

# Social Network Analysis Using Gephi

Written by :

**Rauzan Sumara, Victor A, and Ary Irawan**

## What is Social Network Analysis (SNA)?

In social sciences, social network analysis has become a powerful methodological tool in addition to statistics, network concepts have been defined, tested, and applied in research traditions throughout the social sciences, ranging from anthropology, sociology, business administration and history (Nooy , 2005). Social Network Analysis (SNA) can be described as a study of human relations with the help of graph theory (Tsvetovat & Kouznetsov, 2011).

SNA studies structure relationships that connect individuals or other social units and dependence in related behaviors or attitudes with the arrangement of social relations. The relationship described by nodes, or can be called vertices symbolizes the actor or user and ties or also called edges, links or connections that symbolize relationships between actors (O'Malley & Marsden, 2008). SNA analysis and implementation can be applied to obtain various information. For example, brand awareness or top brand analysis in E-commerce.

To assess top brands using SNA, we can use network properties on SNA. Property on social network can show the relationship between e-commerce brands with users who mentioned about their brand on social media, so that activity can be assessed and the level of user awareness regarding e-commerce brand that is discussed on the social media by the users. Not just e-commerce brand, or any other product, we can use SNA to analysis personal branding in the social media, how the user see us in the social media, or what the impression of the user regarding our personal brand.

## Let's begin the Analysis

The dataset for this Social Network Analysis taken from Twitter using crawling feature from R Studio. For the analysis, I use Twitter data relate to @WHO and #COVID19 tweets in USA. COVID-19 pandemic was confirmed to have reached the United States in January 2020. The first confirmed case of local transmission was recorded in January, while the first known deaths happened in February. By the end of March, cases had occurred in all 50 U.S. states, the District of Columbia, and all inhabited U.S. territories except American Samoa. As of May 27, 2020, the U.S. had the most confirmed active cases and deaths in the world. As of June 5, 2020, its death rate was 330 per million people, the ninth-highest rate globally.

Therefore in this article we want to track tweets related to #COVID19 which account are mostly and actively talking about the coronavirus, to know structure relationships that connect individuals or related behaviors of social relations.

## Crawling Step

The R script to crawl data from Twitter. Using “rtweet” library. For the data limit, using n = 10000 on Saturday 13rd and export it into the csv.

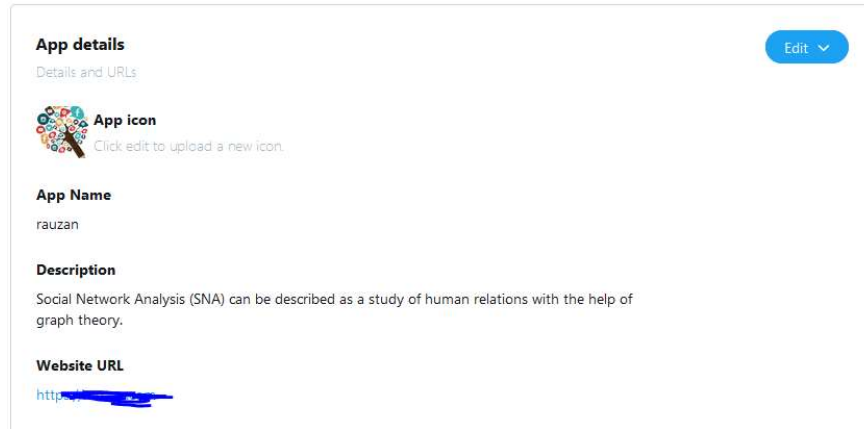
```

1 ## install rtweet
2 install.packages("rtweet")
3
4 ## load rtweet
5 library(rtweet)
6
7 ## store api keys
8 api_key <- "d8Cn..."
9 api_secret_key <- "9hEYog20k1bXNb2..."
10
11 ## authenticate via web browser
12 token <- create_token(
13   app = "rauzan",
14   consumer_key = api_key,
15   consumer_secret = api_secret_key)
16
17 ## search for 10,000 tweets sent from the USA
18 tweets <- search_tweets("#COVID19 OR @who", lang = "en",
19   geocode = lookup_coords("usa"), n = 10000)
20 write_as_csv(tweets, "D:/Dataset")
21
22 ## create lat/lng variables using all available tweet and profile geo-location data
23 rt <- lat_lng(tweets)
24
25 ## plot state boundaries
26 par(mar = c(0, 0, 0, 0))
27 maps::map("state", lwd = .25)
28
29 ## plot lat and lng points onto state map
30 with(r, points(lng, lat, pch = 20, cex = 2.5, col = rgb(0, .3, .7, .75)))

```

But before we do that, we should request access to Twitter through registering our Twitter account to the Twitter Developers. you can read the detail here:

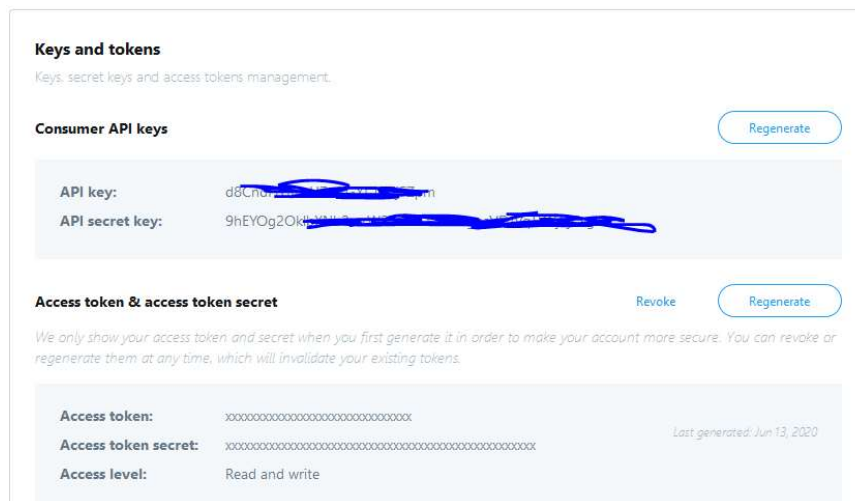
<https://developer.twitter.com/en/docs/basics/getting-started>



The screenshot shows the 'App details' page in the Twitter Developer Console. It includes an 'App icon' placeholder, the 'App Name' 'rauzan', a 'Description' about Social Network Analysis (SNA), and a 'Website URL' field with a redacted URL. An 'Edit' button is in the top right corner.

