DATATHON III : COVID-19

Gandharv Suri
IMT2017017
International Institute of Information
Technology, Bangalore

***Abstract*—In this report we present the results and the methodology used to visualize the data of confirmed, recovered and deaths reported due to covid-19 in various provinces in multiple countries.**

I. Introduction

Recently in 2019, Wuhan, China experienced an outbreak of novel coronavirus [1] and reported exponentially increasing cases. The virus was named as covid-19 by Chinese scientists. The virus is reported to be of the category of β group of coronaviruses. The virus has affected (reported cases) 36,152,949 people and has taken 1,056,318 lives according to [2] till the time of writing this report.
We try to visualize the data publicly available by the World Health Organisation (WHO). We first visualise the time series data starting from January 1st 2020 with the last entry at September 23rd 2020. Later in the report we visualize the data of individual people who were reported to be coronavirus positive and study their travel history.

II. Problem Statement

Use tabular dataset published by the World Health Organization to create new networks and visualize network communities.

III. Data

The data is published by World Health Organization (WHO) and is publicly available. The data consists of following kind of files.

- Time series data for confirmed, recovered cases and deaths reported in various provinces of total of 188 countries.
- Time series data for confirmed cases and deaths reported in various provinces of United States of America.
- Data of patients along with their age, sex and location and recent travel history.

For the current project we have not used the whole data. We used the time series data of the provinces of all countries and did not focus only USA. And we used the data of positive tested patients and study their travel history.

IV. Methodology

For the data of provinces we club the provinces according to the countries. So each country in the data represents a node in the network. The countries which observe the affect of the coronavirus at the same time have an edge between them. We create a network for each of the three data parameters (confirmed, recovered cases and deaths).

We try to study the outbreak according to time weighted average of the function along with time. Mathematically taking an example of a function $f(x)$ the average of the input variable $\tilde{x}$ could be calculated as

$$\tilde{x} = \int_{x=0}^{x=T} xf(x) / \int_{x=0}^{x=T} f(x)$$

Later we cluster in three transition groups to differentiate in the countries who were affected by virus first from the ones affected later.

So for each node we calculate the average time according to the following steps

$$weightedsum = \sum_{i=0}^{i=T} i \times datevalue$$
$$total = \sum datevalue$$
$$averagevalue = weightedsum/total$$

After obtaining the $averagevalue$ for each node we build a complete undirected graph with edge weights according to the average value of the nodes at the ends of the edge. We define two attributes of each edge $weight$ and $distance$.
$weight$ and $distance$ between two edges a and b is defined as

$$weight_{a,b} = 1/|averagevalue(a) - averagevalue(b)|$$
$$distance_{a,b} = |averagevalue(a) - averagevalue(b)|$$

Thus with nodes (countries) where the pandemic nearly together would have comparatively lesser $distance$ between them and $weight$ corresponding to the edge between them would be high.

For the data of patients and their travel history, we map cities which see any travel history between them. So each city/province represents a node in the graph and each edge between the nodes represents that there was a traveler between them. The $weight$ of the edge represents the number of travelers between the cities.

The data was in tabular form with the city, where the patient was tested positive as a single column and the travel history as human written sentences. The travel history was majorly a single city of a two stop path from one city to to another to the final location separated by the keyword 'via'. To extract the locations form the travel history sentences we first remove all the punctuation and the 'stopwords' used in english language. In the remaining text we check for the keyword 'via' to check if the patient travel directly from one location to another or visited another city while travelling. Once the cites are extracted we join the with an directed edge according to the direction patient travelled.

V. Visualizations

A. Confirmed

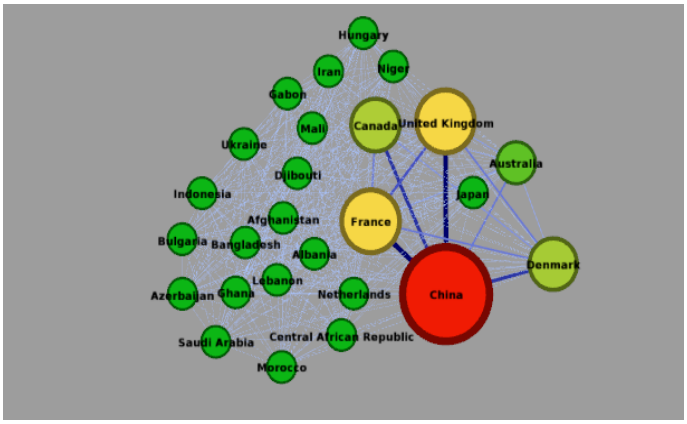

Fig. 1: Confirmed Cases

We visualize the network between the countries for the data of the confirmed cases. Here the size and the color of the nodes is according to the weighted degree ($\sum weight_{ij}$)of the nodes, with the color red being the highest with yellow as intermediate and green being the lowest. The edge color and thickness is representative of the weight of the edge with a sequential colourmap with dark blue as the highest weight and lighter shade as the lower weight.

For Fig. 1 we visualize the filtered out countries with weighted average time in the range [120,175304] units. We observe that China being the epicentre of the outbreak, as expected, and which highly affected countries like France, United Kingdom very strongly and to some extent the countries Canada, Denmark and slightly Australia. We observe that these countries affected strongly to Japan specially Canada and Denmark. Other Asian and African countries were also affected through Europe.

B. Recovered

For the data of recovered patients we follow the similar color scheme and size and color the nodes according to the weighted degree of the nodes. And similar mapping scheme for the edge wrights as well.

