The BRITISH UNIVERSITY IN EGYPT

Infection Spreading Topologies and visualization

Web Topologies
2021CNM601

Alaa.Hesham

Table of Contents

Introduction

Diseases are one of natural disasters. It is always necessary to take precautions against those diseases to limit its spread and protect human health. This is achieved through risk management procedures. One of which is early detection of uncommon/ new diseases as early as possible [1]. For instance, The early reports of H1N1 flu detection in US and China [2] have participated in limiting the spread, raising awareness on how to face the virus. It also gave scientists more time to develop and test medicines to limit or cure this disease. The stages of a diseases spread are [3]

- Outbreak: which means that the diseases is spread among a group on people like a community or a city.
- Epidemic: when the spread of a virus is larger than expected and it is infecting several communities (ex: a country or several cities).
- Pandemic: when the spread of the disease is harder to control, and it is spread across the globe in several countries. (This is a challenging stage that represents an extensive spread)
- Endemic: when the disease is constant across the globe in specific countries only(ex: malaria).

Covid Dataset

Covid was announced to be considered as pandemic in March 2020[4] as it has reached the third stage of disease spread. The dataset presented in this report was constructed using analysis on data from Facebook between Jan 2020 to May 2020[5]. The purpose of collecting these dataset was to determine if covid can be detected from posts instead of tracing symptoms and cases from medical reports. The data(posts) were gathered and preprocessing. The posts that were selected belonged to users that expressed having covid. These dataset were constructed through several models; each corresponds to a certain language to determine and compare the common factor. Using LDA topic modeling algorithm as well as other clustering algorithms. The data resulted of this model

was later transformed into graph format for representation using graph gexf topic modeling algorithm.

Problem statement and Objective

The infection of diseases has causes symptoms and affected many humans. In some cases of late stages of infections(pandemics) the severity of the diseases had cased death cases. This report presents visualization graph of the infection of covid as a disease in several stages. Covid disease was chosen as it is a recent pandemic and the data on its evolution from early stages of the infection are available in the current dataset. The purpose of this graph is to highlight the factors and the insights of the stages of infection spread as well as explain the relationship between these factors.

Network Representation Process

1. Loading the datasets

2 datasets were collected regarding covid spreading. They were collected from Facebook users in English languages. The first dataset was collected between Jan2020 and 29th of Feb 2020 and the second was collected between March 202 to April 2020. The data were collected in graph format with gexf extension. The datasets were loaded in working spaces on Gephi as the following

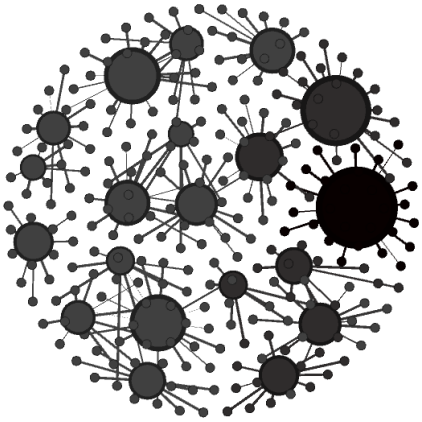


Figure 1: first dataset containing Covid Spread between Jan-Feb 2020

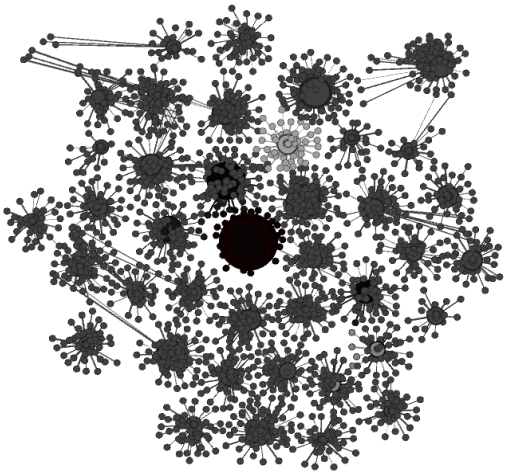


Figure 2:Second dataset containing Covid Spread between March to April 2020

2. Analyzing the datasets

The first dataset of covid FEB 2020 was an undirected graph composed of 281 Nodes and 261 edges. These numbers indicates that the graphs do not contain a large number of degrees since the difference between edges is minimal. The second graph of dataset collected in March 2020 has recognizable number of nodes and factors. Number of nodes are 3910 and edges are 3870. the difference is minimal which indicates that the average

degree will be very minimal. The objective of this report is to analyze and visualize the given graphs. Section 2 is concerned with analyzing the graph and section 3 will demonstrate the possible visualization layouts.

2.1 labeling dataset

for a better understanding of the dataset, the dataset is labeled to be able to identify the core nodes as the following.

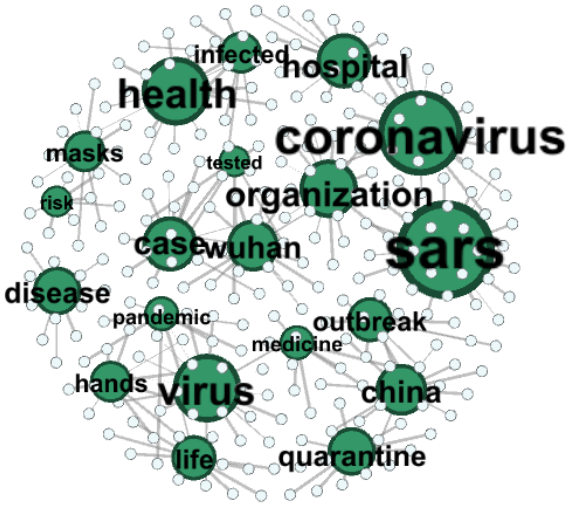


Figure 3:Covid-19 feb labeled undirected graph

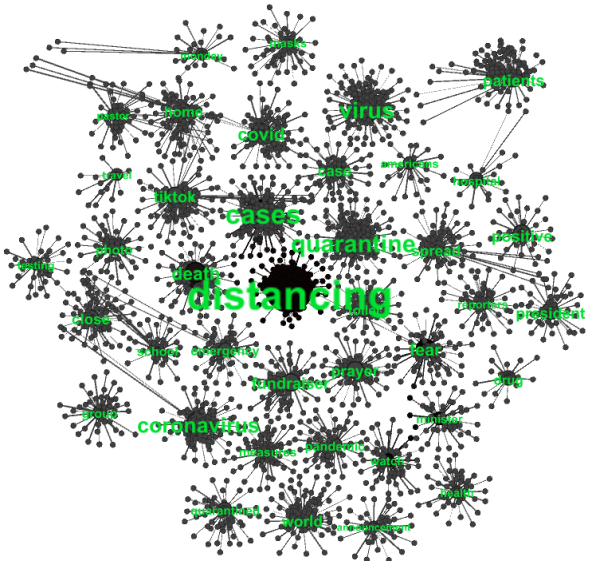


Figure 4:covid 19 march undirected labeled graph

2.2 Coloring Graph

2.2.1 coloring based on role

in order to understand the graph and the labeled nodes, nodes were colored for more depth of information. They can be colored according to a specified attributes. In this case, datasets were colored based on their label/role for instance “Wuhan” and “China” are both places, as a result they are both colored the same. Similarly, “virus” and “Sars” were colored the same. Also, nodes related to actions or words related to pandemics like “outbreak”,” quarantine “,” cases”,” masks” were colored the same .. etc.