Field	Value
App Name	rauzan
Description	Social Network Analysis (SNA) can be described as a study of human relations with the help of graph theory.
Website URL	http://[redacted]

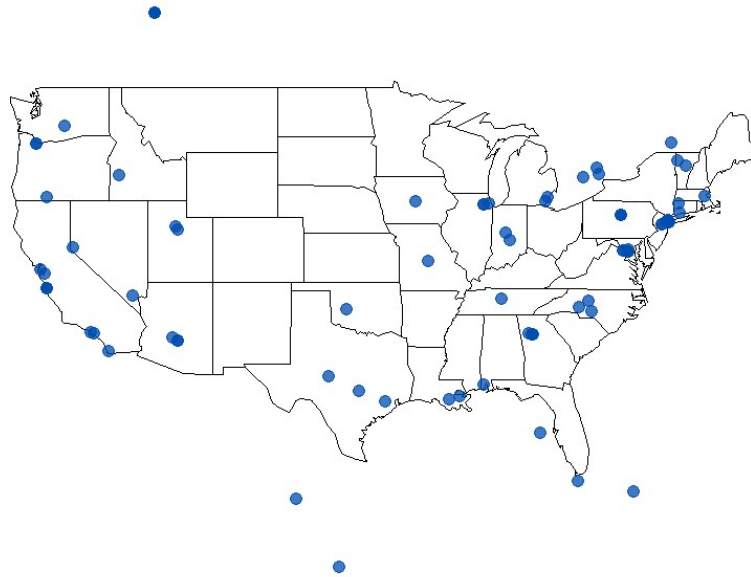
After that, we can get out API key and API secret key and use that as input of R code.



The screenshot shows the 'Keys and tokens' page in the Twitter Developer Console. It displays 'Consumer API keys' with a 'Regenerate' button. Below, it shows 'Access token & access token secret' with 'Revoke' and 'Regenerate' buttons. The access token and secret are redacted, and the access level is 'Read and write'.

Field	Value
API key	d8Cn[redacted]
API secret key	9hEYOg2Ok4[redacted]
Access token	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Access token secret	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Access level	Read and write

There are 96 variables and 10000 observation, and now we also visualize where tweets are from which areas they are. It is quite interesting because we can understand how separable our data is. We prefer random tweets data in order to get result which can represent the population. As you can see in the picture bellow, our tweets that we have are separately from different place and are retrieving randomly.



From 96 variables, we only need two which is text (tweet from someone who mention or retweet #COVID19) and screenName. Therefore first of all, we have to clean our data. In this case, we were using OpenRefine, more detail here <http://openrefine.org/>

## **Gephi Implementation**

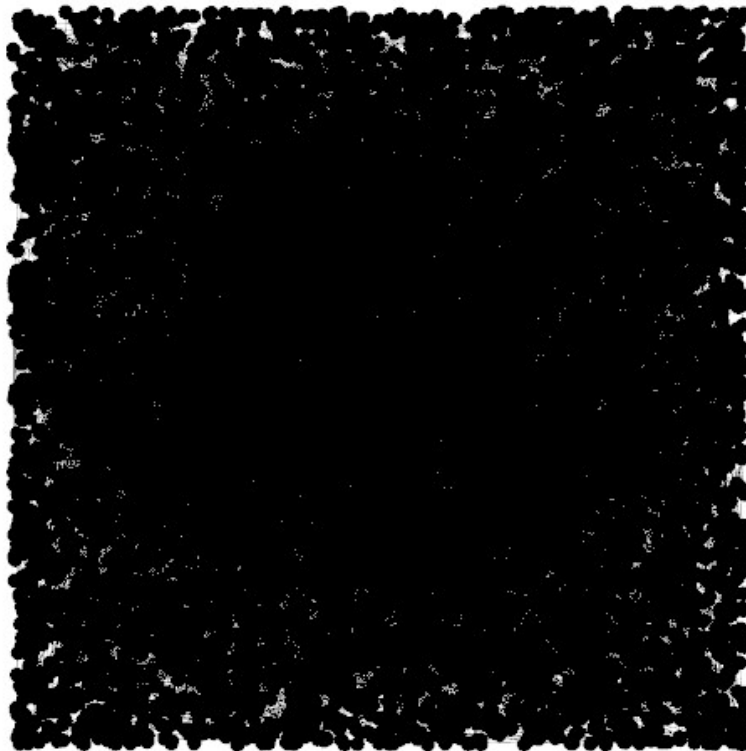
Gephi is an open source application for network exploration and manipulation. A network module that will be developed can be processed by importing, visualizing, mapped, filtered, manipulated and exported in Gephi. Data processing in Gephi is carried out with the following steps:

1. Import network data sets that have been created before using help spreadsheet in OpenRefine. Data sets that can be used are only data sets with extensions .csv text based.
2. Select the visualization algorithm to be used. This algorithm functions to determine governance the location of the nodes that will be visualized in the sociogram. Note that algorithm selection affects the form of network visualization that will be generated.

