Datathon - 3

[1]Arjun Verma : IMT2017008
CS732/DS732: Data Visualization
Course Instructors : Prof. Jaya Sreevalsan Nair
Technical Report - 3

Abstract. *This technical report contains a brief overview of the methodology involved in the visualization of the provided tabular datasets published by the World Health Organization to create new networks and visualize network communities. The report further attempts to make certain inferences from the visualizations generated.*

1. Introduction

The provided dataset contains information about different COVID related stats across different provinces in different countries. These include a number of variables such as deaths, recovered and confirmed cases along with basic demographic information. The tools that were experimented with for the purpose of this assignment include **Gephi, Networkx, i-graph and Mongkie**. The final tool that was opted for making inferences was **Gephi** as it provides a great sense of comfortability. *This assignment was proceeded by keeping in mind an overarching goal of the study of the disease with the sole purpose to track the spread of disease and how counter measures could be taken to lessen it's spread.*

2. Tools and Methodology

2.1. Tools

The exhaustive set of libraries used for generating the final inference visualizations involve :

- os
- numpy
- pandas
- networkx
- Gephi

The usage of the aforementioned tools would be elaborated upon in the methodology section. Note, only those involved in the final generation of visualizations have been mentioned. The ones that were experimented with but discarded have been omitted.

2.2. Methodology

In this section, we provide the distinct characteristics involved in each technique and the reasonings behind them. However, before we dive into the individual techniques, let's have a clarity on the basic workflow of the code.

2.2.1. General Methodology

For visualization of each of the networks, we have three stages represented by three files of code to be run to get to the final output. The workflow proceeds in this order :

- First, the data from the csv files provided by WHO are pre-processed to capture only relevant information for the purposes of visualization.
- Secondly, these pre-processed csv's are then used to generate Node and Edge Lists to provided to Gephi.
- Lastly, Gephi is used to generate visualizations and make inferences from them.

A basic framework is represented as :

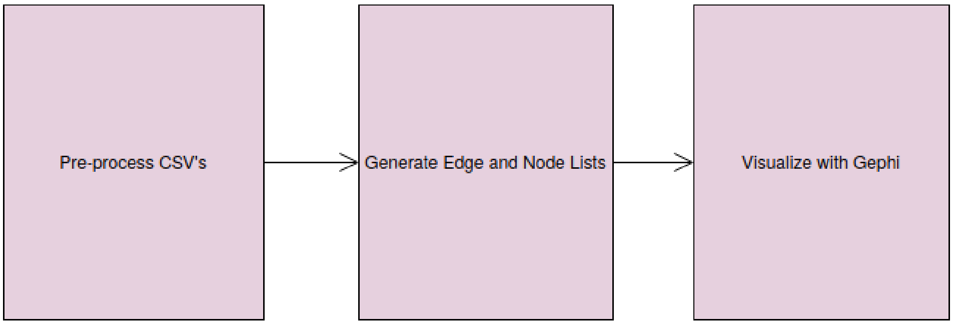

Figure 1. Framework

Apart from this, the warm up that was conducted was with the Dolphin Communities Dataset and the key takeaways from it were :

1. **Gephi** is a better tool over **Networkx** for the purposes of this datathon.
2. The idea of connecting provinces to form communities presented subsequently in the report was captured from here.
3. A **Gephi** submission on this dataset is also given in the Submission folder, however I am choosing to omit the inferences on this one as it is a widely studied dataset and the inferences can be essentially picked up from any jolly website on the net.Rather, I have presented what were my takeaways from this warm-up which later helped me in the actual datathon.

2.2.2. Methodology for COVID Visualizations

Distinct techniques (characteristics) for the network visualizations :

- Layouts Tested : Force Atlas 2, Yifan Hu Province, Fructerman Reingold, OpenOrd
- Color Map for Nodes and Edges : Gradients of singular colors
- Edge Weighting Technique : Interval Creation
- Interval Scale : Logarithmic
- Values Utilized : Provinces, Countries, Deaths, Recovered, Confirmed

Reasonings for important characteristics of the visualization :

- The case study of utmost importance that one can think of while going through such a dataset is keeping a track of the disease. Visualizing the flow of disease and it's epicentres is of high priority to control the spread of such diseases. To study this effect, I wanted to go for an inter-country province analysis, according to different dates of the year. Firstly, one simply uses the groupby() functionality from pandas to create subsets of the dataset according to the different dates on which the data is collected and then studying it for the date you want. For the purposes of this report, we're specifically focusing on the interval of April, particularly the **9th of April, 2020** when countries were going under full fledged lockdowns.

