CS732/DS732: Data Visualization – Datathon 3

Kunika Valecha

October 06, 2020

Contents

Part I
Karate Club Dataset

1 Problem Statement

To generate and study Node-Link Graph visualizations on COVID-19 data set using the python libraries like networkx and tool Gephi.

2 Data Description

The data source is Karate Club Dataset.This is the well-known and much-used Zachary karate club network. The data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is undirected. An often discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers.

2.1 Metadata

- name: Zachary karate club
- code: ZA
- url: karate club data
- category: HumanSocial
- description: Member–member ties
- long-description: This is the well-known and much-used Zachary karate club network. The data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is undirected. An often discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers.
- entity-names: member
- relationship-names: tie
- extr: ucidata

3 Data Reading

- **In Gephi** input the .gml file as it is. The gephi's default interpreter reads the nodes and edges from it.
- **In Python** read through pandas, convert it into a dataframe.

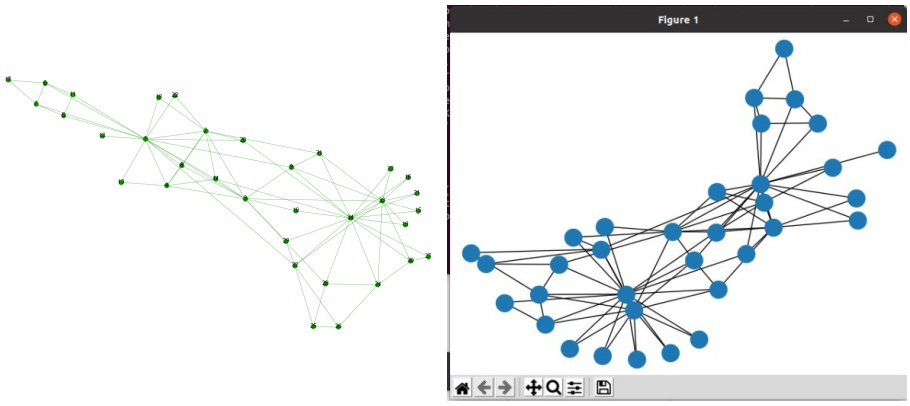
4 Undirected Graphs

An undirected graph is graph, i.e., a set of objects (called vertices or nodes) that are connected together, where all the edges are bidirectional. An undirected graph is sometimes called an undirected network. In contrast, a graph where the edges point in a direction is called a directed graph.

When drawing an undirected graph, the edges are typically drawn as lines between pairs of nodes.

One can formally define an undirected graph as G=(N,E), consisting of the set N of nodes and the set E of edges, which are unordered pairs of elements of N.

5 Nodelink Layout for Karate Club dataset



(a) Gephi output

(b) Networkx output

Figure 1: Figure a, b showing undirected graphs for karate club dataset.

Part II
COVID-19 Dataset

1 Problem Statement

To generate and study Node-Link Graph visualizations on COVID-19 data set using the python libraries like networkx.

2 Data Description

The data source is COVID-19 dataset.

- **COVID-19 Metadata**: The tabular datasets published by the World Health Organization to create new networks (similarity networks, correlation networks, etc.) and visualize network communities.
- The data is in tabular format and is contained inside multiple csv files in multiple formats.
- One of the formats is the time-series format in which the number of cases(deaths, recovered, and confirmed) are being recorded day-wise and stored in accordance with there positions in world coordinates.
- The time-series data for United states particularly is also available.
- The data has been recorded from 22nd January, 2020 to 23rd September,2020.
- Three categories of cases are recorded in three separate files viz.
  - time_series_covid_19_deaths.csv,
  - time_series_covid_19_recovered.csv, and
  - time_series_covid_19_confirmed.csv.

3 Data Reading and Cleaning

- Since the dates of data are represented by columns and rows represent the States and their respective countries worldwide, so the data has been transposed in order to labels of nodes in nodelink diagram as name of countries.
- Due to presence of a large amount of States the node-link diagram might look cluttered. Also, for some of the observations the information of the particular state is absent, leaving a NaN value in the data frame, this is why the countries are considered to be a better choice in order to represent the labels on the nodes.