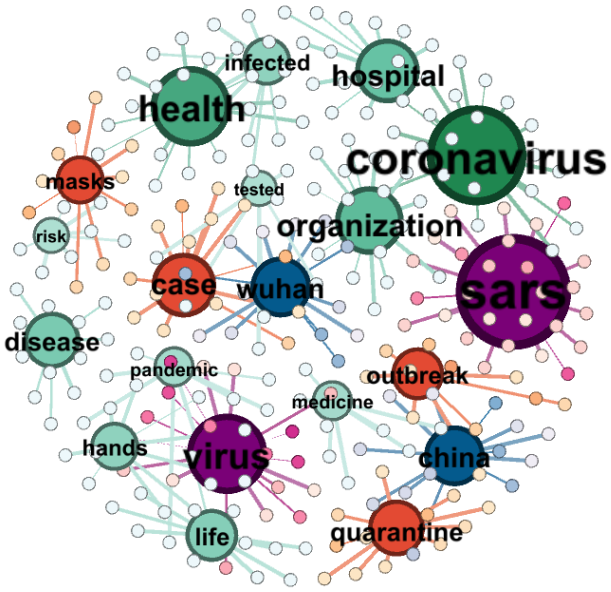


Figure 5:covid Feb dataset colored based on role

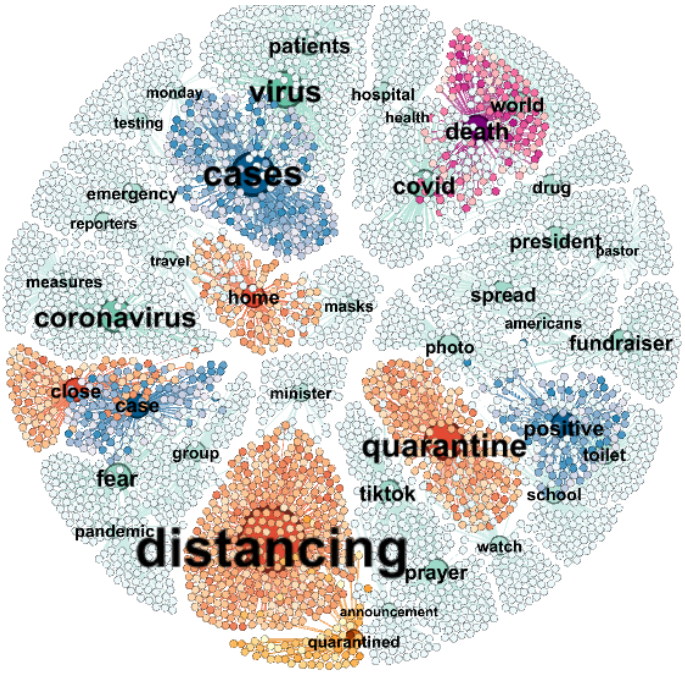


Figure 6:colored labeled graph covid march 2020

On observing this graph colored based on role. Positive and cases have same color which indicates similar or related meaning. Also, it is noticed that death is associated

with world. Moreover, distancing, close and quarantine has similar role/meaning. Additionally, most of the other colors are related to the patient or virus which indicates their strong relationship.

2.2.2 coloring based on degree

on coloring the first graph based on the degrees. It colors the majority of the graph based on the more probability of the same degree. In this case most of the graph has nodes with one degree which was predictable.