3. Personalize the network that has been visualized. In this process settings will be made color display, shape, labeling of nodes in the network and also can control the thickness of an edge between nodes and naming the edge too.
4. Calculate network property values. In this research, the value is calculated network property attributes in the form of Total Node, Total Edges, Average Degree, Average Weighted values Degree, Network Diameter, and Number of Communities.
5. Displays the ranking of the nodes that have the highest influence or interaction value in the network. This step can be done in two ways, first by looking directly in the windows data table, and the other is to configure the display by changing the size of the node or label node in the network visualization image according to the order of values owned by these nodes.

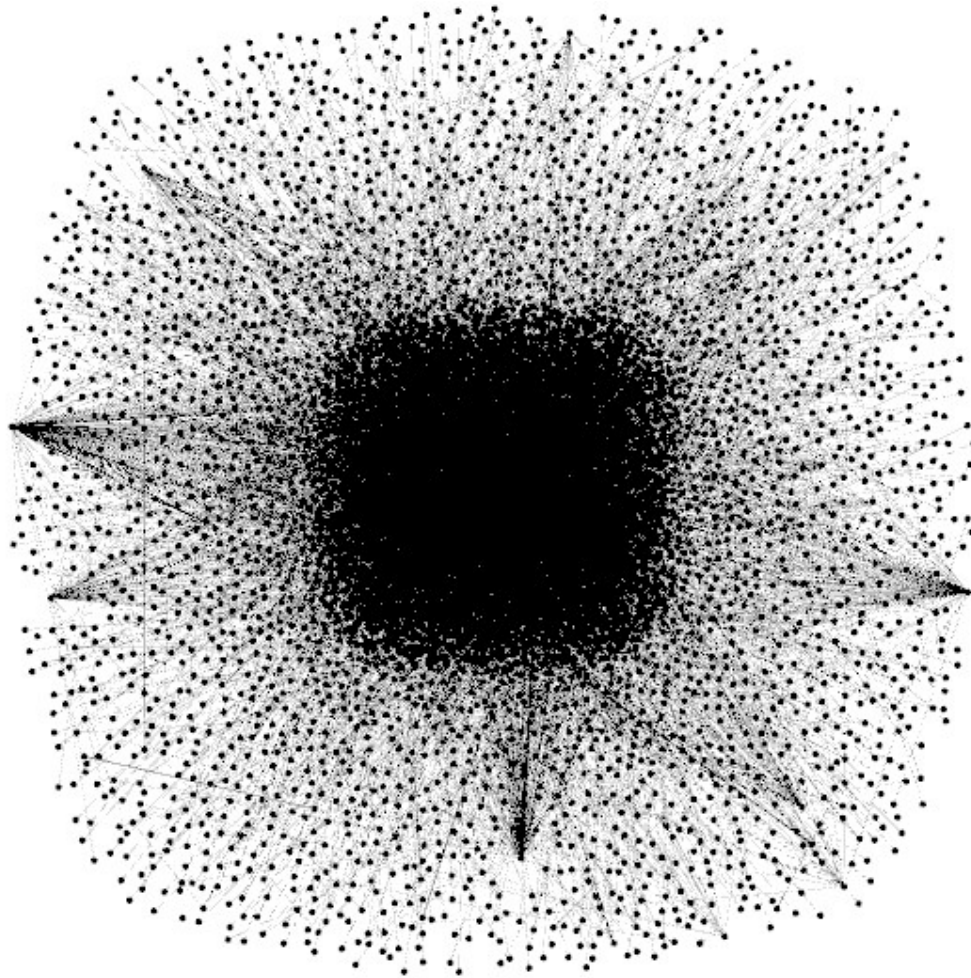
## **Network Visualization**

After we successfully import the graph, this is the first look of the graph. Kind of messy and difficult to interpret. The next step is to change the layout.



## Force Atlas Layout

Based on the graph above, we want to make it more interpretable. I try to change the layout as “Force Atlas” with “Repulsion strength” at 20 000 to expand the graph.



Now, the graph looks more clearly. There are several clusters in the network, and we can also see the nodes and edges in each group. However, we still cannot see which nodes that is the center of the graph and with whom they interact the most. The next step is to see the name of each node and give it the color too, so we can look the graph clearly.

## Betweenness Centrality

First, run the in the statistic field. It is the maximal distance between all pairs of nodes and in the appearance field, it will have the **Betweenness Centrality** in the ranking field. Choose it and change the size min 20 and max 119.

