COVID-19: THE RISE OF ANTI-VAXX INFORMATION ON TWITTER AND THE IMPACT ON VACCINE HESITANCY.

Nadezhda Manuilova,

University of Padova, Department of Information Engineering
Network Science, February 18, 2022

TWITTER ANALYSIS

a. TWITTER DATA

For our analysis, we concentrate on discussions on Twitter around the rise of anti-vaxx information from January of 2018 to January of 2022. Twitter is a directed graph social networking platform in which each user can choose to follow a number of other users and can also be followed by other users. As a result, the "follow" connection is asymmetrical, it requires no necessary acknowledgement, and is primarily used to receive all public messages posted by any user. We used an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles. Twitter's search operators let you scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags and trends, sort out sensitive information from Tweets like email and phone numbers.

Consequently, in our analysis we collected tweets relating to certain topics, hashtags & trends. In particular, our main "keywords" were: Antivax and VaccinesKill.

To collect the twitter data, we create our own application with help of twitter API. For this, we create a developer account and register our app. After we receive a consumer key and a consumer secret: these are used in application settings and from the configuration page of the app we require to make analysis in Twitter. The API variable is now our entry point for most of the operations we can perform with Twitter. The API provides features to access different types of data. In this way, we can easily collect tweets and store them in the system.

Firstly, we start to collect tweets with key word Antivax with two different intervals of time (January of 2018 and January of 2022). Secondly, we collect tweets with key word VaccinesKill with two different intervals of time (September of 2020 and January of 2022).

The work starts with setting up Twitter API, Download data with special dates for each tweet and inspect data. Then, we convert the data into Pandas DataFrame because this format helps us easy manipulation of the data as well as saving and loading data. After we filter the dataframe and keep only columns we are interested in ["created_at", "id_str", "text", "user", "retweeted_status", "quote_count", "reply_count", "retweet_count", "favorite_count", "entities"]. Next step we build the networks from our data, which we download from Twitter.

Extracting words. For these we clean the Tweet text: remove punctuation, hashtags, mentions, urls. Also, turn all letters to lowercase to be similar to each other. We loop through the dataframe and then through the words in the clean text. We add the words as keys to dictionary and use their frequencies as values.

We use the network library (Python library that enables network science analysis of the data) to create our network and extract edge list. In addition, we loop through our dataframe and connect words and hashtags. Which words and hashtages appear together in the same Tweet we connect. After all steps our Node Lists and Edge Lists are created.

Since we save our tweets, node lists and edge lists.

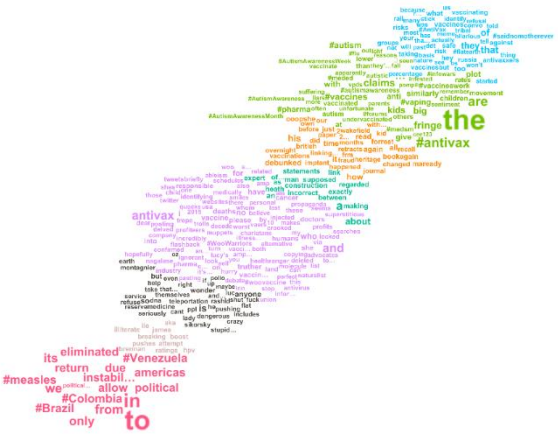
b. VISUALIZATION OF NETWORKS

The most convenient open-source software for visualizing and analysing large networks graphs is Gephi. Gephi uses a 3D rendering engine to display graphs in real-time and to speed up exploration. It can be used to explore, analyse, spatialize, filter, clusterize, manipulate and export all types of graphs. How we use Gephi:

1. Download the software on computer and create a new project in the start window.
2. Import nodes.csv and edges.csv files which we download from Python
3. Choose an undirected network as the network type.
4. Put PageRanks and Modularity algorithms, pick the palette and the partition option to alter the color of the nodes based on the community they belong to and the PageRank rate they have.
5. PageRank range eliminate nodes, which are not significant in the network and to view the network more clearly.
6. Choose MDS Layout, choose Modularity class and at the end, select the Label Adjust plugin.