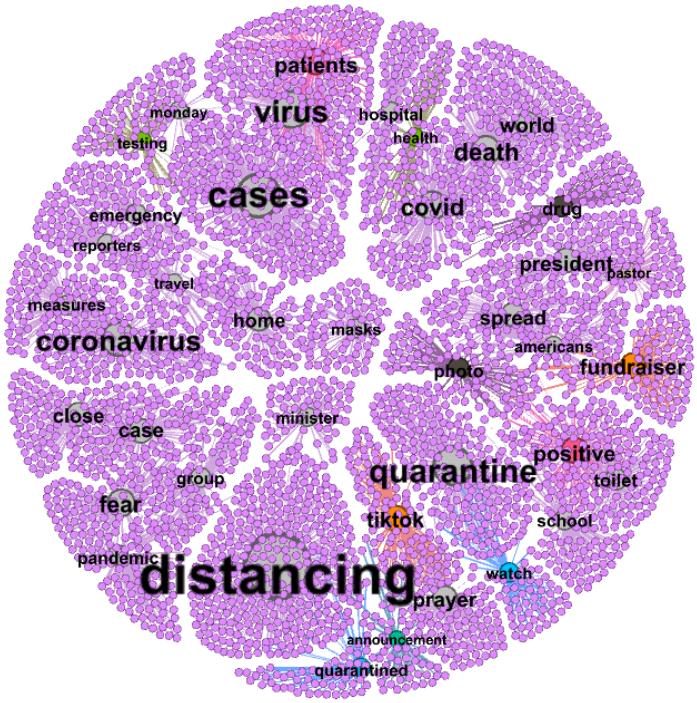


Figure 7:labeled graph covid March 2020 colored based on Degree

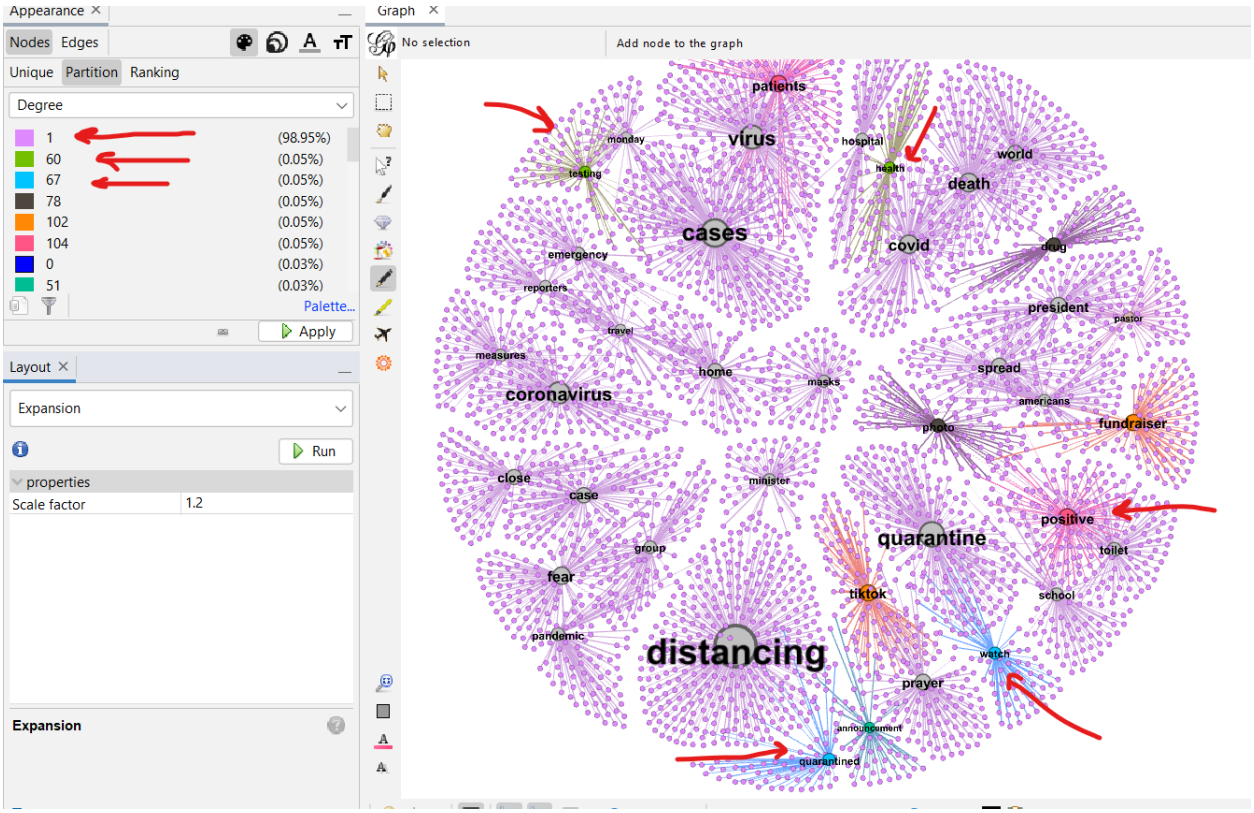


Figure 8: colored graph based on degree

Figure 8 illustrates that based on the degree and its probabilities they are assigned to a color. This color represents the nodes that follows this criteria. As shown that 98% of the nodes in the graph were in violet color as they are nodes of degree one. Similarly other degrees with red arrows are illustrated in the graph. For instance, positive node is colored in pink has degree 104 which is greater than quarantine node. This is reflected in the color of the edges and the size of the nodes. This explains that nodes “distancing” then “cases” and “quarantine” were the biggest degrees. This indicates their high influence on the graph

On analyzing the scale of degree on the first graph which is noticed to have a smaller number of nodes and degrees in the graph It is noticed that the biggest nodes are Sars and covid.

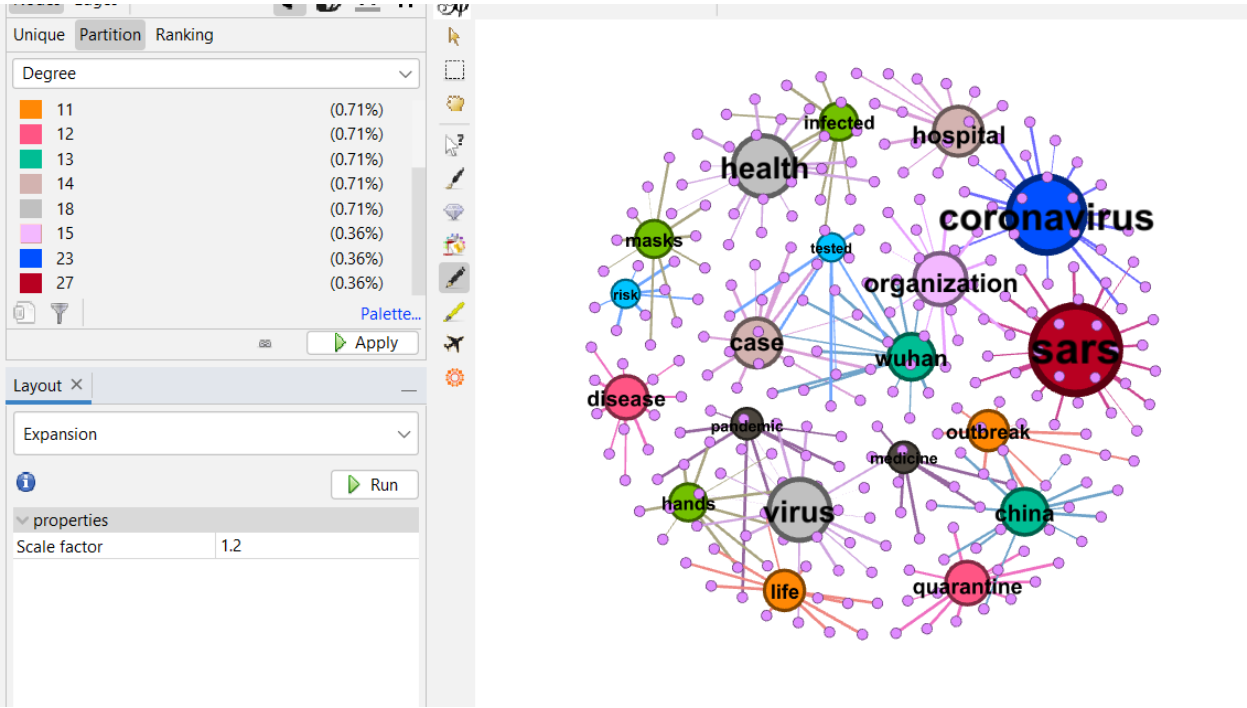


Figure 9: colored graph based on degree for Feb 2020

2.2.3 coloring based on cluster

on clustering based on Leiden algorithm, clustering large group of nodes was challenging. Numerous clusters were detected. It could not demonstrate a clear structure to the graph. Furthermore, more processing and filtration for the graph will be needed.