## Graph Distance Report

---

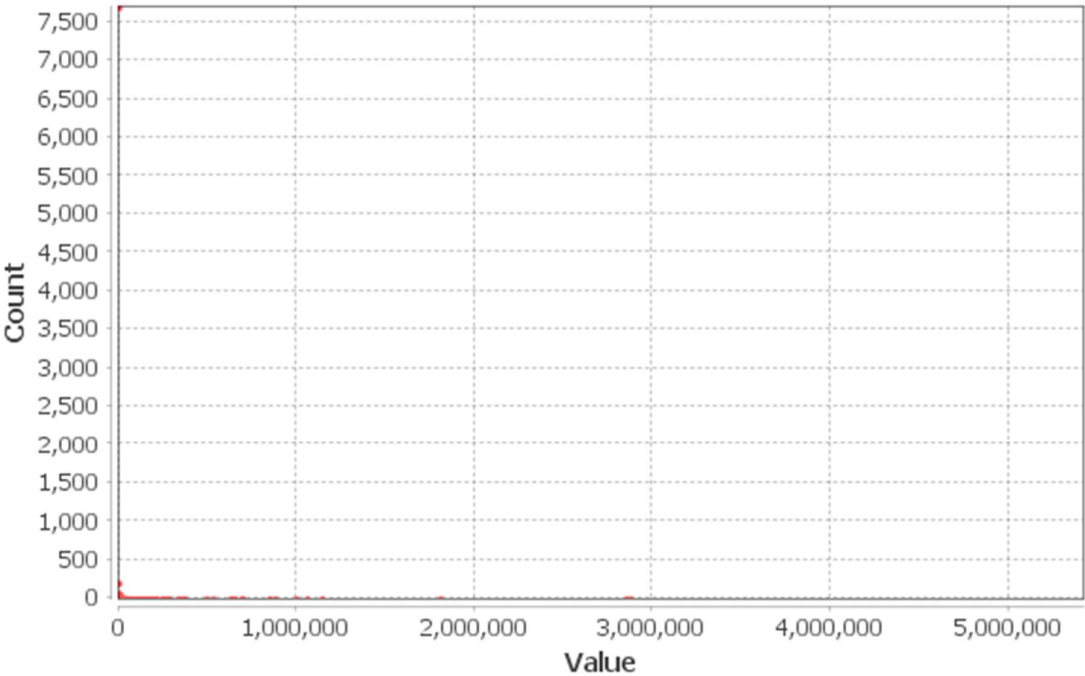
### Parameters:

Network Interpretation: undirected

### Results:

Diameter: 16  
Radius: 0  
Average Path length: 5.409916116080412

### Betweenness Centrality Distribution

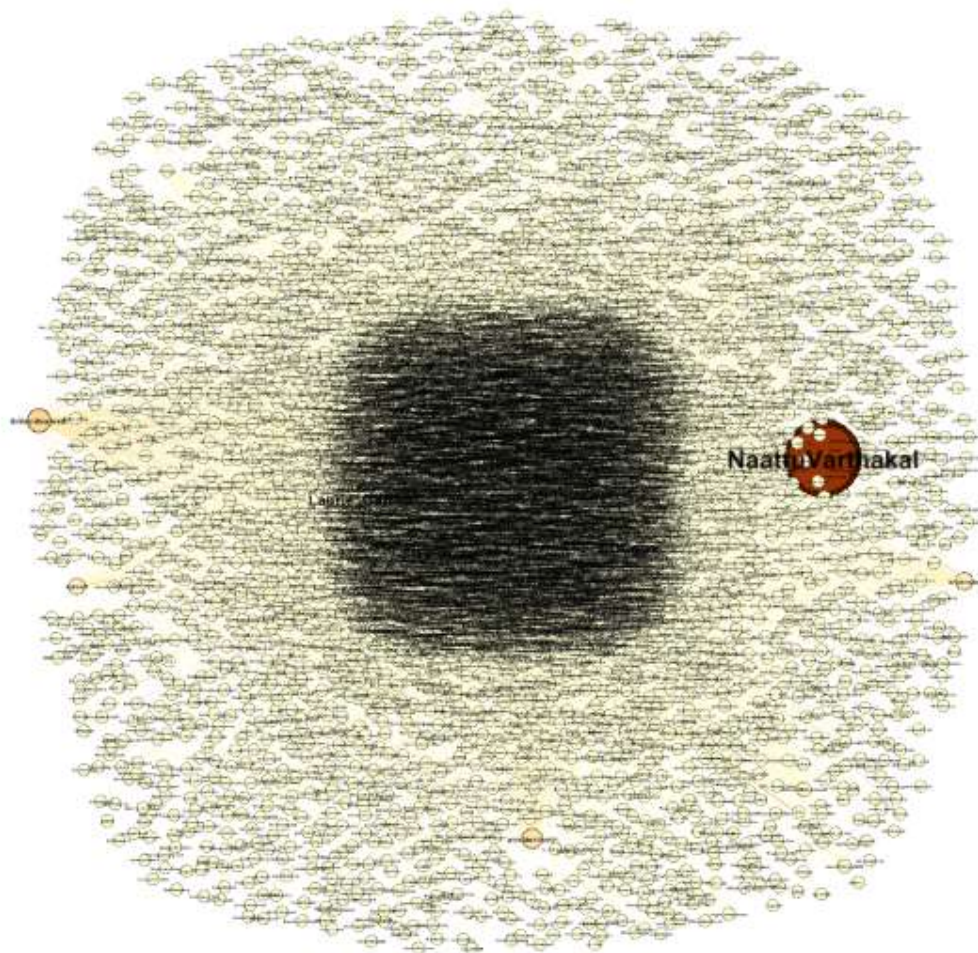


### Algorithm:



Ulrik Brandes, *A Faster Algorithm for Betweenness Centrality*, in Journal of Mathematical Sociology 25(2):163-177, (2001)

When it is finished, the metric displays its result in a report such as diameter: 16, radius: 0, and average path length: 5.4099. Metrics generates general reports but also results for each node. Thus three new values have been created by the “Average Path Length” algorithm we ran, such as Betweenness Centrality, Closeness Centrality, and Eccentricity. But in this case, we choose “**Betweenness Centrality**”. This metrics indicates influential nodes for highest value.



We can see that there are in different size and the edge also different in the level of thickness of the line. In this step, we Implement Betweenness Centrality with size between 20 to 119. It



can be said that Betweenness is a symbol of “strength” or “influence” of a node in social networks. Because the node is a bridge to another node.