- For the purposes of our **inter-country province analysis**, I first grouped the datapoints by countries they are located in. All members of these group would be connected with each other via edges and then the spread of the disease in each country would be tracked via the connections through provinces. Next up comes up determining the weighted edges for these connections. The weights for the edges were based on the average of variables for each province for the study in focus. For example, when creating the *Deaths Network*, the edges between provinces are determined by the average number of deaths in both the provinces and so forth and so on for the rest of the attributes.

- The created nodes and weight csv's are then imported into Gephi to generate visualizations. We first begin by removing seemingly disjoint points from the network by doing a *degree filtering* and removing nodes with degrees less than equal to three. Different Layouts are then tested to make different inferences.

- Now comes the selection of appearance for our visualizations. For this purposes, we first compute the *network statistics* as so graciously provided to us by Gephi. We then use values computed from these statistics such as betweenness centrality; degree centrality as ranking attributes and choose a singular color gradient to map to these values for nodes. The weights of the edges are mapped to another singular color gradient for edges. For the purposes of *Deaths Network*, the nodes are displayed along a gradient of blue while the edges are displayed along a gradient of red. For the *Recovery Network*, the edges are displayed as a gradient of green. Note these are different networks across different visualizations and it's not that green and red are being used together to keep the visualization color blind friendly.

- With the visualizations in place, we now proceed to see the sample visualizations generate and would then proceed with the inferences.

2.2.3. Sample Visualizations

1. The **figure below** depicts the individual country clusters being formed. I am aware the size is extremely small however please bear with me and just focus on the 7 clusters (5 are quite clearly in focus, one will have to strain hard to see the other two) being formed depicting the number of countries in focus after filtering. We will magnify these images and focus on individual inferences from the countries as we move along.

Figure 2. COVID country clusters in the Deaths Network

2. The **figure below** shows the visualization after the application of the Fructerman Reingold Layout to visualize the *Deaths Network*.

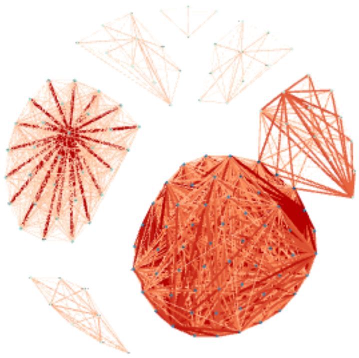

Figure 3. Reingold Layout for the Deaths Network

3. The **figure below** shows the visualization after the application of the OpenOrd Layout to visualize the *Recovery Network*.

Figure 4. OpenOrd Layout for the Recovery Network

4. The **figure below** shows Mainland China in focus in *Recovery Network*. An interesting thing to focus on in theh *Recovery Network* is that the nodes on the boundary have been colored darker while the nodes in the central are lighter. This was deliberately done using **invert color scale** to better visualize that the lesser connected nodes in the recovery network exhibit lesser degree of recovery and hence are more ”dangerous” zones.

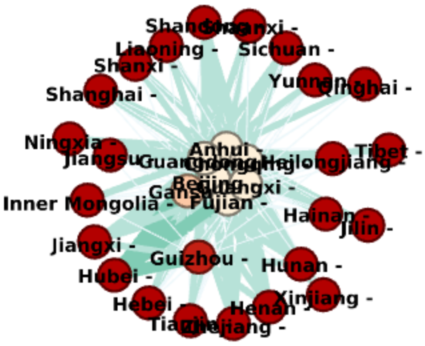

Figure 5. Mainland China in focus in *Recovery Network*

2.2.4. Inferences

Now, that we have had a look at sample visualizations, let us look at some interesting inferences that ne can make from the visualizations.

1. Let us begin with the explanation of the most absurd thing noticeable. The combination of 2 clusters Canada and USA in the *Deaths Network*. Why is it that all other countries are separate but these countries conjoined? The reason for this is the cruise ship **Diamond Princess** which contains members from both USA and Canada. Our Dataset apparently captures the nationality of the people on board and hence **Diamond Princess** ended up being a node in both the cluster resulting in the conjoining. When one focuses on the color of the edge weights, one notices that these are in proportion with the actual observed stats. The same explanation goes for the rest of the two common nodes, **Grand Princess** and **Recovery**, where **Recovery** most probably depicts a common quarantine centre.

