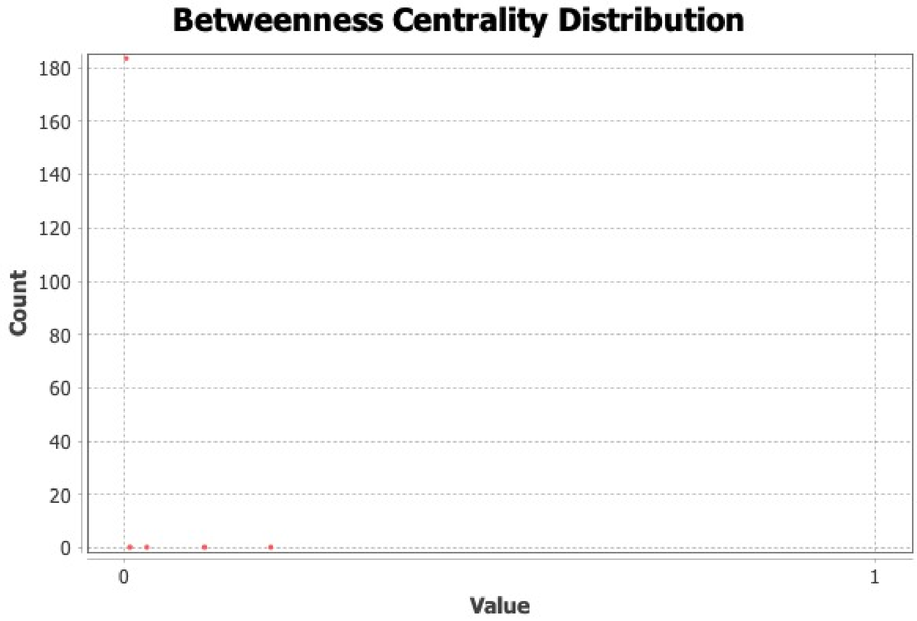


Figure 5: Betweenness Centrality Distribution

Closeness centrality is a measure of the average shortest distance from each vertex to each other vertex. Closeness centrality is a way of detecting nodes that are able to spread information very efficiently through a graph. The closeness centrality of a node measures its average farness (inverse distance) to all other nodes. Nodes with a high closeness score have the shortest distances to all other nodes.

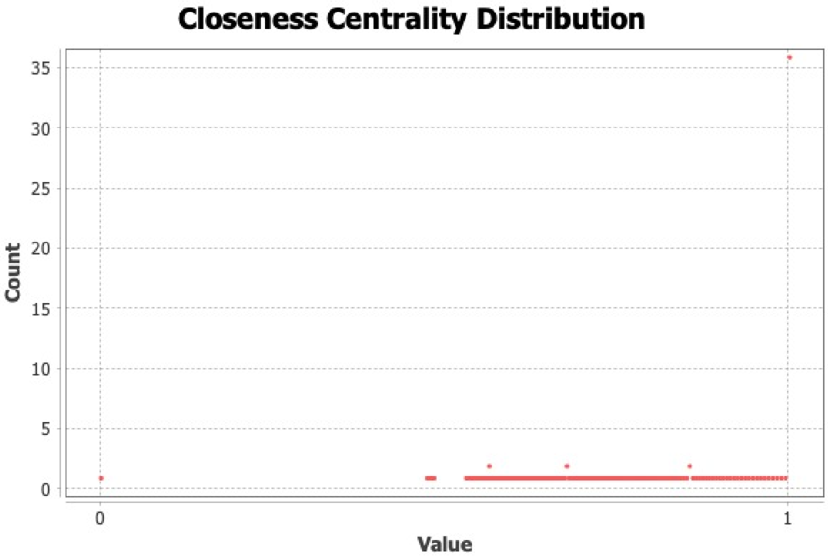


Figure 6: Closeness Centrality Distribution

Harmonic centrality is a variant of closeness centrality that was invented to solve the problem the original formula had when dealing with unconnected graphs. As with many of the centrality algorithms, it originates from the field of social network analysis.