## **Modularity reports**

After that, it can indicate that the node (the big one) have high controls collaboration between, disparate clusters in a network; or indicate they are on the periphery. But we cannot see which person that belongs to the nodes and how strong they interact with each other. Therefore we implement “**Modularity reports**” to the graph. Therefore in order to detect communities in network, we would like to colorize clusters using the Louvain method or usually called “Modularity”. Based on the result below, we have Modularity: 0.955 and Number of Communities: 1340. In the same way, we create filters that can hide nodes and edges on the network. We will create a filter to remove leaves, i.e. nodes with a single edge.

# **Modularity Report**

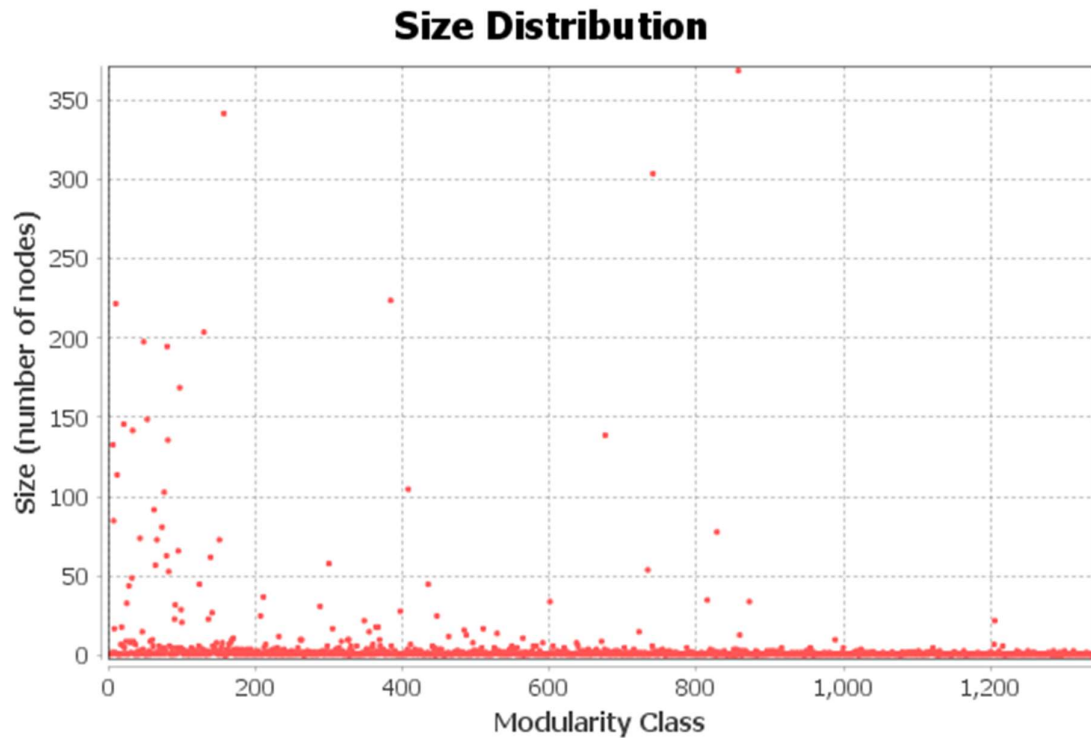
---

## **Parameters:**

Randomize: On  
Use edge weights: On  
Resolution: 1.0

## **Results:**

Modularity: 0.955  
Modularity with resolution: 0.955  
Number of Communities: 1340