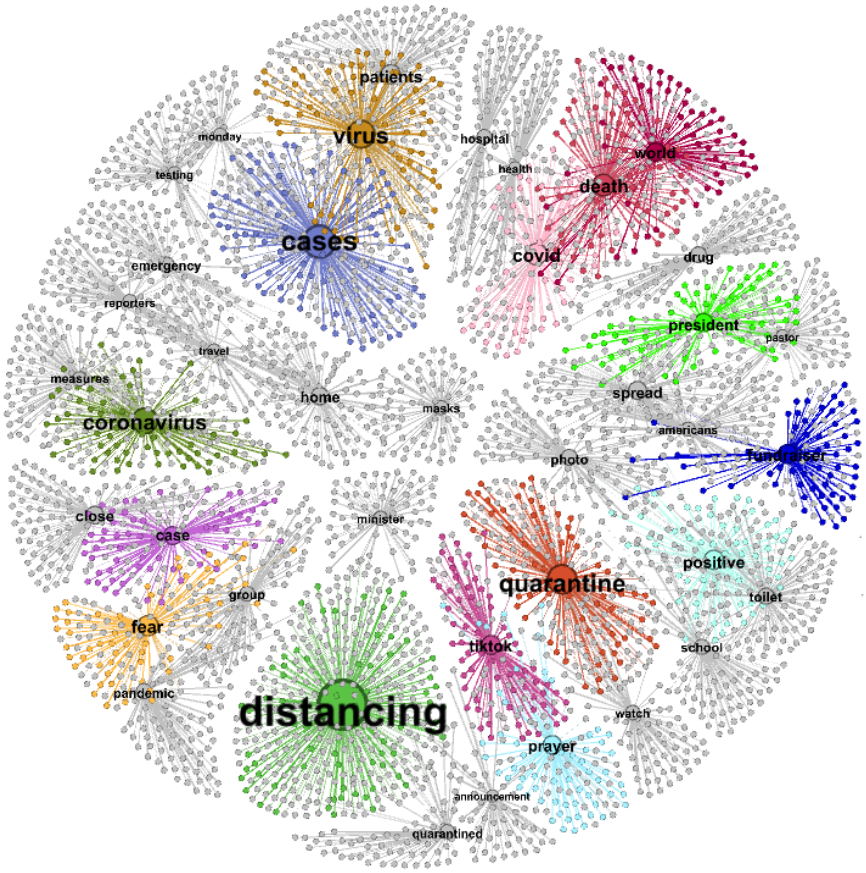


Figure 10:Clustered Graph for coved march 2020

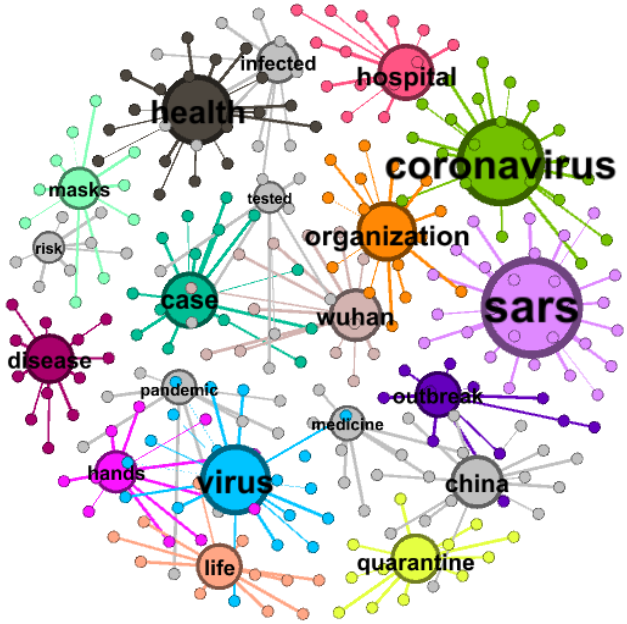


Figure 11:colored graph based on clustering for Feb 2020

It is observed based on comparing the 2 graphs that number of clusters is shorter on the Feb 2020 graph which is relative to the size of the graph. On the other hand, the graph of March 2020 is a bigger scale and clustering was challenging and it requires other better algorithms for handling large scaled networks.

2.3 graph statistics and filtering

Graph coloring and labeling were implemented to emphases the elements and highlight their roles. In other words, it gives a context to the graph and the placement of nodes. Statistics and filters helps extracts useful information by highlighting some insights and statistics.

2.3.1 Filter: Degree range

Degree range is a filter in topology section. It is used to determine the nodes with highest degree by counting edges or selected the preferred degree. This is useful in highlighting the nodes with high influence on the graph.

Figure 12 shows the degree range in March 2020 dataset. On observing the figure the highest degree is degree of 334 in the query. However the filter showed that the biggest possible nodes are “distancing “and “cases”. It indicates that these are the most important nodes that has an effect on the whole graph.

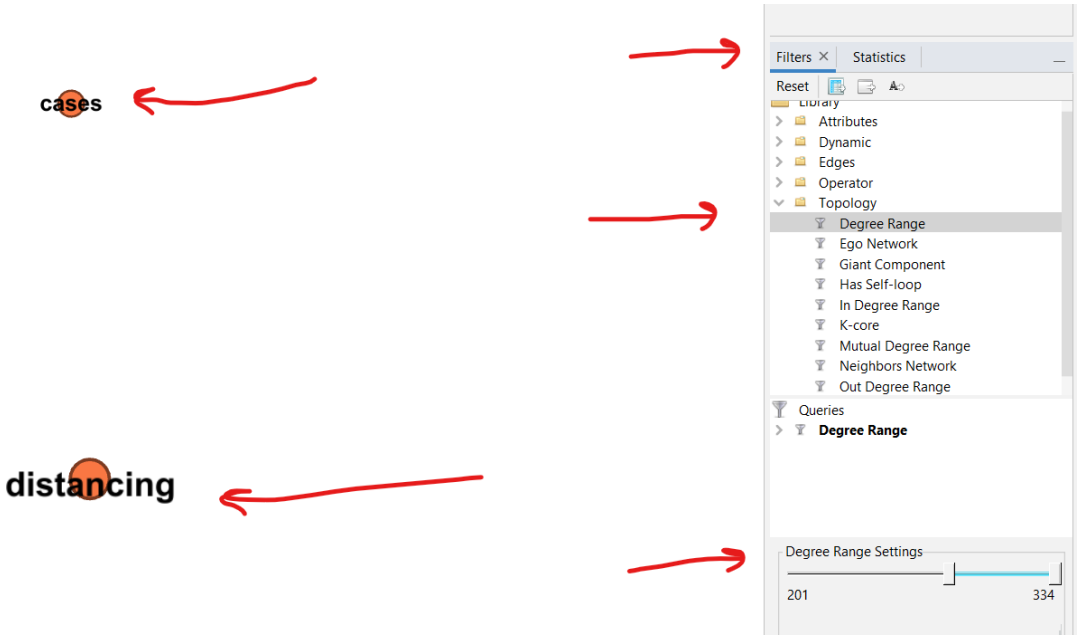


Figure 12:Graph march 2020 - filter degree range

On the other hand , figure 13 shows a different degree range relative to the scale of the graph which is smaller compared to figure 12. The highest degree range in figure 13 is 27. The influencing nodes are “health” “virus” “covid” “Sars”. More specifically, “Sars” and “Covid” which indicates that these nodes are the ones that connects the whole graph.