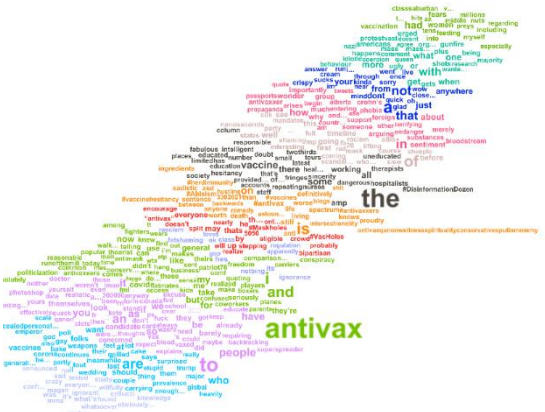
After analyzing the data, we received related words with antivax, such as #Brazil, #Colombia, #Venezuela, #kids #dangerous. After analyzing the results, we found that in 2018 cases of yellow fever were reported in Brazil, Colombia, Venezuela. The Brazilian, Columbian, Venezuelian health

authorities took the necessary measures, including vaccination, laboratory testing and vector control. At the end of 2018, vaccination against yellow fever was started in the countries. Preventive vaccination against yellow fever of persons traveling to countries endemic for this infection is the only and most reliable means of preventing the disease. In accordance with the requirements, persons who have been vaccinated have received an international vaccination certificate, which is valid for a lifetime. From this moment people start discussing vaccination problem in Twitter. It can be noticed that most of the words are of American origin. This also means that the problem of vaccination was concentrated in North and South America at this period of time.



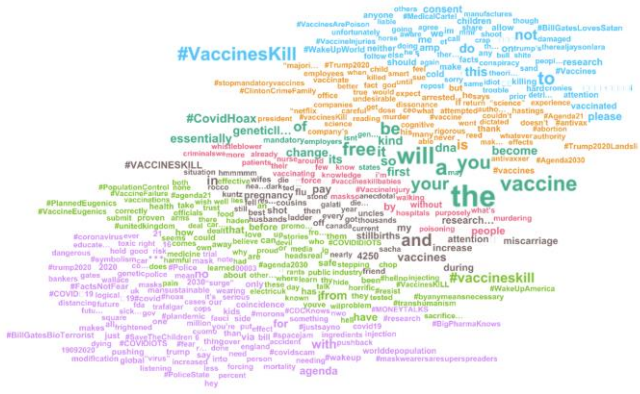
Graph with the data from Twitter #Antivax in January 2018

In 2019 a new disease was born and it spread very fast around the world. After a huge number of people began to die and people develop a new vaccine from this disease in the last months of 2020. In the beginning of 2021 the number of vaccines has increased significantly due to the emergence of a new disease in all countries. Disease was spread all over the world and that is why the plot of data much larger than in 2018 year. Since vaccines are new for all mankind we can see the following linked tweets #global #knew #confused #life #realize. People was confused and do not know what to expect. They try to search information and share their thoughts in Twitter.



Graph with the data from Twitter #antivax in January 2021

We can see that the number of searches on Twitter about VaccinesKill has increased significantly. Because it was a new vaccine for people, they did not know what the effect it might have on a person. That is why we see such linked tweets #women #influence #happens #dangerous #society.



Graph with the data from Twitter #VaccinesKill in September 2020

In 2022 the situation significantly changes. The plot decrease. And the most important question for people now is why the vaccine should be mandatory for everyone. We see the linked tweets #passport, #mandate, #again, #everyoneshould.