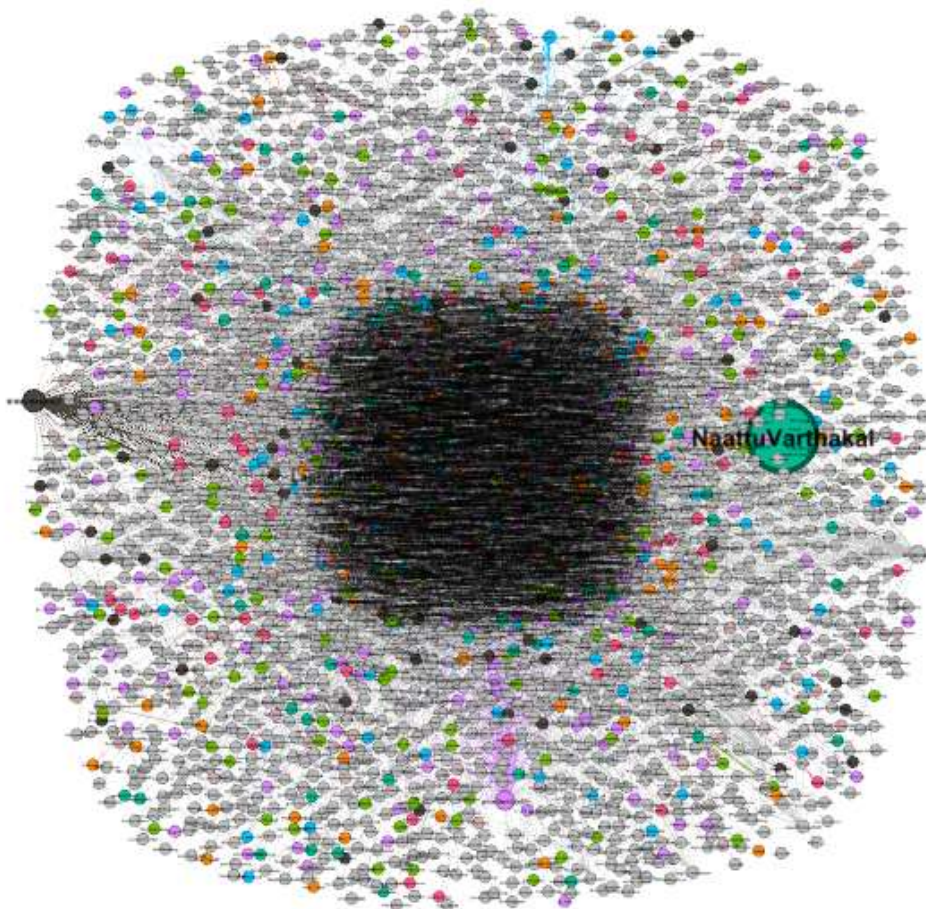


## Algorithm:

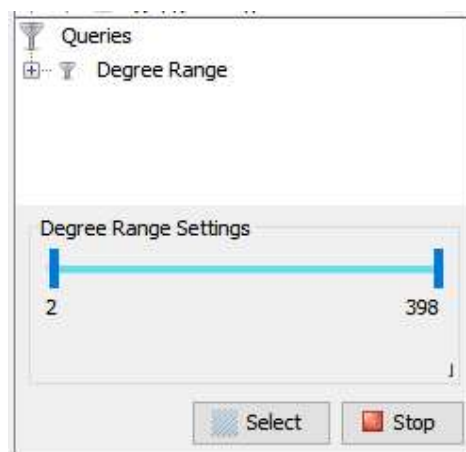
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, *Fast unfolding of communities in large networks*, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

Appearance		
Nodes	Edges	
Unique	Partition	Ranking
Modularity Class		
722		(5.06%)
842		(4.23%)
131		(3.19%)
360		(2.57%)
5		(2.55%)
127		(2.34%)
806		(2.06%)
409		(2.01%)
537		(2.01%)
298		(1.72%)
465		(1.7%)

The colors in Modularity Reports indicate that different communities determined by this algorithm and basically it will show which routers are more densely connected between each other than to the rest of the network.

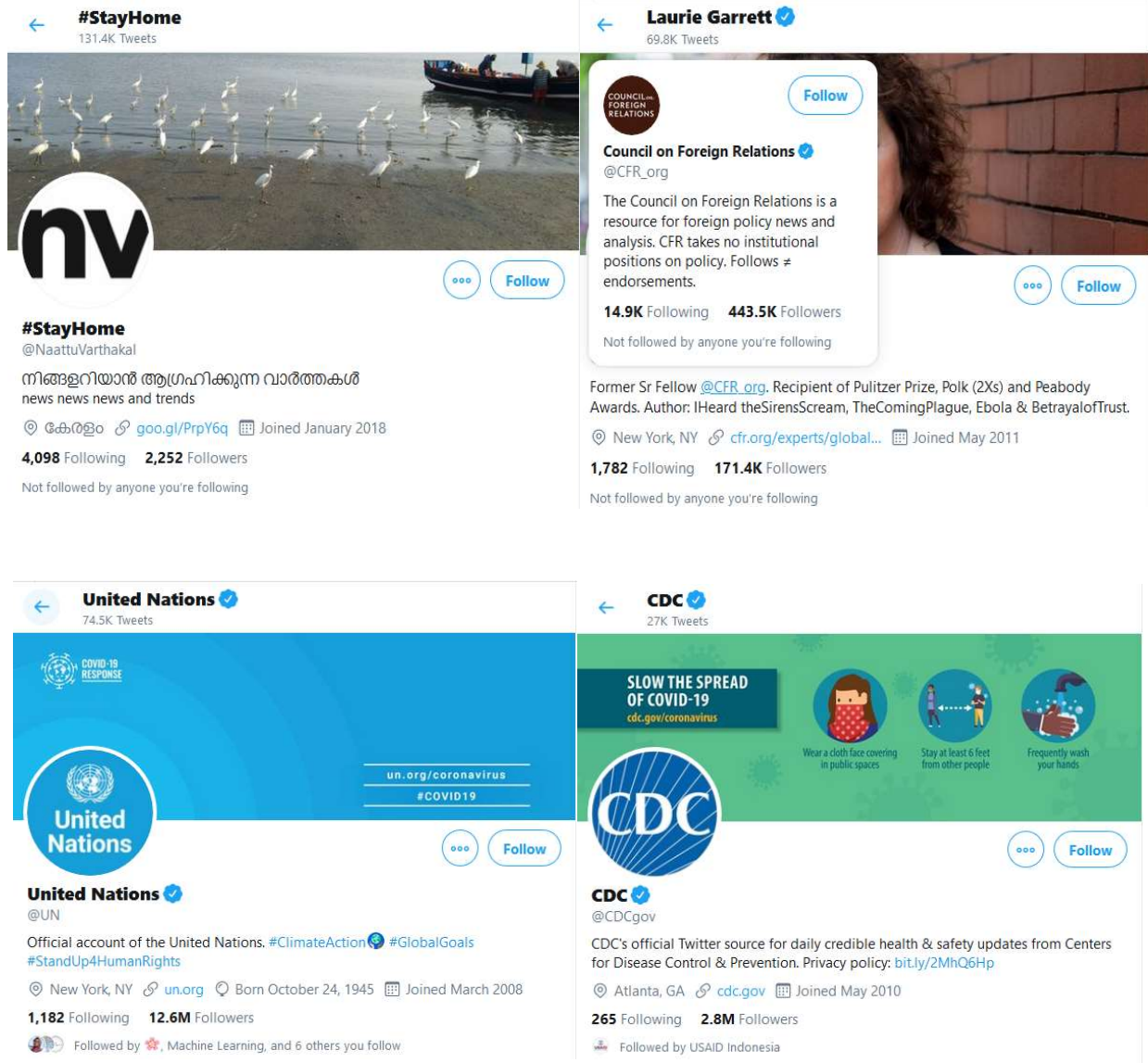


Because it created so many communities, here we try to filter degree range starting from 2. After that we will have the new graph as followed picture,





work of the United Nations are guided by the purposes and principles contained in its founding Charter.