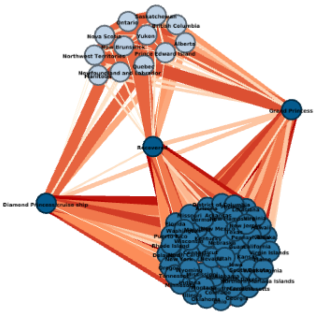

Figure 6. US and Canada in the Deaths Network

2. Taking the Reingold figure from sample visualizations again, one thing to notice here is the stark contrast in ”redness” of countries, where USA and Canada are severely red, with USA somewhat spreading that ”redness” to Canada and meanwhile the apparent ”declining redness” for Mainland China. These again are in alignment with the stats observed in real life during the timeframe of **April**.

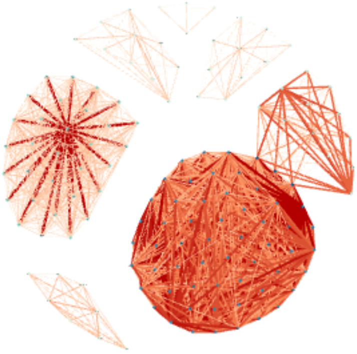

Figure 7. Mainland China is on the left while US is the big red blotch with a small protruding Canada on the top

3. When we focus upon USA, it is just as we expected where **new York** and **Washington** are racking up the reddest areas. These are two of the busiest hubs in the USA and the network captures that. It is these areas which would have had to be locked first to prevent the spread of the virus. One can see from the red edges from these two provinces to a number of other provinces and see how these two are the common factor in the spread of cases being observed.

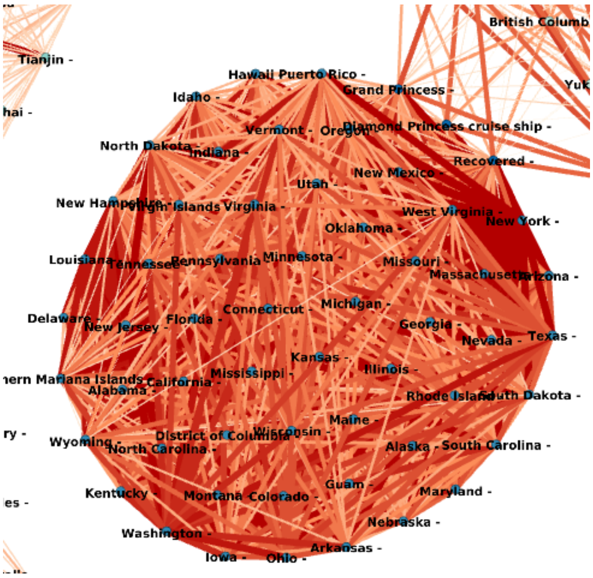

Figure 8. Purely focusing on USA

4. Lastly focusing on the most obvious thing that everyone was expecting. When we focus on Mainland China, it is none other than Hubei province, which is the sole curator of all problems. Wuhan, the onset of the virus is the capital of Hubei and the graph could not have more beautifully portrayed Hubei as the epicentre.

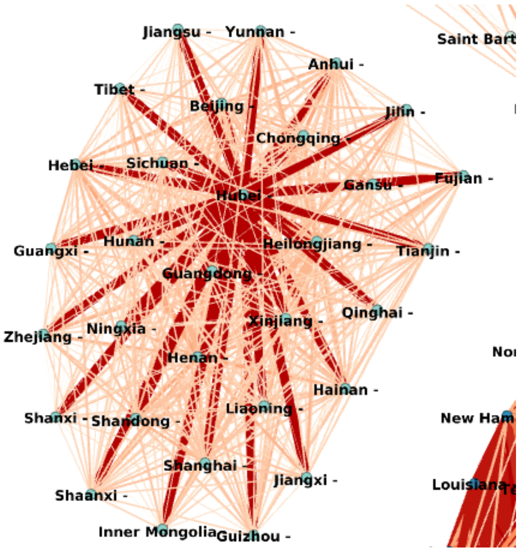

Figure 9. Purely focusing on Mainland China

5. Let us now have a look at the recovery data. When one sees it, it kind of looks empty and somewhat suspicious considering that the one major cluster being shown is of Mainland China. However, on careful evaluation one must recall that the almost all countries apart from China were on the steep rise of the curve. There were barely any recoveries during this stage. Only countries which would have been effected for a while would have any recoveries and this is what is portrayed in the below graph.