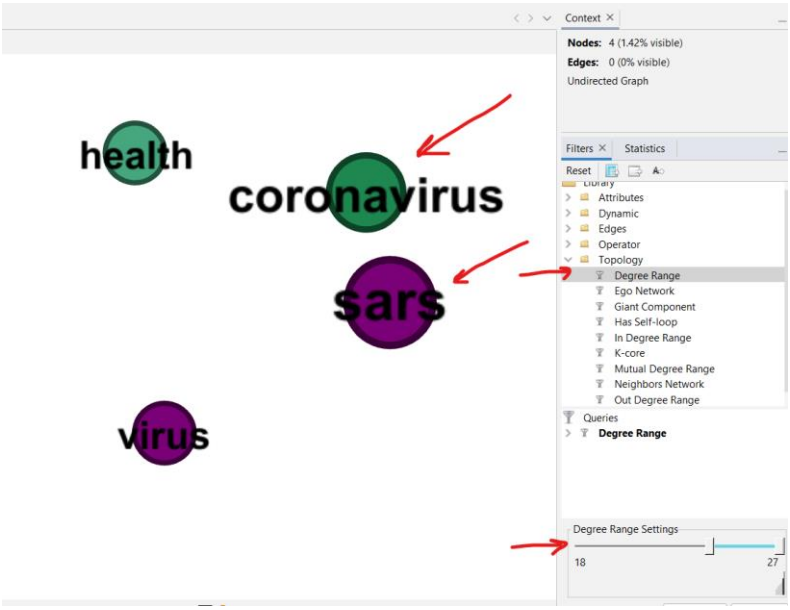


Figure 13: Feb 2020- filter degree range

2.3.2 Detecting betweenness centrality:

In the previous point the degree range has highlighted the important nodes. Yet, how much information are being transformed was unclear. To determine that on undirected graph, betweenness similarity was calculated to determine how many nodes can appear in a path. In other words, it tests the reachability between nodes. The results of betweenness centrality on figure 14 shows that most components were colored in purple which means that these components are connected with each other. Therefore, they do not need a path other indirectly connected nodes to reach the target. Additionally, the connected component report detected 20 nodes of type connected components.

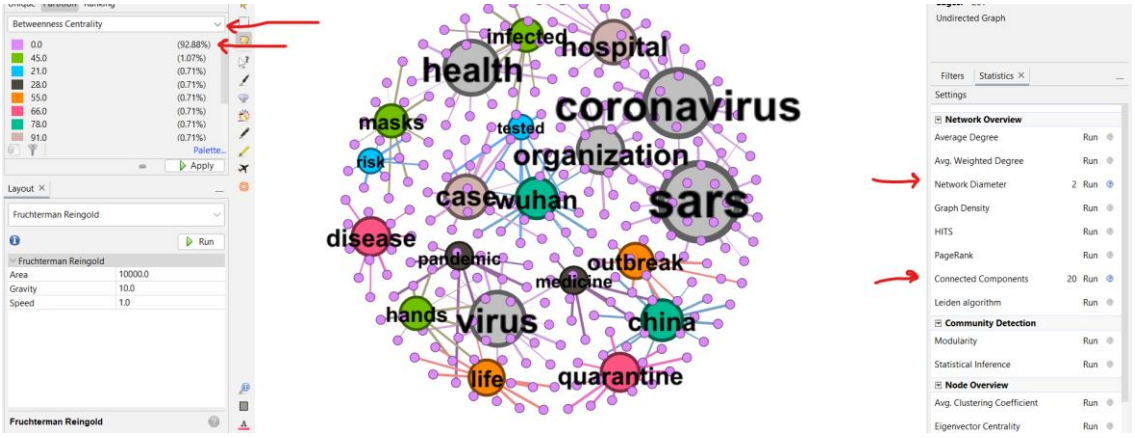


Figure 14:statsisctis report on graph Feb 2020(betweenness centrality)

The betweenness in the second graph is higher compared to the first graph ,resulting 98%. This indicates less path of several nodes and this is logical since the graph has several average degree 1 which means that most nodes are connected one to one. As shown in the figure below

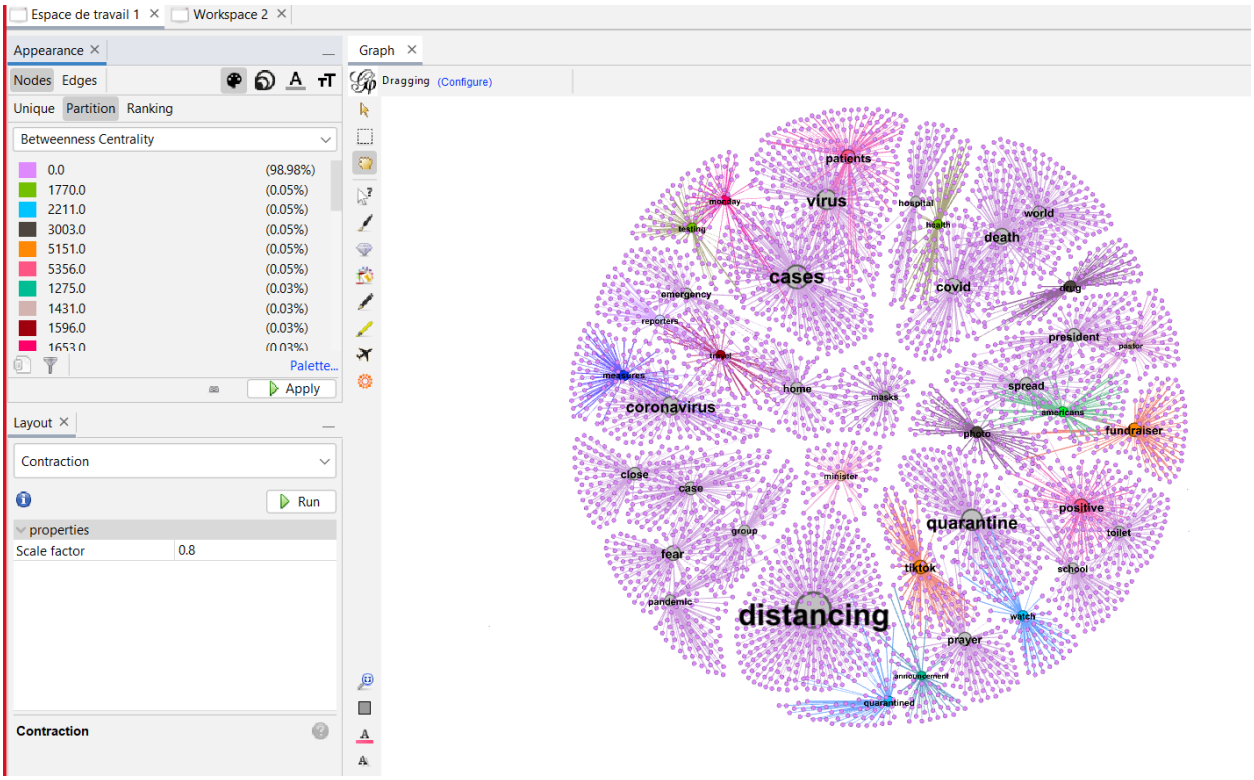


Figure 15:statistical report on March 2020 detecting betweenness centrality

2.3.3 Detecting closeness centrality

Detecting closeness was implemented to test how close are each node with each other, it had proved that the graph has more chances of being a connected graph. As the results in figure 16 illustrated high score in closeness. Which means that there are no nodes between the source and destinations nodes.