In the graph above, we see that **Laurie Garrett** is the center of certain nodes and we can also note that there are also 2 nodes mostly interacting with **Laurie Garrett** which is **Nuria2407** and **Nikluk**.



Furthermore we also can see that United Nations has its subnetwork, and **Kenzytweets** account mostly made retweets from United Nations. According to any retweet that **Kenzytweets** does, there are several accounts are mostly interacting with **Kenzytweets** such as **AminaJMohammed**, **TheUNTimes**, **MelissaFleming**, **IMFNews**, **ASteiner**, **TheWordLesson**, and **UN\_News\_Centre**.

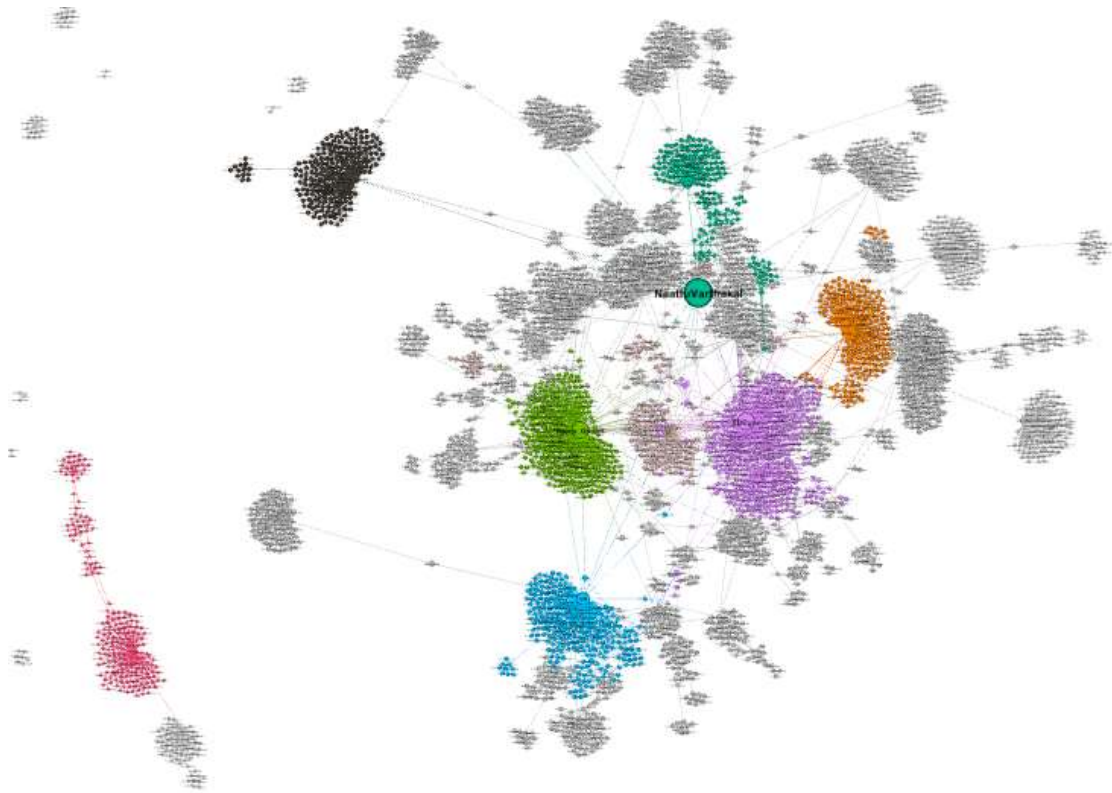




And then the last subnetwork, the graph above, we can see that **CDC** makes tweets and will be retweet by **SaludHEALHinfo**, **CDCemergency**, **Nuria2407**, **timwisn31**, **MayoClinic**, **tedwisn3** and **Nuria2407**.