- As the the rows of the dataset are unique from each other due to State identity, but as we consider the countries for the basis of our labelling then there might occur repetitions as several state can have a common country. In order to avoid this repetitions, we have taken the sum of all the values belonging to a particular country on a specific date.

4 Network Data Visualisation

Network Visualisation or Network Graph is often used to visualize complex relationships between massive data. A network visualization displays undirected and directed graph structures. A network is a collection of data where the entities within that data are related through the principles of either connection or containment.

4.1 Adjacency Matrix

An adjacency matrix is a square matrix used to represent a finite graph. The elements of the matrix indicate whether pairs of vertices are adjacent or not in the graph. Here we have an unweighted and undirected graph where the adjacency matrix is a zero diagonal symmetric matrix with zeros representing no edge and ones representing the edge.

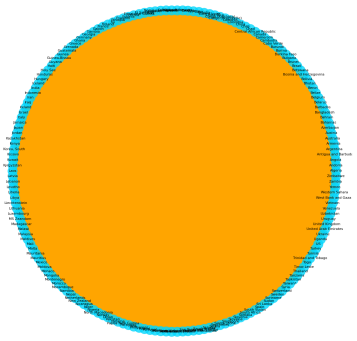
4.2 Node-Link Diagram

Graphs are frequently drawn as node-link diagrams that illuminate relationships between entities, which are nodes and links, respectively. Nodes nothing but vertices are represented as disks, boxes, or textual labels. Links nothing but edges are represented as line segments, polylines, or curves in the plane connecting vertices.

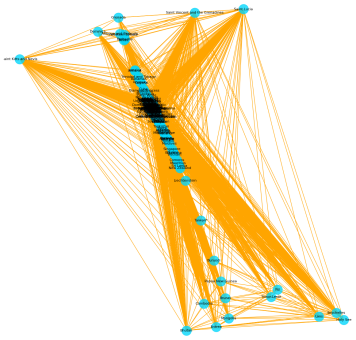
4.3 Network Visuals breakdown

4.3.1 Visualising Adjacency Matrix

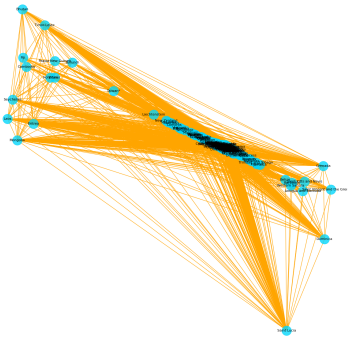
In order, to perform the Nodelink visualisation, a correlation or adjacency matrix has been calculated which finds the positive and negative correlation amongst the countries. If positive and negative correlations are not segregated, then our visualisations will not be informative enough to give correct dependencies amongst the countries and the will look like *figure.1* .



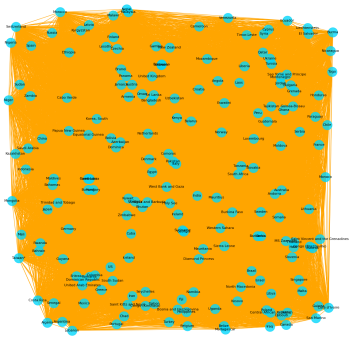
(a) Circular Network Plot



(b) Spring Network Plot



(c) Reingold Network Plot



(d) Random Network Plot

Figure 2: Figure a, b, c, d showing all the correlations amongst the countries w.r.t to death cases trends