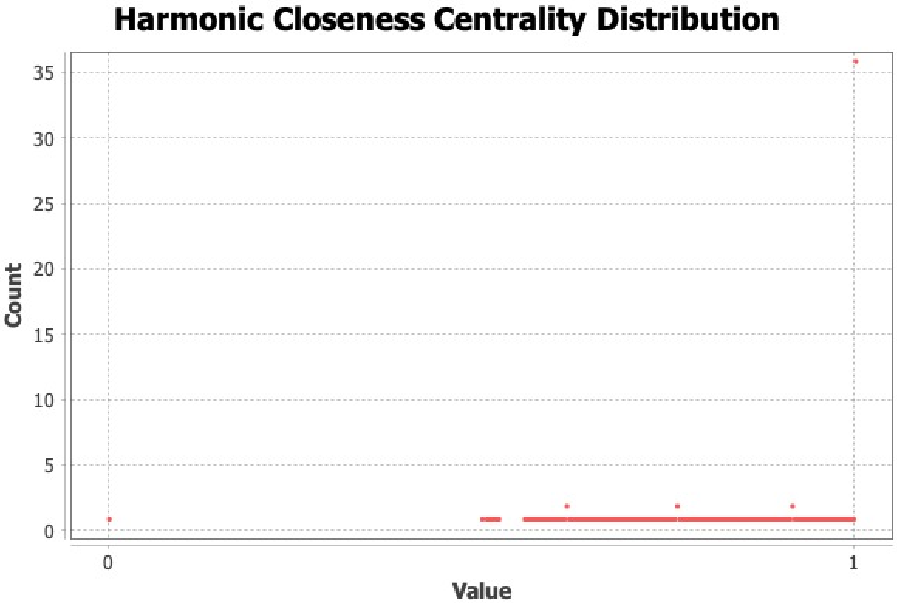


Figure 7: Harmonic Closeness Centrality Distribution

Eigenvector Centrality

A person with few connections could have a very high eigenvector centrality if those few connections were to very well-connected others. Eigenvector centrality allows for connections to have a variable value, so that connecting to some vertices has more benefit than connecting to others.

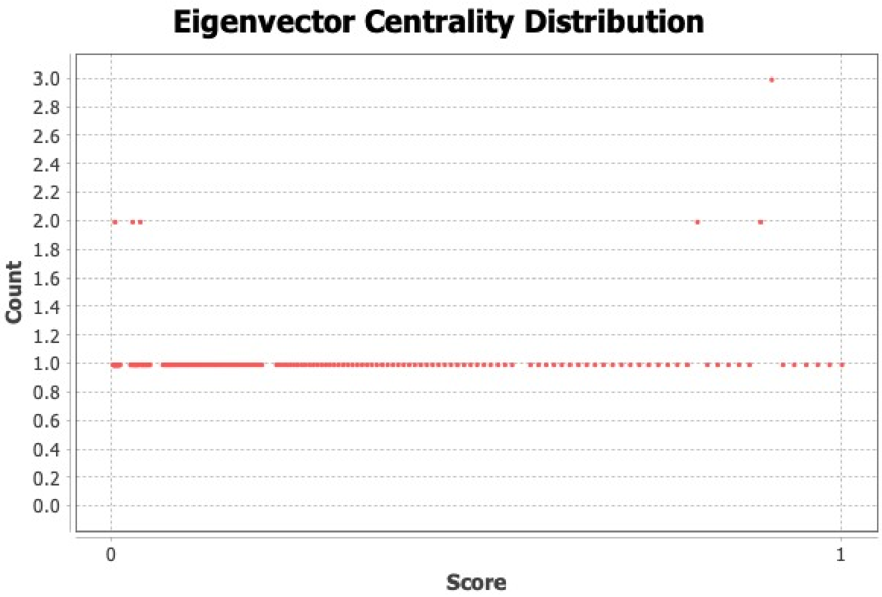


Figure 8: Eigenvector Centrality Distribution

Weighted Distribution

The edge nodes degree can be computed by taking the sum of the weights of all edges that link a given node.