Figure 10. We have a cluster of Australia on the left and Mainland China on the right

6. Focusing on Mainland China again, but this time in *Recovery Network*, we see that it is the three provinces which were the epicentres four months ago from the time frame with the most recoveries. This is expected as it is these provinces which would have resolved most cases over time. Another thing to notice is the hub being formed in the centre, it is this recovery hub, from where medical emergencies were being sent out as the ones who had dealt with the virus the longest had some fighting supplies which were being transported.

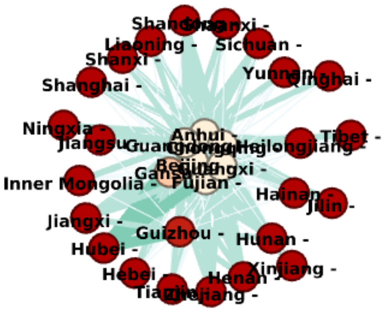

Figure 11. Mainland China in Recovery Network

3. Other notes

1. Note, the inferences presented were being looked for across a number of dates and the same visualizations were run a number of time. After careful consideration, **April** was chosen to be the most optimal timeframe to be presented in the network as it was the period of boom in the number of cases in the rest of the world.

2. Obviously, the next point that comes to mind is a dynamic visualization and yes, attempts were made at it however it was unsuccessful. Generating an animation by taking multiple screenshots seemed to be too time cumbersome a task and hence was chosen to be omitted.

3. As a substitution to the dynamic graphs, correlation networks were considered by gathering data across dates and also to capture correlation among attributes. However, as mentioned in the code comments as well, this did not prove much helpful in making inferences for the question being studied. A **sample image** of the correlation network between certain attributes is given below just for trials' sake.

4. Keeping comments from the last Datathon in mind, I have added references as a separate section rather than hyperlinks. I still however chose to go with bullet points as it gives a better structure to my inferences and methodology. (Just a personal bias for the purposes of assignments/datathons, hope it's not an issue.)

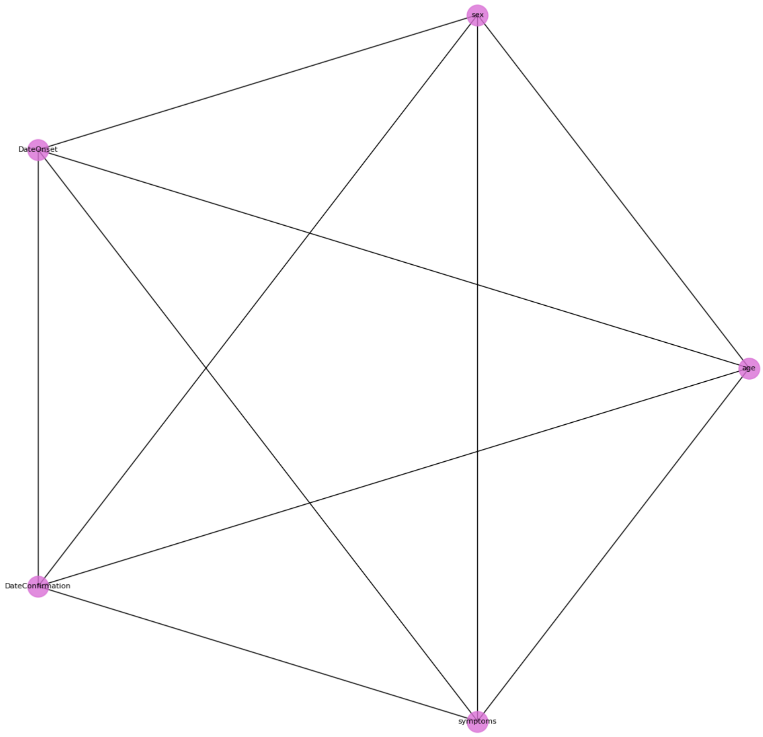

Figure 12. A sample correlation network of attributes

4. Conclusion

In conclusion, it is worth to mention that the report has successfully elaborated on all the deliverables. The tools and methodology section entails the questions :

1. Which parts of the dataset were you able to use, and how have you been able to use?

2. Which visualizations did you choose, why, what technologies (Python libraries, others) did you use for the visualizations ?

The indicative tasks of making inferences from the network communities has also been elaborated upon. The key takeaways from the warm-up has also been talked about.

5. References

1. For correct choice of Gephi Layouts
2. Hubei Timeline
3. Diamond Princess and Grand princess Stats
4. A timeline of COVID stats across the World.