Graph with the data from Twitter #VaccinesKill in January 2022

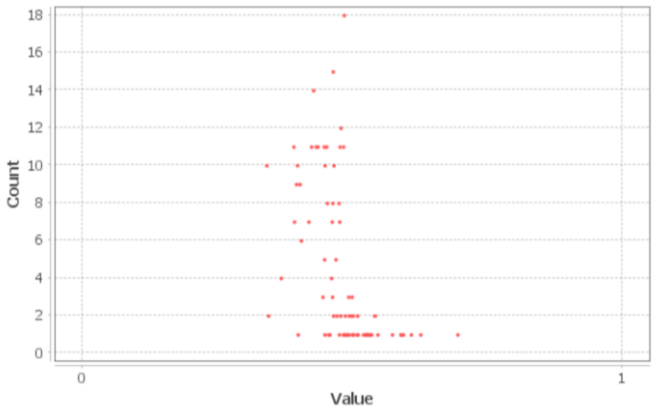
From these graphs we can see that there is a connection between Antivax and VaccinesKill in the network extracted from Twitter. All people are concerned about the impact the vaccines will have on human health.

c. ANALYSIS
CENTRALITY

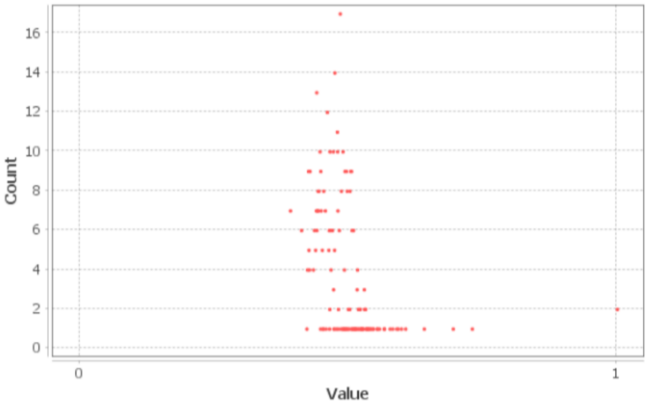
Network analysis is used to measure centralities of the network to see which nodes are more important. There exist many centralities in Gephi:

1. Degree Centrality is used to measure how many links connected with a node, a high degree centrality which means the node connected with many links.
2. PageRank is used to now the number of connections on a node and also the quality of the links.
3. Betweenness Centrality it is a metric for identifying the principal bridge-nodes across communities based on the number of shortest routes traveling through each node.
4. Closeness Centrality it is a metric for identifying the principal bridge-nodes across communities based on the number of shortest routes traveling through each node.

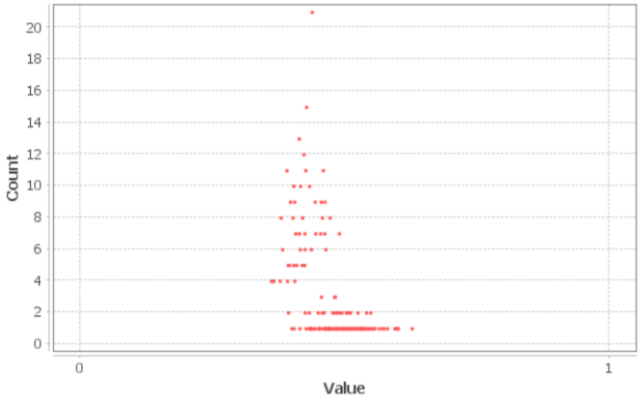
We make analysis of closeness Centrality of #antivax and # VaccinesKill. The avarage distance from a given starting node to all other nodes in the network is different.