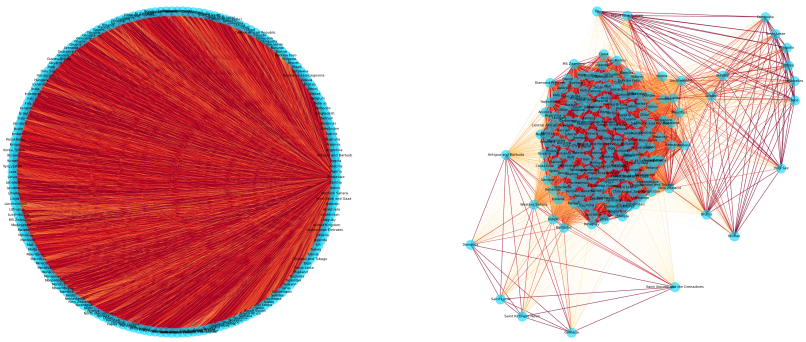
4.3.2 Correlation

It is very important to separate out both the correlation layouts. The positive and negative correlations implies:

- **Positive Correlation** connects the countries which portrays a similar trend in the change in number of cases i.e. if recovery cases have risen in comparison to previous stamp for a country then the other countries show similar raise will get connected to each other. Here, an important parameter is extent of correlation which signifies to what extent two countries are exhibiting a similar trend. For example, if two countries shows a raise of recovery rate in range of 1000 to 1200, then they are highly correlated and vice-versa.
- **Negative Correlation** connects the countries which portrays an opposite trend in change in number of cases i.e. if recovery cases have risen in comparison to previous stamp for a country then the other countries show a fall in recovery rate with similar rate will get connected to each other. Here, an important parameter is extent of correlation which signifies to what extent two countries are exhibiting an opposite trend. For example, if two countries shows a raise and fall of recovery rate in range of 1000 to 1200 respectively, then they are highly correlated and vice-versa.

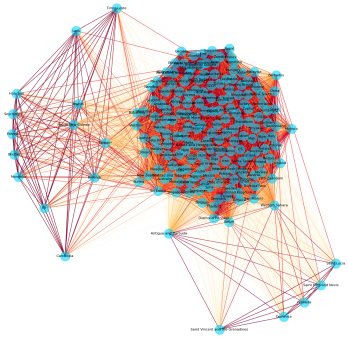
4.3.3 Setting correlation threshold

In order to generate some deducible network graphs, it is very important to give preference to highly correlated edges and hide the edges have lower correlation weights. Otherwise, the graph generated can be very cluttered due to high number of nodes and consequently high number of edges. Samples of few such networks are shown below in *figure.2 and figure.3*