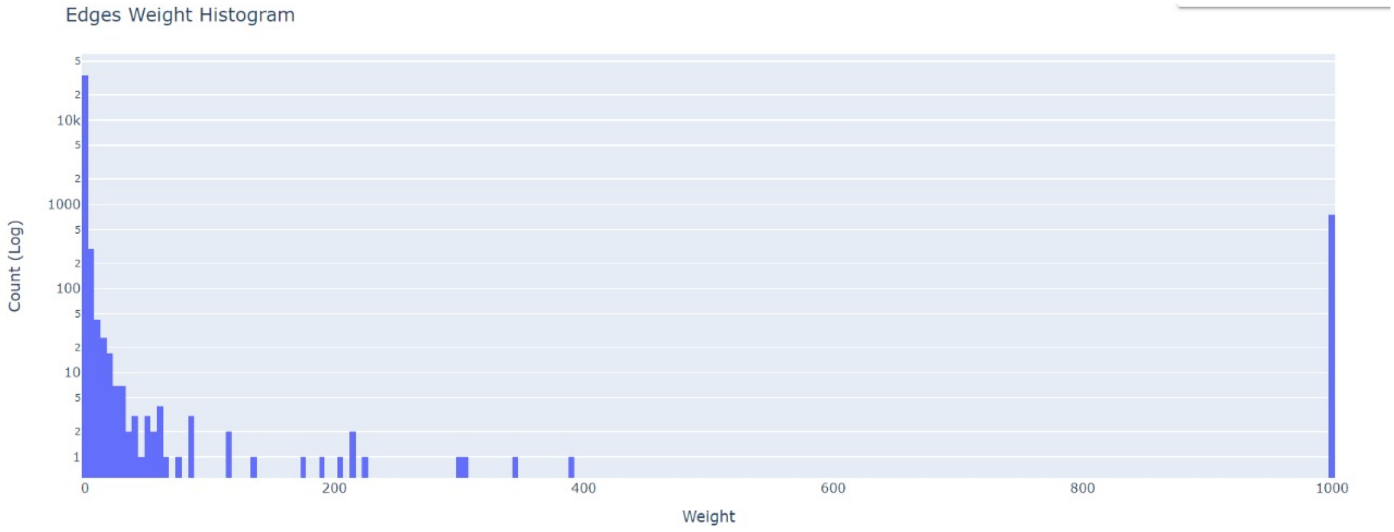
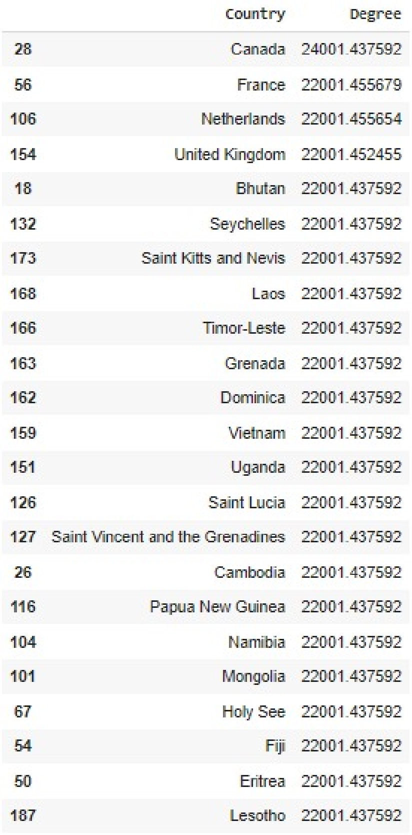


