Understanding World Population Correlation through Network Analysis

I. Background

Population growth of a country depends on number of factors like economic growth, health care, government policy, life expectancy, fertility rate, education, immigration, socio-economic factors, etc.
We aim to find the 4 countries that are highly correlated to a target country. This is achieved by first finding the top 4 highly correlated countries and then using linear regression to predict the future population. The second method to find the 4 high impacting countries on the target country's population is by using Lasso regression. The benefit of lasso regression is that by increasing the multiplying factor, we can reduce the dimension of the predicting independent variables to 4. The coefficients of the variables obtained through lasso regression are used as edges in the following network analysis.

II. Dataset

The dataset consist of yearly population of 212 countries from 1960 to 1999.

III. Network Analysis

A. Introduction

The first step in Gephi Network Analysis is to import two spreadsheet files of nodes and edges respectively. In our case, we have a matrix representing the coefficients as the relation of the target country with the other 4 countries. These coefficients are used as edges and their magnitude acts like weights. The name of the country is the node in our analysis.

Figure 1 is a graph without implementing any network algorithm. As we can observe there is no discernable pattern.

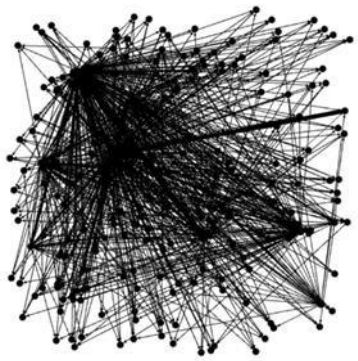

Figure 1. Gephi Network Graph without implementing any algorithm

B. Analysis and Insights

The Fruchterman Reingold Layout algorithm is a force directed graph algorithm. Their purpose is to position the nodes of a graph in two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible, by assigning forces among the set of edges and the set of nodes, based on their relative positions, and then using these forces either to simulate the motion of the edges and nodes or to minimize their energy.[1]

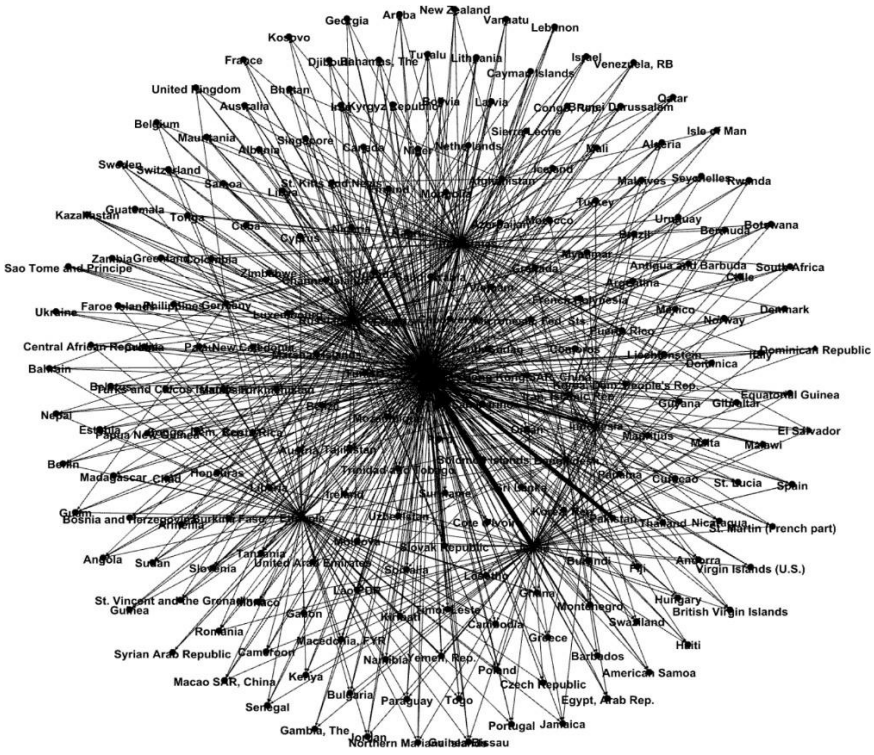

Figure 2. Network Graph Using Fruchterman Reingold Layout.