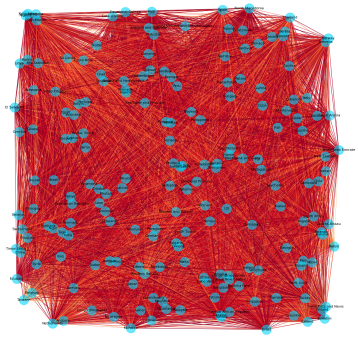


(a) Circular Network Plot

(b) Spring Network Plot

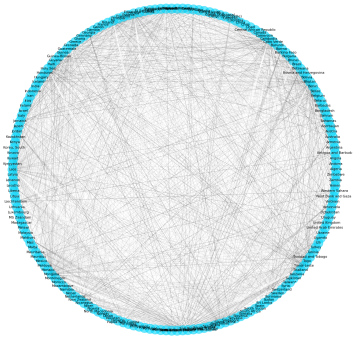


(c) Reingold Network Plot

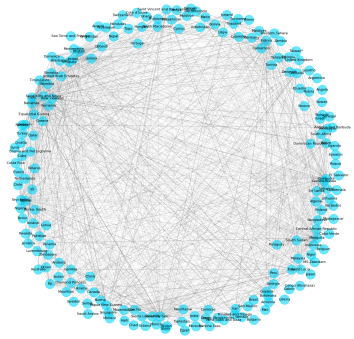


(d) Random Network Plot

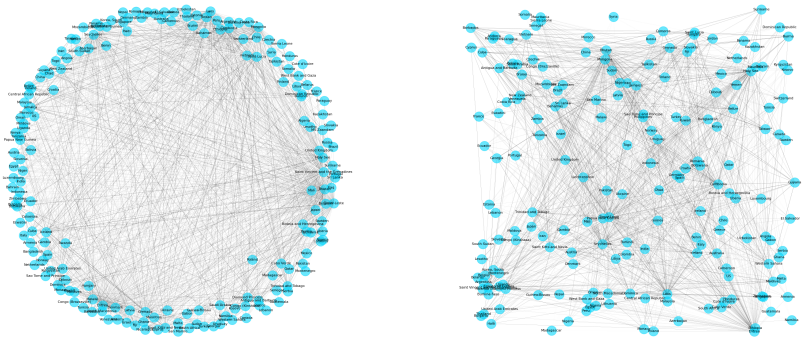
Figure 3: Figure a, b, c, d showing all the positive correlations amongst the countries w.r.t to death cases trends



(a) Circular Network Plot



(b) Spring Network Plot



(c) Reingold Network Plot (d) Random Network Plot

Figure 4: Figure a, b, c, d showing all the negtive correlations amongst the countries w.r.t to death cases trends

Setting threshold is another important aspect. As shown in figure, sometimes it is difficult to put all correlations. So, to overcome this problem, we can set a threshold value for correlation weight. All correlation weights in graph vary within range of 0 to 1. According to the density of the graph one can set the threshold values. For, ***positive correlations*** , all weights below threshold are made hidden in the graph and for, ***negative correlations***, all value above threshold are made hidden.

4.3.4 Styling the edges based on their weightss

is a step where we extract the weights of edges, so that they work as the edges' widths in the graph. Given that correlations are very small, we have modified each one using the function 1+abs(x))**2 , so that they don't look too thin in the graph. We have used the scaling colour gradients for positive and negative correlations.

4.3.5 Styling the nodes based on the number of edges linked (degree)

The final step is to style the nodes based on how many edges it is linked to (also known as the degree of the node). For this, we unpack all the nodes in node sizes and scale it to x so that they look bigger.

4.3.6 Setting the distances amongst the nodes

the defalut distance between two nodes is 1/sqrt(total number of nodes). As the number of nodes in this graph is 187 which makes extremely small distances

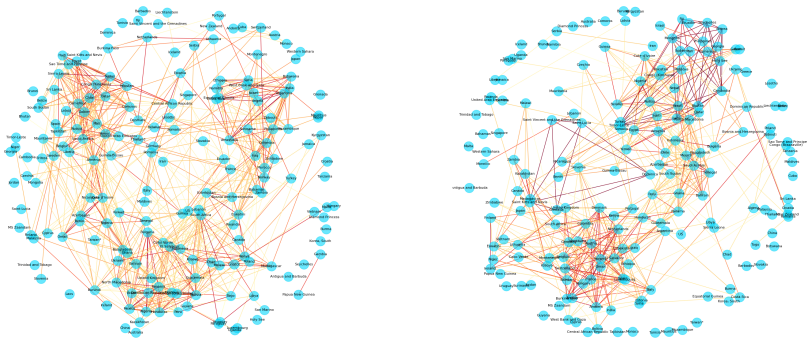
between the nodes making it difficult to view nodes distinctly. Thus, the distance has been manually set amongst the nodes by giving argument to function nx.spring_layout(k=1) where k is the distance set.

5 Nodelink layouts for COVID-19 on World data

For the the visualisation of spread of COVID-19 throughout the World three csv files from given dataset are being used in order to make an inference on confirmation, death, and recovery rates. The csv files have the tabular data representing the day wise increase in cases regionally. The columns represent the dates and the rows represent the regions(viz. States and Countries) where the cases have been spotted.