Figure 9: Edge Weight Distribution



	Country	Degree
28	Canada	24001.437592
56	France	22001.455679
106	Netherlands	22001.455654
154	United Kingdom	22001.452455
18	Bhutan	22001.437592
132	Seychelles	22001.437592
173	Saint Kitts and Nevis	22001.437592
168	Laos	22001.437592
166	Timor-Leste	22001.437592
163	Grenada	22001.437592
162	Dominica	22001.437592
159	Vietnam	22001.437592
151	Uganda	22001.437592
126	Saint Lucia	22001.437592
127	Saint Vincent and the Grenadines	22001.437592
26	Cambodia	22001.437592
116	Papua New Guinea	22001.437592
104	Namibia	22001.437592
101	Mongolia	22001.437592
67	Holy See	22001.437592
54	Fiji	22001.437592
50	Eritrea	22001.437592
187	Lesotho	22001.437592

Table 4: Degree Distribution of Countries greater than 20,000

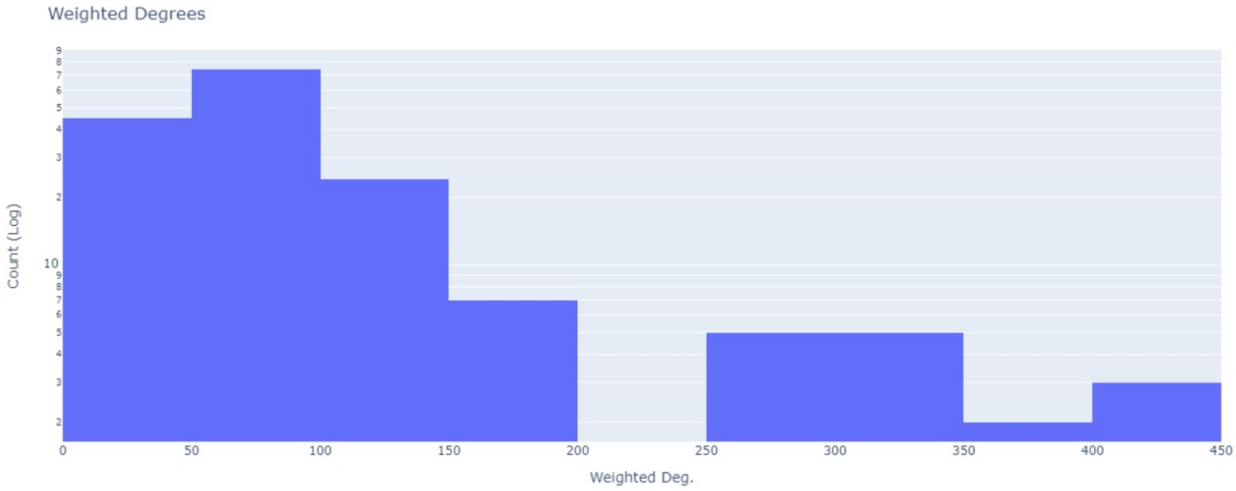


Figure 10: Weighted Degree Distribution of other Countries lesser than 20,000

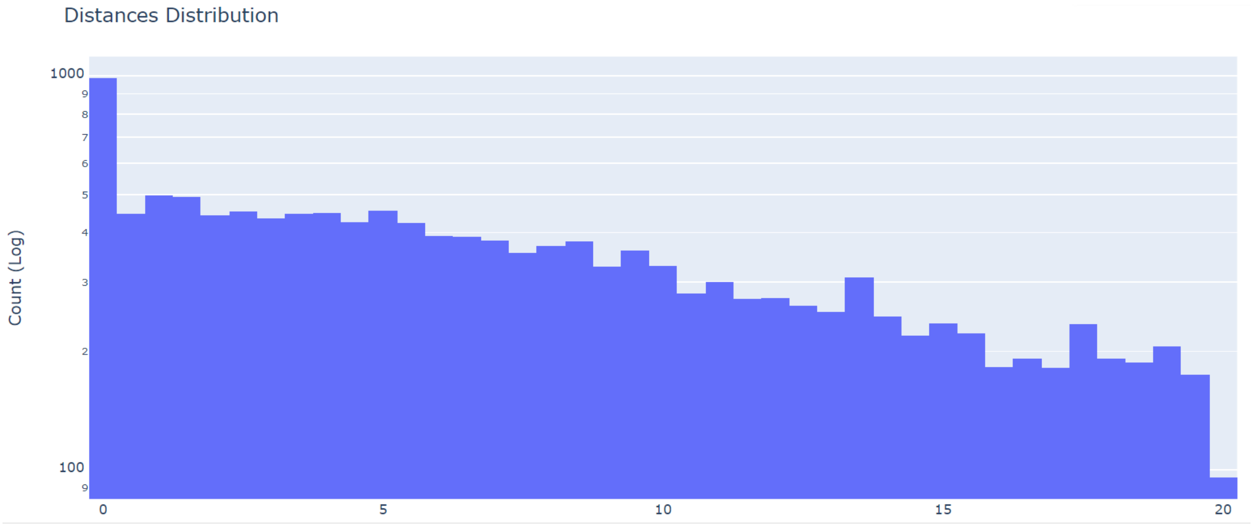


Figure 11: Distance Distribution of Countries lesser than 20

Above metrics that are performed to find edge distance distribution and degree distribution because it can be applied to find good clustering algorithms for finding transmission groups / communities that lead to widespread of COVID19 virus. So nodal degree distribution tells us that there are a few countries that have been the main transmission centre for COVID19. Further can be explored through clustering those values.

Other graph metrics

- **Average degree** : Average degree is simply the average number of edges per node in the graph.
    - Total Edges/Total Nodes=Average Degree
- **Average Weighted Degree :** Average sum of weights of the edges of nodes. The graph is designed in such a way that, weight of an edges represents, how many times that edges is traversed between a pair of nodes
- **Diameter of a network:** It is the shortest distance between the two most distant nodes in the network.
- **Density** is defined as the number of connections a participant has, divided by the total possible connections a participant could have.
- **Modularity** is a system property which measures the degree to which densely connected compartments within a system can be decoupled into separate communities or clusters which interact more among themselves rather than other communities.

Graph Metric	Value
Average Weighted Degree	3993.424
Average Degree	95.495
Average Path Length	1.392
Density	0.511
Diameter	3
Modularity	0.019

Table 5: Other Metrics

Clustering for community detection

Weight of the graph is the inverse of the absolute value of the difference between the mean times, and is considered to form an adjacency matrix which is used as a metric for formulating the clusters using hierarchical clustering. The weight of each edge signifies how inflated the spread is from the respective country to the other.

The following dendrogram was formed which was used to choose the number of clusters. The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold.