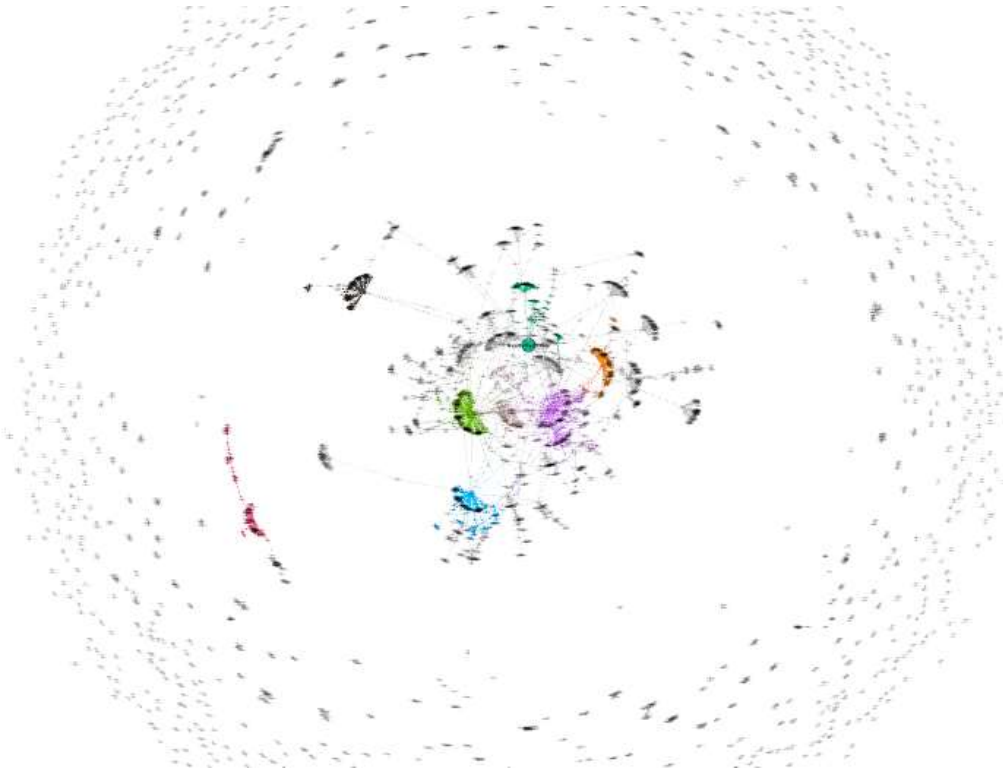
## Force Atlas 2 Layout

Based on the graph below, I want to make it more interpretable. I try to change the layout as “Force Atlas 2” with “Prevent Overlap” to expand the graph.

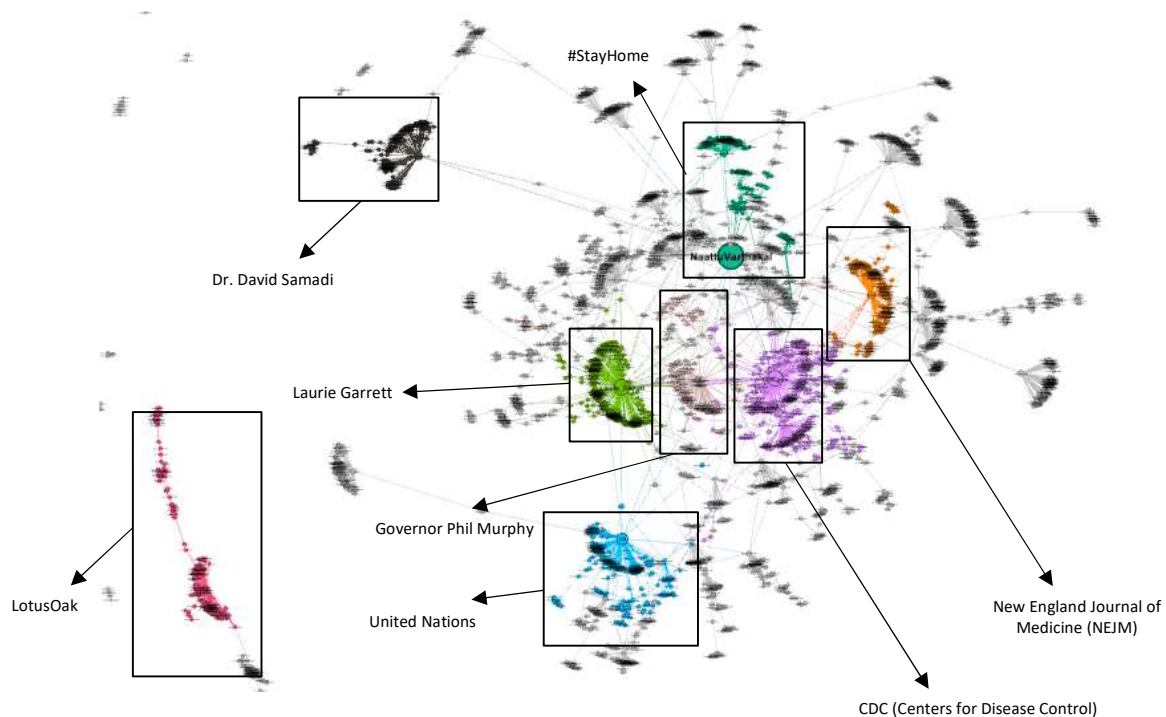


Now, the graph looks more clearly. There are several clusters in the network, and we can also see the nodes and edges in each group. However, each cluster has different sizes, it will represent how big their influences are. Therefore, we will present it in percentages later.

## Yifan Hu Proportional Layout



There are 2 part of graph which are inner graph and outer graph. Inner graph has play a part in inviting other people to aware of corona virus and campaigning to stay at home during the pandemic. Besides, outer graph is representing the people who are affected by campaign. They made tweets related to covid19 to support headline about staying at home during the pandemic.



As you can see that the first cluster which is **CDC**, official Twitter source for daily credible health & safety updates from Centers for Disease Control & Prevention, (purple color) has 5.06% in influencing to the society. The second cluster which is **Laurie Garrett** (green color) has 4.23% in influencing. The third cluster which is **United Nations** (light blue color) has 3.19%. Moreover, we can see **New England Journal of Medicine (NEJM) account** (orange color) has 2.55% of the impact. The New England Journal of Medicine (NEJM) is the world's leading medical journal and website. Published continuously for over 200 years, NEJM delivers high-quality, peer-reviewed research and interactive clinical content to physicians, educators, and the global medical community. Also the other person comes from political background, **Governor Phil Murphy** for example, he is 56th Governor of the great State of New Jersey. It has 2.01% of the impact to the community.



Besides, there is person who comes from doctor background named **Dr. David Samadi**. He was joining @FoxNews medical a team and expert in prostate cancer and his research focused on robotic surgery. He has 2.57% in influencing to the society through twitter during the pandemic.

← **Dr. David Samadi** ✓  
16K Tweets



**Dr. David Samadi** ✓  
@drdavidsamadi

was @FoxNews Medical A team. #RoboticSurgery . Expert in #ProstateCancer.  
#MAGA samadimd.com @newsmax contributor 🇺🇸🇺🇸🇺🇸🇺🇸

📍 Manhattan, New York 🔗 prostatecancer911.com 📅 Joined March 2009

4,602 Following 74.7K Followers

Not followed by anyone you're following