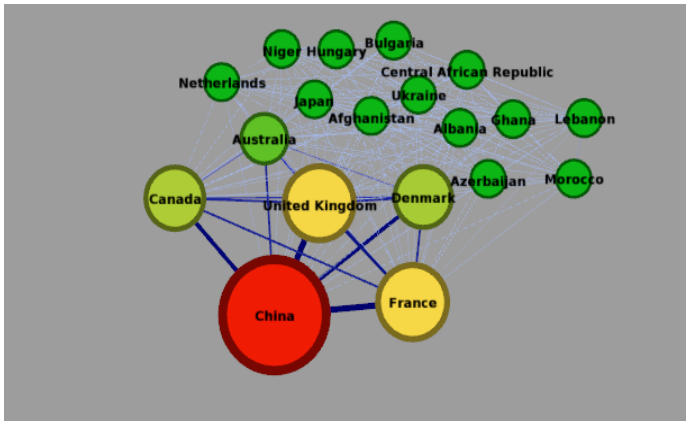

Fig. 2: Recovered Cases

In Fig 2. we have filtered the countries with the weighted average time in the range [130,175300] units. We observe that the count of number of recovered cases for countries China, France, and United Kingdom were very close. The countries Canada and Denmark also are close to the recovery rates with China. Along with that we observe that Australia is moderately linked with China, Canada, Denmark and United Kingdom.

C. Deaths

For the deaths reported due to coronavirus we follow the similar color mapping and size mapping for the nodes according to the weighted degree of the nodes. And similar mapping scheme for the edge wrights as well.

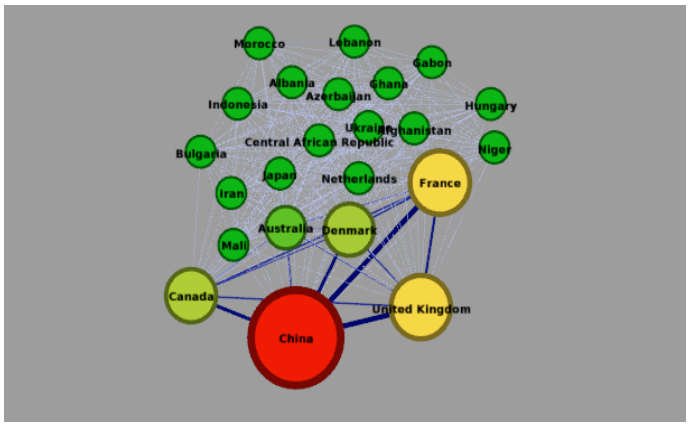

Fig. 3: Deaths

In Fig. 3 we have filtered the countries according to the weighted average time in the range [125,175304] units. We observe that China takes the leading role in deaths reported and we observe similar affect in France and United

Kingdom. The three countries are observed to affect Canada and Denmark and Australia. We observe that these countries further are connected to other African and Asian countries and few European countries.

D. Travel

We visualise the data of patients who were tested positive. The method used to construct the graph was explained in the methodology section.

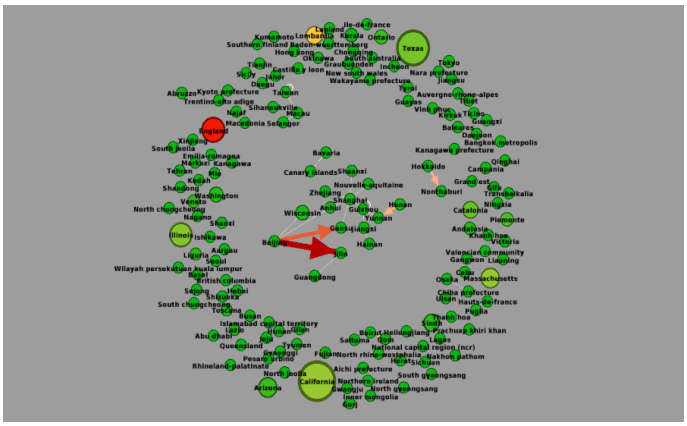

Fig. 4: Deaths

We have not filtered the nodes in this case. We opted for a different size mapping and color mapping scheme for this case. The size of the node is mapped to the number of confirmed cases and the color of the of the nodes represents the number of deaths reported in the province with a red being the highest and dark green being the lowest, and yellow indicating intermediate number of cases reported in the province.

For the edges, we follow a sequential colormap with red being the highest weight which indicates highest number of travellers between the two cities.

We observe that England as a whole saw the largest number of deaths reported due to the coronavirus, the next we observe the Lambardia with intermediate number of deaths. Along with that we observe that Texas and California have the highest number of cases reported but the number of deaths are low for the two provinces.

Observing the travel history of the patients we observe the highest travellers from Beijing to Jilin and Gansu. We also observe that many travellers left Shanghai for different locations. Henan to Yunnan and Hokkaido to Nanthburi also observe many travellers.

The data for the travel history wasn't really rich in our opinion so the network formed isn't dense. Also the data total confirmed cases and death for all the provinces wasn't available.

VI. Acknowledgements

Medium blog post — Valeria Cortez — Link
Kaggle Nootebooks — G-Dant — link

A. References

[1]Muhammad Adnan Shereena, Suliman Khan, AbeerKazm, Nadia Bashir, Rabeea Siddique *etal* "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses"

[2] Worldometer/coronavirus — link