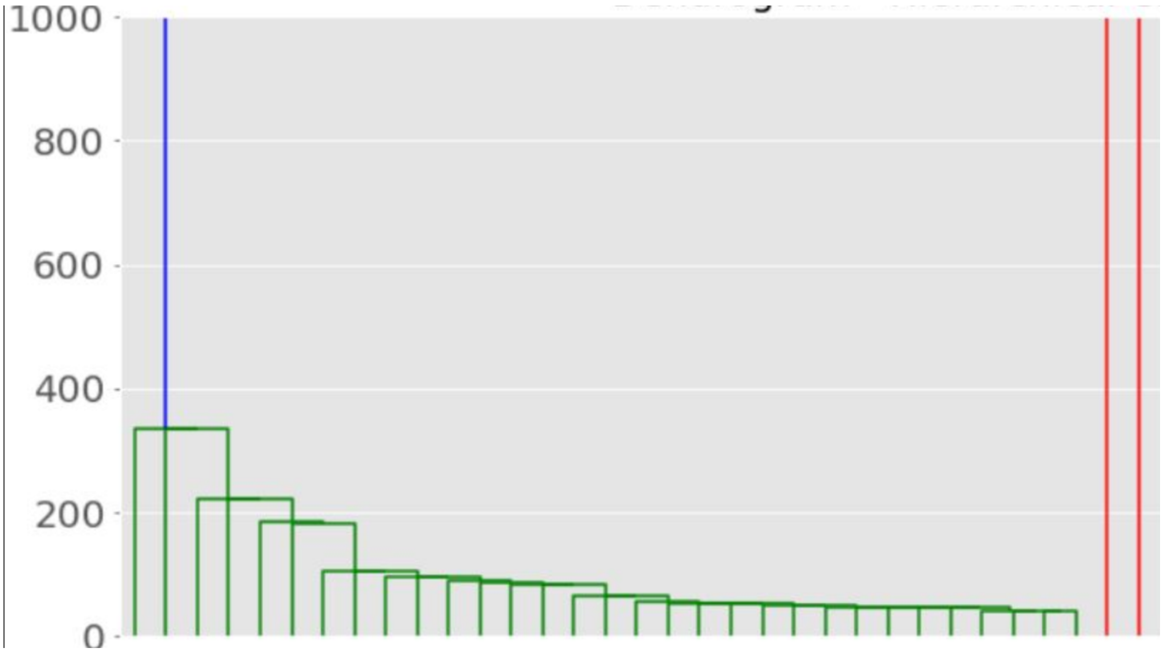


Figure 21: Dendrogram based on the weights of the edges

As we can see, if we take a threshold equal to 200, we will have 3 clusters, which is a reasonable number of clusters and, in this case, we will have 2 outliers (the figure helps us to find a good trade-off between the number of clusters and the number of outliers).

1. **Cluster 0:** The Majority with Origins

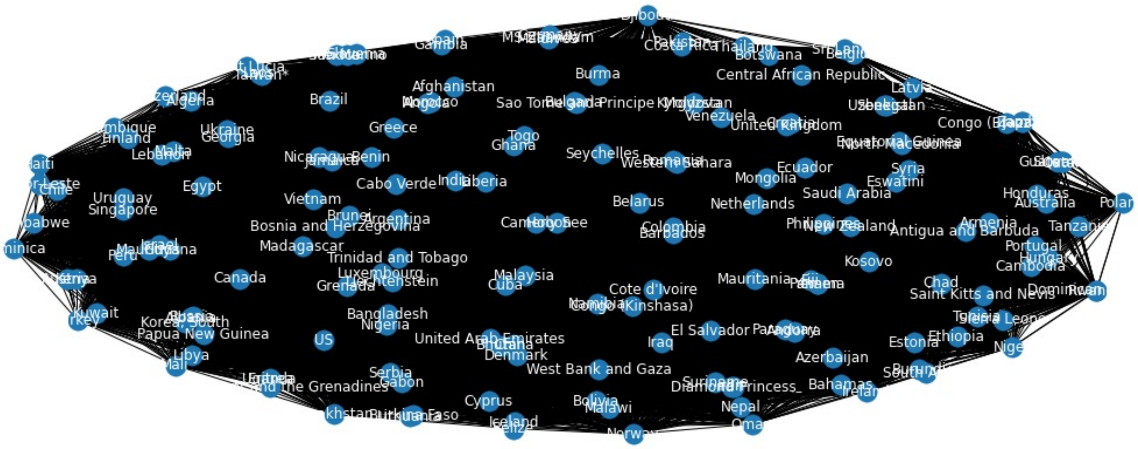


Figure 22: Cluster 0

The first cluster takes a huge number of countries when compared to others. They are formed by a huge group of countries but it is reasonable since China is seen to be a part of the cluster and hence has a strong influence in this cluster.