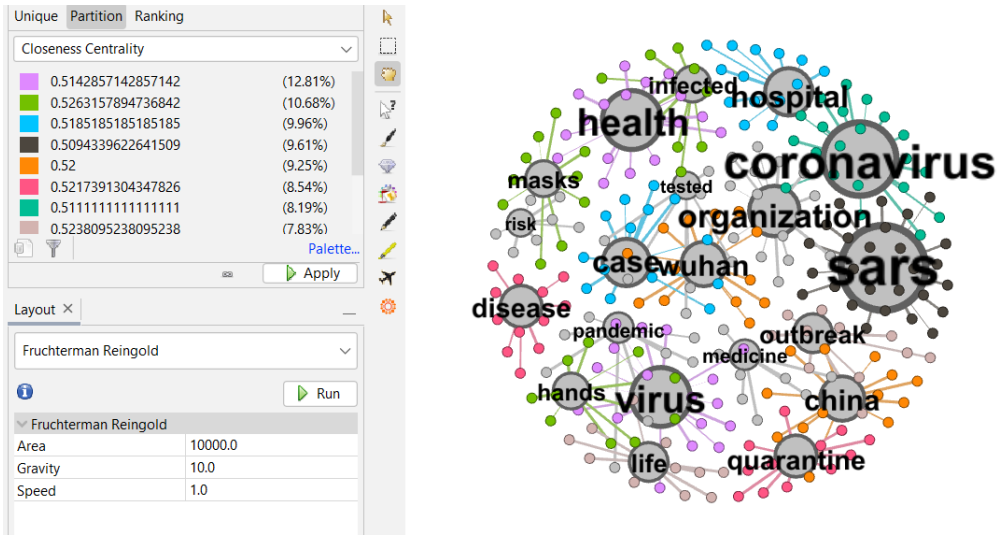


Figure 16: statistical report on graph Feb 2020(closeness centrality)

Similarly in Figure 16 : march 2020 dataset had high scores of closeness too. The colors indicate how are some components connected with their surrounded nodes in a direct association. As shown in figure 17

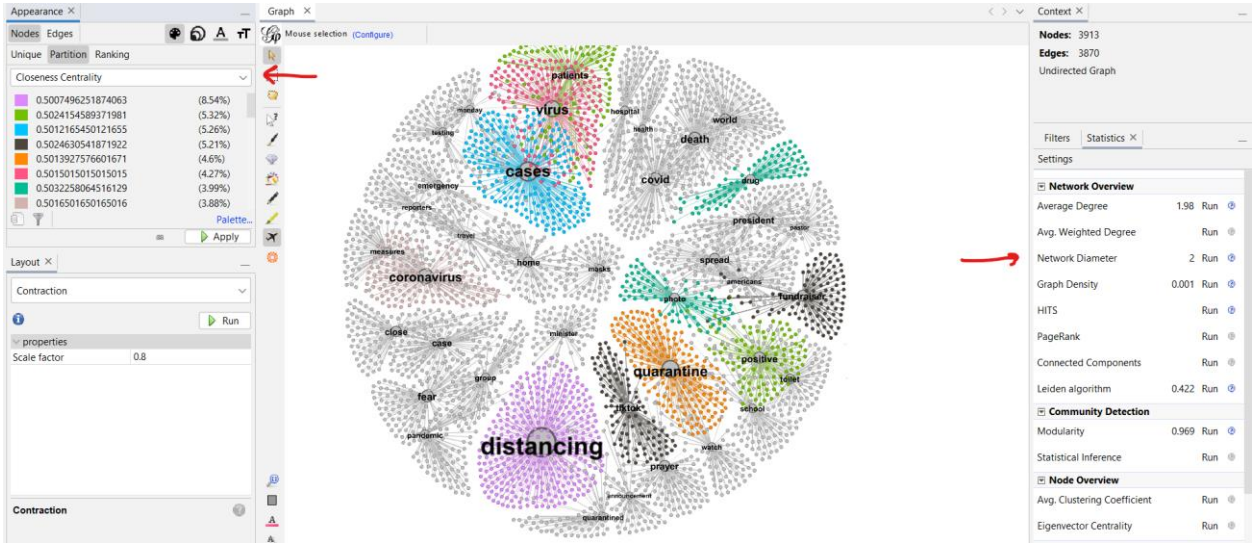


Figure 17: statistical report on graph March 2020 (Closeness Centrality)

Additional note:

Using HITS algorithm for determining hubs and authority is not recommend it gives in-accurate results. It is best worked with directed graphs to give value to the nodes in order to decides which can be authority and hubs.

2.4 graph layout(visualization)

layouts are different types of visualization perspectives that helps understanding deeper information. It aims to link relationships and behaviors of the nodes. The following sections, different layouts were chosen each highlights a perspective.

2.4.1 OpenOrd Layout

This layout was chosen as it is great at visualizing large-scaled graphs in scale of thousands to millions. It works well with undirected graph. It visualized the network from clusters perspective. The figure below shows

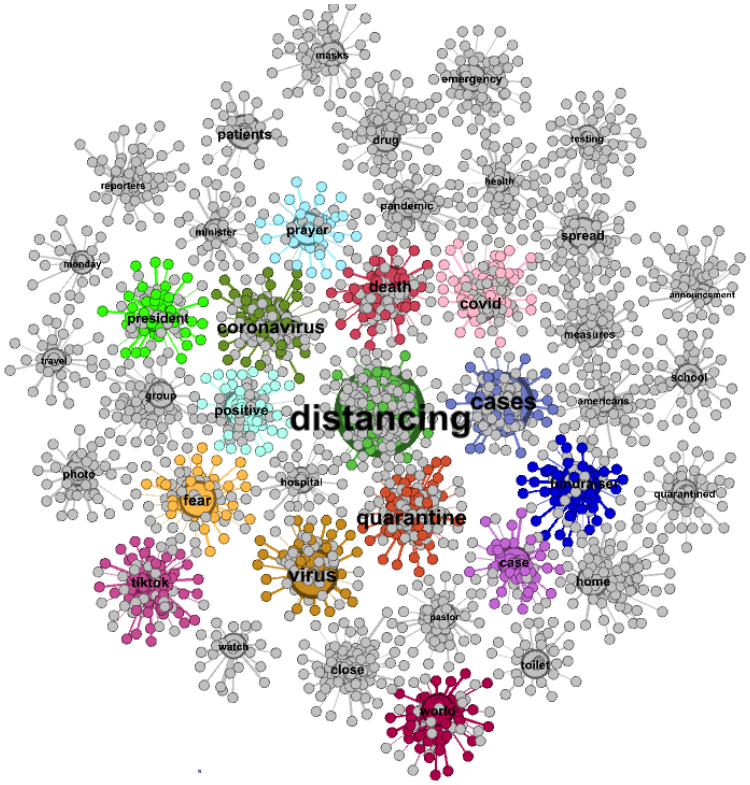


Figure 18:March 2020 OpenOrd layout

The figure below shows the association of the nodes in a circular shape. It shows an interaction (strong relationship) between “distancing” cases” “quarantine” “covid” ”death” as a primarily actor. while the rest of the nodes are associated partially with their nearest neighbors.