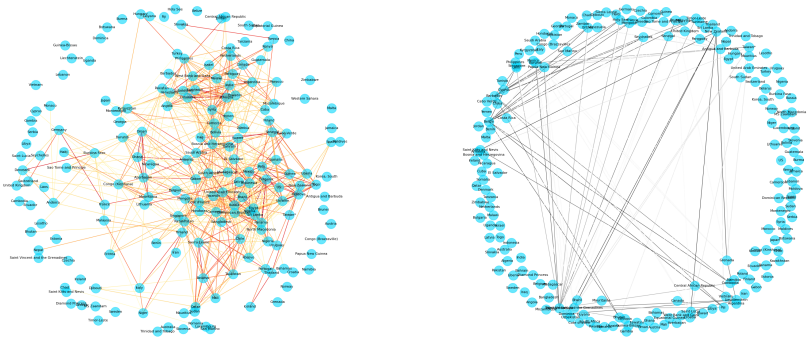
After following the above procedure, the following inferences and visualisations has been obtained.

- **Confirmed** Clusters of countries with similar trend: [Armenia, Haiti, Egypt, Sierra Le one, Sudan, Saudi Arabia] , [Mexico, Panama, Algeria, UK, Bulgaria, Guatemala, Indonesia,North Madonna, Kosovo],[India, Syria, Argentina, Venezuela, Ethiopia, Zimbabwe, Botswana]
- **Deaths** Clusters of countries with similar trend: [Denmark, Columbia, Switzerland, Andorra, India, Oman, Norway, Netherlands, Austria, Italy, UK, Ireland, Finland, Madagascar] , [Yemen, Chile, Haiti, Brazil, Cambodia, Fiji, Enteria, Mongolia, Ukraine, Bhutan, Qatar]



(a) Confirmed Cases

(b) Death Cases



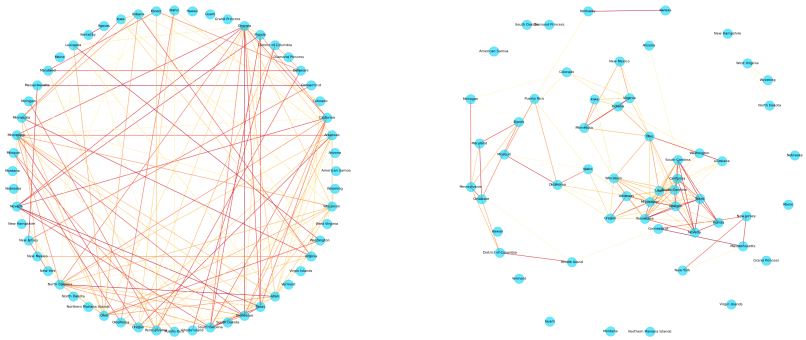
(c) Recovered cases (d) Death cases

Figure 5: Figure a, b, c showing positive correlation amongst the countries for confirmed, death, and recovered cases respectively. Fig. d shows the negative correlation amongst countries for death cases. All these plots are on Reingold layout.

- **Recovery** Clusters of countries with similar trend: [Togo, Chile, Sudan, Liberia, US, Ukraine, Kuwait, Congo, Russia, Poland], [Costa Rica, India, Ethiopia, Syria, Philippines]
- **Deaths** Clusters of countries with similar trend:[Barbados, Mongolia, Liechtenstein]
- No clusters found for recovery and confirmed cases for negative correlations.

6 Nodelink layouts for COVID-19 on United States data

For the the visualisation of spread of COVID-19 throughout the US two csv files from given dataset are being used in order to make an inference on confirmation, death rates. The csv files have the tabular data representing the day wise increase in cases regionally. The columns represent the dates and the rows represent the regions(viz. States and Cities) where the cases have been spotted.



(a) Circular Layout

(b) Spring Layout

(c) Reingold Layout

(d) Random Layout

Figure 6: Figure a, b, c showing positive correlation amongst the US states w.r.t. trends in confirmed cases for COVID-19 .

7 Best Visualisation

The best suited for the world COVID-19 data set is the Fruchterman Reingold layout and Spring layout as they depicts the clusters properly which other graphs fails to potray. However, the US dataset being lesser dense can be analysed through other methods as well.

8 Technologies Used

- networkx library python
- matplotlib python