2. **Cluster 1:** The Major Europe, the Middle East and African countries

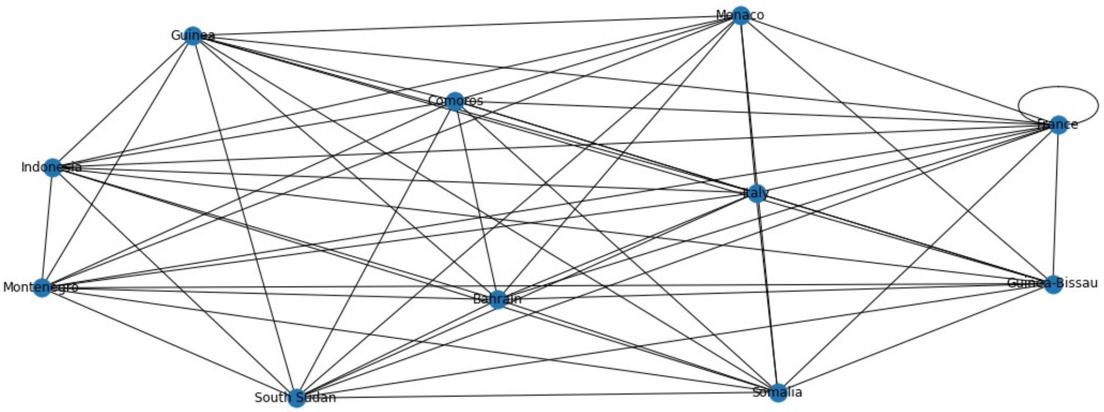


Figure 23: Cluster 1

We can also notice a strong presence of European countries like France and Italy along with several other Middle East and African countries.

3. **Cluster 2:** Example of a Minority Cluster

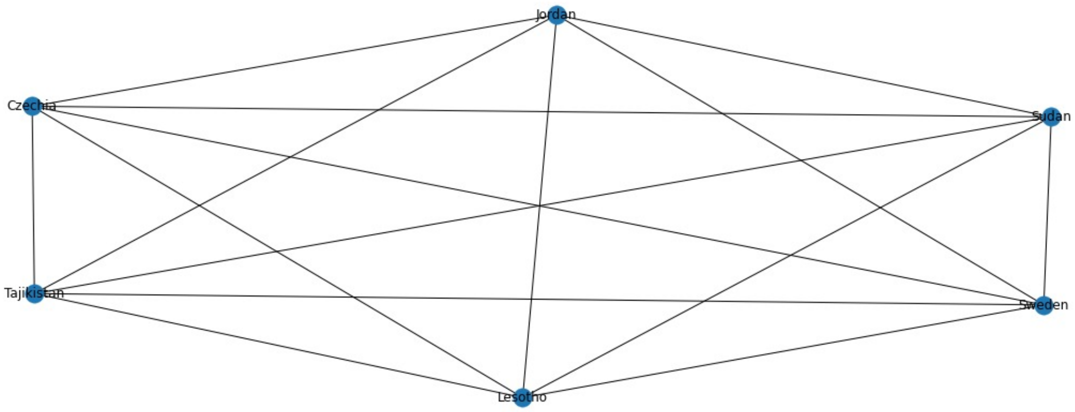


Figure 24: Cluster 2

Conclusion

Community detection in the spread of coronavirus across the world shows significant communities around Italy, China, and middle eastern countries which are the countries with the highest out-degree.

A given country should always be aware about new diseases in its "graph main neighbours". This is not a perfect analysis since some countries do not have a good testing rate and are passing by sub-notification problems but it's nice to see how we can transform transmission curves in instants of time.

It is important to point out the limitations to the SNA approach. In a majority of cases, datasets are incomplete. There are often problems associated with fuzzy boundaries and not knowing, in advance, who to include or not to include. Most importantly, there is lack of recognition for the dynamics of the network phenomenon: networks are not static. They evolve over time.

Appendix

1. Dataset Link:

   https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

2. Our Code Base:

   https://colab.research.google.com/drive/1-OAXhmag-bQ3csS1cjtiQXTtg2dOS1cv?usp=sharing