Additional note:

This layout is not recommended to be used with the first dataset because it is in a smaller scale. Using this algorithm will not provide useful insights.

2.4.2 Force Atlas 2 Layout

Force atlas 2 was chosen instead of force atlas because it is an enhanced version that handles networks better in repulsion condition. It uses force in simulation the reactions (attractions of nodes). It uses scale parameters and other tunning settings for better visualization. It was applied in experiment 1 on graph Feb 2020 and once again on graph March 2020 in experiment 2 In this experiment it was applied twice.

First experiment on the figure below, it tests the behavior of nodes in graph under a large force that was requested in the tuning parameters. The figure below shows that the nodes were grouped in a circular shape under force . it shows the that virus is dependent on sars and sars is dependent on covid and covid is found in organizations, more likely hosipitals…etc

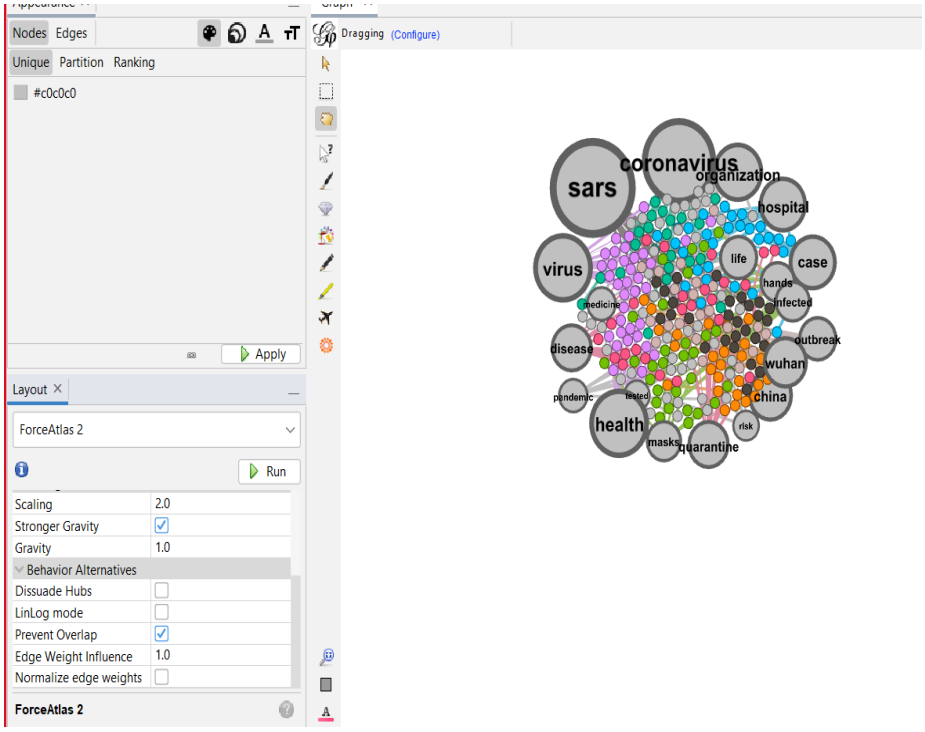


Figure 19: Force Atlas 2 lay out for Graph Feb 2020

Second Experiment: in the figure below, the tuning settings were “prevent over lapping” to have a clear image of the interaction of nodes. Also “ scaling” was 2 it was unnecessary at the moment to increase the graph parser. On running the algorithm it is observed an association between “distancing” “ virus” “covid” on an applied force.