Figure 2 shows how all the countries are related to each other as per the Fruchterman Reingold Layout Algorithm

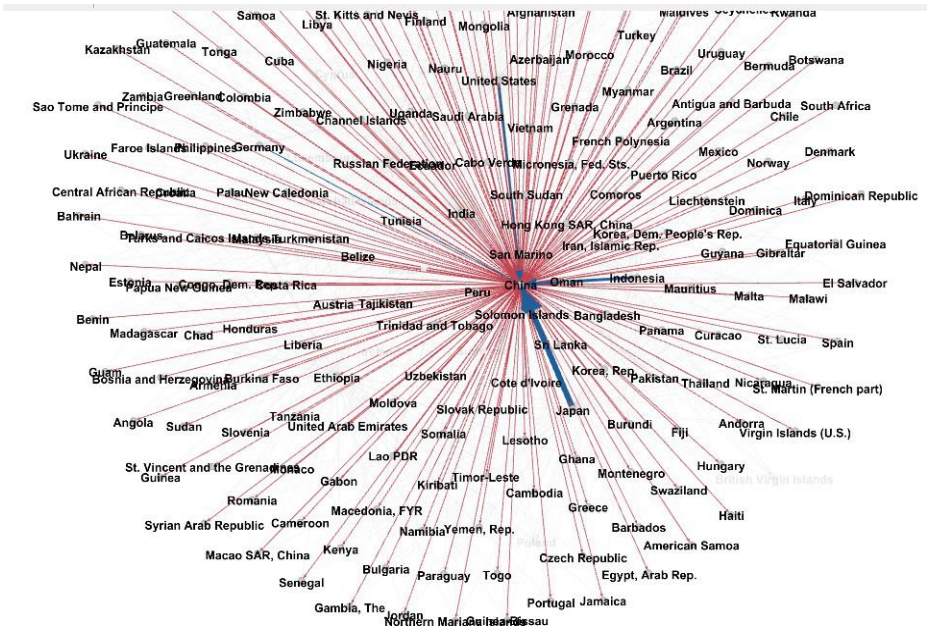

Figure 3 Correlated Countries for China

In figure 3, blue arrow represents the countries that are correlated to China. As per the algorithm, thicker the arrow, higher is the correlation. Therefore, Japan is the most correlated country to China followed by Indonesia, United States and Germany. From this insight we can develop a hypothesis that countries with closely related economic status (GDP, etc) are correlated (China, Japan, Germany and United States have similar economic status).

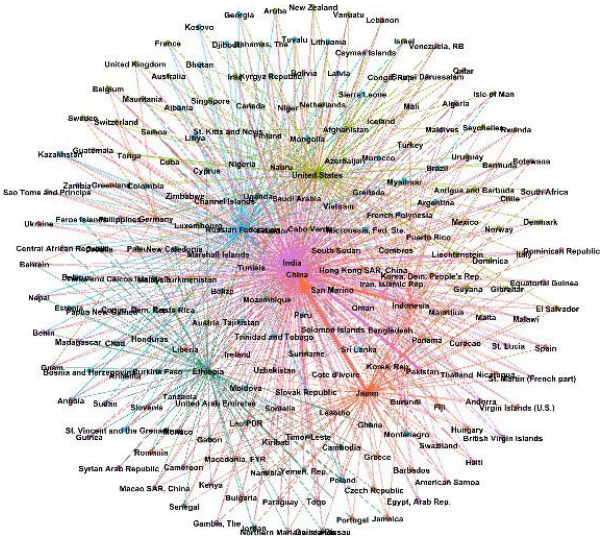

Figure 4 Communities in World

Figure 4 shows various communities/clusters that were obtained after applying modularity to the network. The below table shows the type and percentage of communities.

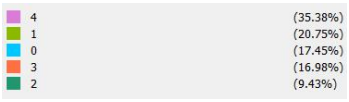

Figure 5 Community and percentage

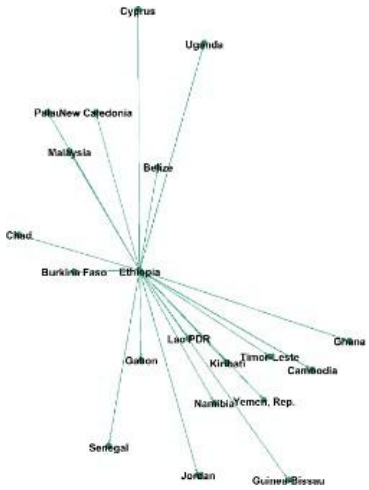

Figure 6 Cluster 2

The clusters could be used to develop hypothesis if there is any relationship between various countries and their population growth. Few of the countries in the above clusters have geographic proximity.

IV. Conclusion

Network analysis using Gephi is helpful to explore the structural properties of correlated variables. We can safely generalize that relationship with population growth of some of the countries are affected by geographical proximity, socio-economic level, standard of living etc. Thus, Network analysis can be employed to understand complicated problems and to find inherent patterns among representative nodes.

V. References

[1] Force-directed graph drawing (https://en.wikipedia.org/wiki/Force-directed_graph_drawing)

[2] https://gephi.org/tutorials/gephi-tutorial-layouts.pdf