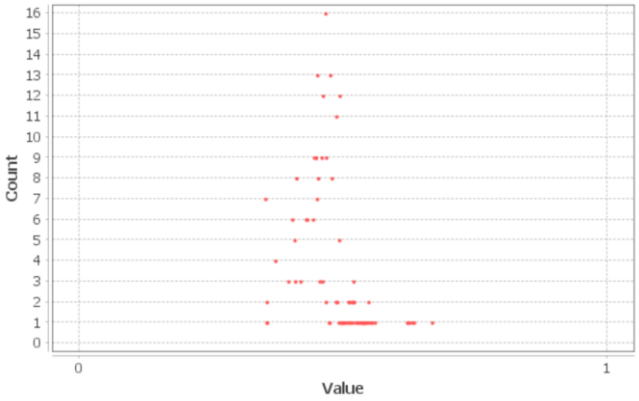
Closeness Centrality distribution of data from Twitter #Antivax in January 2018



Closeness Centrality distribution of data from Twitter #Antivax in January 2021



Closeness Centrality distribution of data from Twitter #VaccinesKill in January 2020



Closeness Centrality distribution of data from Twitter #VaccinesKill in January 2022

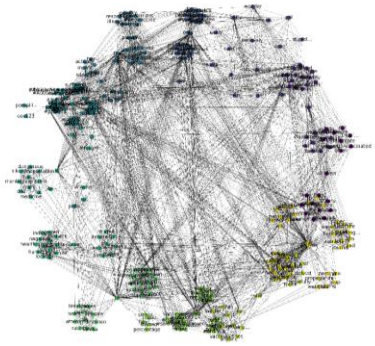
d. Results of Twitter

The Louvain method is an algorithm to detect communities in large networks. It maximizes a modularity score for each community, where the modularity quantifies the quality of an assignment of nodes to communities. This means evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network.

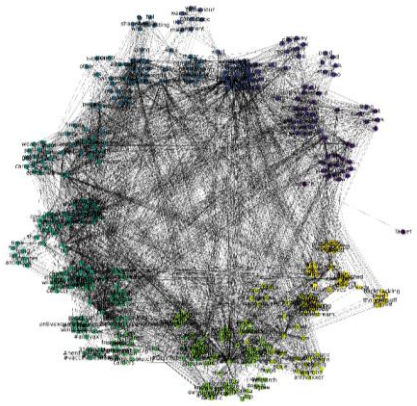
The Louvain algorithm is a hierarchical clustering algorithm, that recursively merges communities into a single node and executes the modularity clustering on the condensed graphs.

Louvain is a hierarchical clustering algorithm. That means that after every clustering step all nodes that belong to the same cluster are reduced to a single node. Relationships between nodes of the same cluster become self-relationships, relationships to nodes of other clusters connect to the clusters representative. This condensed graph is then used to run the next level of clustering. The process is repeated until the clusters are stable.

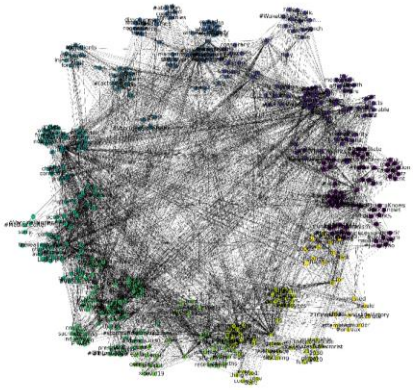
In order to demonstrate this iterative behavior, we present such complex graphs.



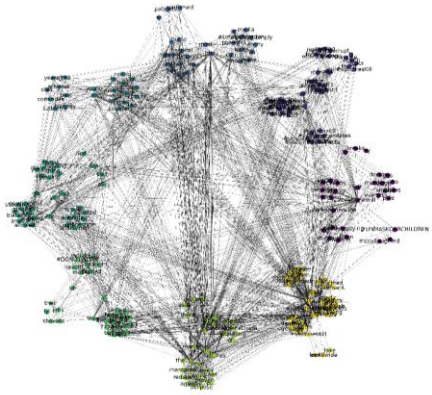
The Louvain method of data from Twitter #Antivax in January 2018



The Louvain method of data from Twitter #Antivax in January 2021



The Louvain method of data from Twitter #VaccinesKill in January 2020



The Louvain method of data from Twitter #VaccinesKill in January 2022

Authors

Nadezhda Manuilova: Twitter Data, Twitter Python code. Visualization of networks. Analysis Centrality. Results of Twitter. Gephi plots.