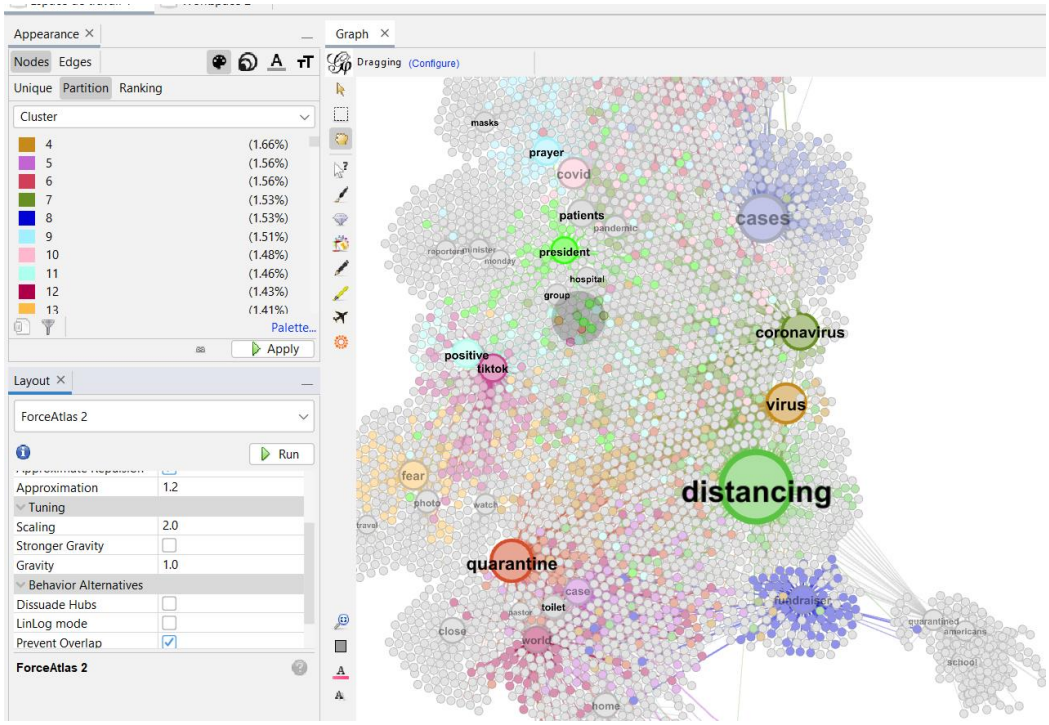
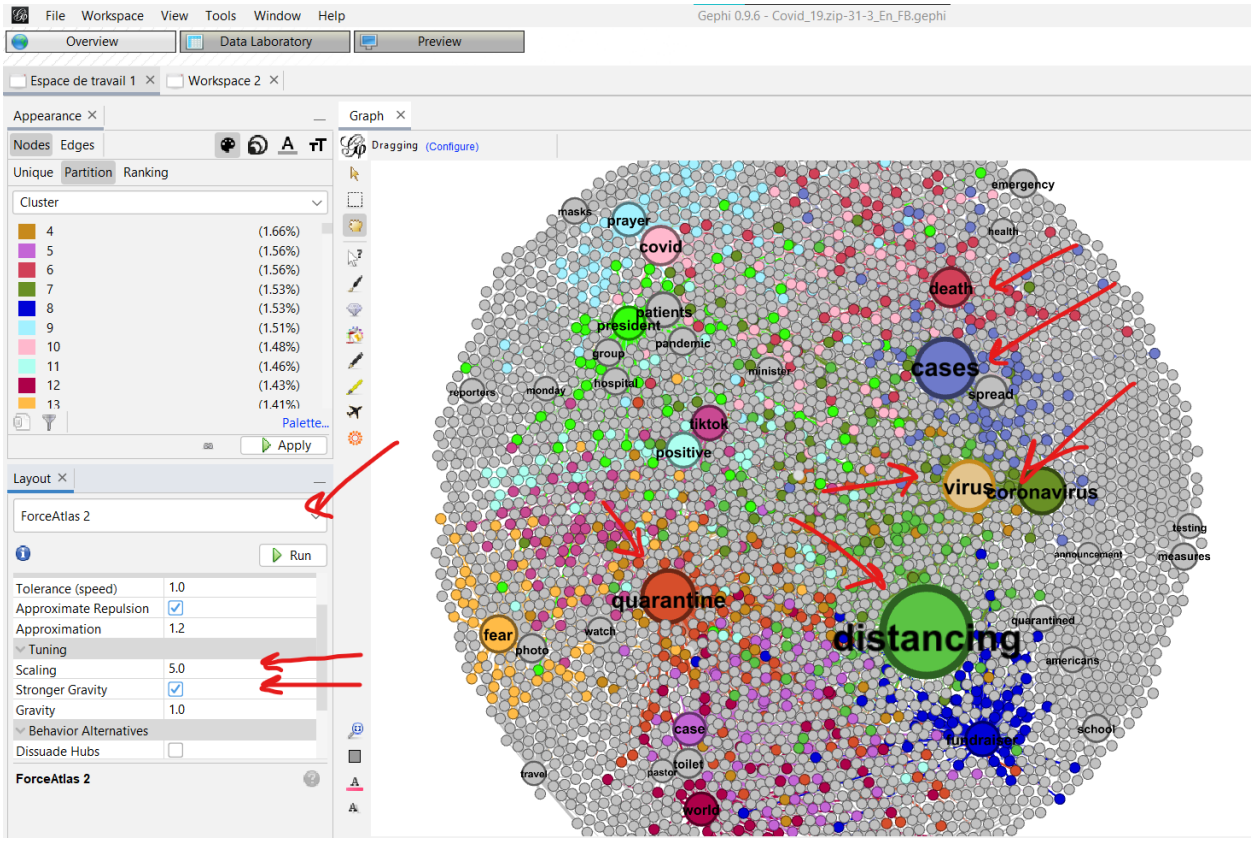


Figure 20:Force Atlas 2 Layout experiment 2

For further testing , the settings were tuned to have an intensive situation(applying more force) . as shown in the figure below



The association between “distancing and virus and covid remained. Which indicates that the relationship is still strong through the increase of time. However, the simulation shown an association or movement of nodes” cases “ ” death” and” spread” which indicates that there were still cases and there were more death .

Conclusion

the spread of diseases is dangerous. it affects millions of lives and in some cases it can be responsible for death. as a result , understanding and analyzing the stages of infections of a diseases is a necessary. this report presents the stages of infection to covid diseases. for this report 2 datasets were selected. the first was conducted in feb 2020 from Facebook posts related to covid. the second dataset was conducted in march 2020. the first dataset contained 281 nodes and 261 edges . while the second contained 3870 nodes and 3910 edges. the goal of this report is highlight the factors and the insights of the stages of infection spread as well as explain the relationship between these factors.

in the first section of analysis , the dataset was labeled according to what each node represents and then it was colored based on (cluster , degree , role) . on observing the colors of the graph based on role virus and sars had the same color and covid had a different color. it is logical in the first dataset as the disease was in outbreak stage. in the second dataset covid and virus had the same color and so is quarantine , distance and home as these were the precautions facing the infection. On observing the dgree most of them had a high probability of first degree which is predictable since the difference between edges and nodes were minor it indicates that the structure of the graph was centralized

when applying filtering and statistics, to determine the centers of the first graph , it was observed that cases and distancing , while the second dataset had covid , virus , health and Sars. For further understanding on calculating betweenness it was minimal and that is logical constituting the structure of centered point . and so for the closeness, it was minimal as most nodes are connected to a centralized node. on applying connected component report 20 component was found (centralized nodes).

for visualization several layouts were implemented. OpenORD was applied and it was concluded that interaction (strong relationship) between “distancing” cases” “quarantine” “covid” ”death” as

a primarily actor. When applying force atlas , on 3 experience . the first experiment on feb 2020 concluded that nodes were grouped in a circular shape under force . it shows the that nodes were dependent in a sequential form. the second experiment concluded that an association between “distancing” “ virus” “covid” on an applied force. and the third experiment on applying intensive force concluded that relationship is still strong through the increase of time. However, the simulation shown an association or movement of nodes” cases “ ” death” and” spread” which indicates that there were still cases and there were more death .

References

[1] Lee K, Agrawal A, Choudhary A (2017) Forecasting influenza levels using real-time social media streams. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI), pp 409–414

[2] Ding H, Zhang J (2010) Social media and participatory risk communication during the h1n1 flu epidemic: A comparative study. China Media Research 6:80–91

[3] Epidemic, Endemic, Pandemic: What are the Differences? | Columbia Public Health," *Epidemic, Endemic, Pandemic: What are the Differences? | Columbia Public Health*, Feb. 19, 2021. https://www.publichealth.columbia.edu/public-health-now/news/epidemic-endemic-pandemic-what-are-differences#:~:text=What%20does%20Endemic%20mean%3F,in%20certain%20countries%20and%20regions. (accessed Jul. 13, 2022).

[4] W. Director General, "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020, Mar. 11, 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed Jul. 13, 2022).

[5]]A. Amara, M. A. Hadj Taieb, and M. B. Aouicha, "Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis - Applied Intelligence," *SpringerLink*, Feb. 13, 2021. https://link.springer.com/article/10.1007/s10489-020-02033-3 (accessed Jul. 13, 2022).