Challenge 3:
Insights from World Population Visualization

Kishan Shah, Shirish Pandagare, Siddharth Ajit

***Abstract*- This report explains the relationship among various countries through world population data using network analysis in Gephi.**

I. BACKGROUND

Population growth of a country, in general depends on number of factors like economic growth, health care standards, government policy, life expectancy, fertility rate, education system, net-migration, social and culture factors etc. These common factors responsible for the population growth provides a little intuition that the population of one country may depend on the other. To draw some conclusion regarding the dependency of the population of one country on another, we are performing the Machine learning technique on the World Population data collected by the World bank. The data consist of yearly population of 258 countries for last 75 years. It is our purpose here to dig up some substantial information about the population of the country.

Our aim for this study is to deduce 5 important countries which plays a vital role in predicting the population of a given country. To serve our purpose, for finding the 5 essential predictors (Country), we are using the linear regression technique. Because of the inter-association of the variables (country), the data is highly prone to multicollinearity. Hence to circumvent this problem, LASSO regression technique was used to obtain the best five coefficients corresponding to the significant countries. Further, we have used Gephi for the interpretation and visualization of the results. The inference from the graphs are discussed in the next section.

II. NETWORK ANALYSIS

A. Introduction

The first step in Gephi Network Analysis is to import two spreadsheet files of nodes and edges respectively. Node file includes Id and Labels and Edge file includes Source, Target and Weight. In our case, 0 to 257 are Ids, and name of the countries are labels in a node file. Source in edge file includes corresponding Ids of 5 or less countries which are important to predict the population of one target country and weight indicates the coefficient of source countries to explain target country.

Figure 1 is the network graph that we get after importing the node and edge file. This graph is the basic initial network generated without performing any analysis.

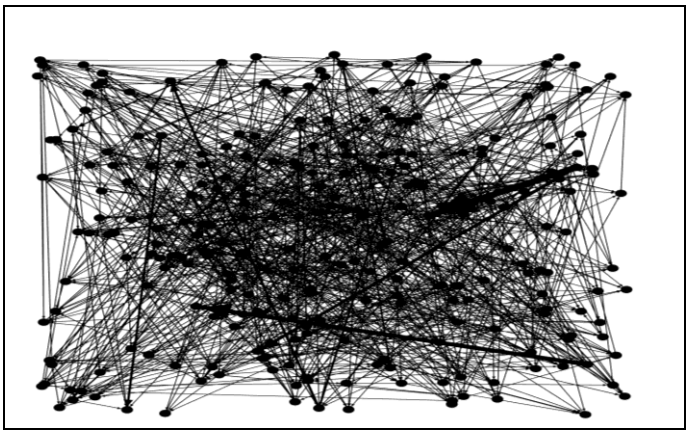

Figure 1. Gephi Netwrok Graph without Analysis

B. Analysis and Insights

The Fruchterman Reingold Layout algorithm is a force directed graph. Forces are allocated to the edges and nodes, these forces are further used to represent the graph such that the overall energy of the network is minimized[1].

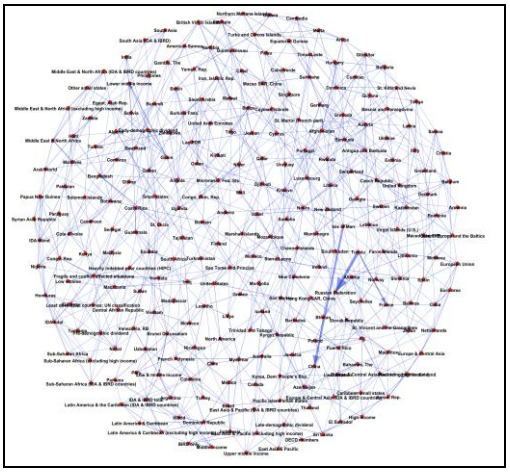

Figure 2. Netwrok Graph Using Fruchterman Reingold Layout.

From figure 2 represents how one country is related with other countries in terms of population parameters. Network graph in figure 2 is generated using Fruchterman Reingold algorithm.

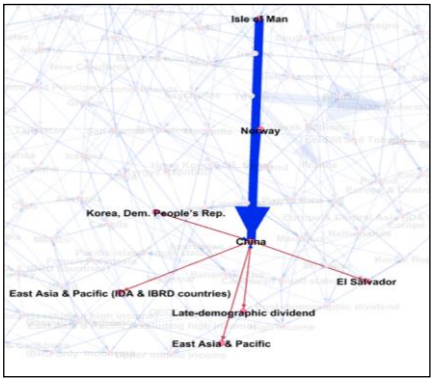

Figure 3(a) Important Countries Representing China

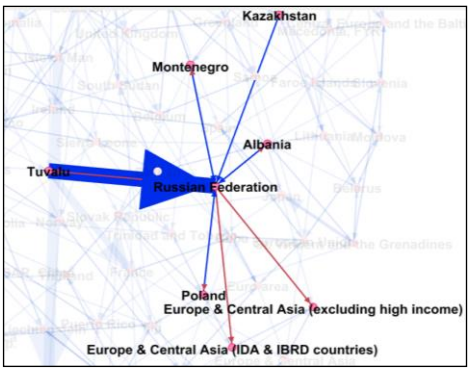

Figure 3(b) Important Countries Representing Russian Federation

After carefully observing figure 3(a), we can say that Isle of Man is very significant country to represent China as arrow has higher thickness as compare to other countries. Higher thickness represents Isle of Man has higher coefficient as compare to other countries which represent China. Same way in figure 3(b), Tuvalu is very significant to represent Russian Federation as it has higher thickness (means Tuvalu has a higher regression coefficient) as compared to other countries which represent Russian Federation.

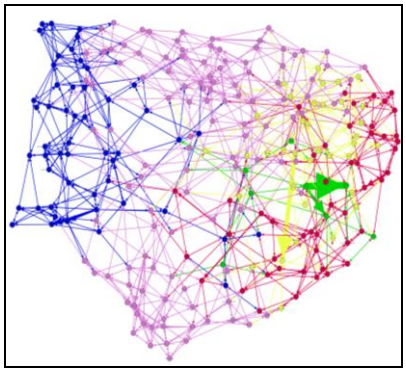

Figure 4(a) Yifan Hu Layout with Modularity

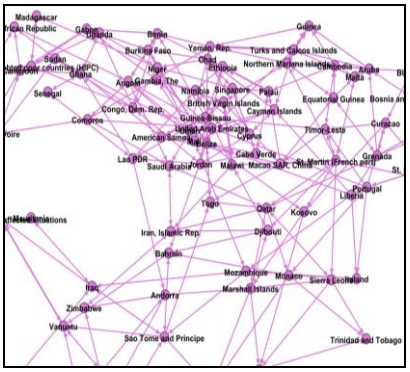

Figure 4(b) Magnified View of Modularity=2

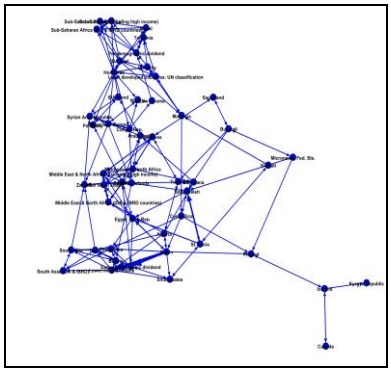

Figure 4(c) Magnified View of Modularity=4

The network representation above is clustered according to the modularity ranging from 0 to 4. Modularity is used to detect sub structures within networks. Higher modularity indicates stronger connection within the sub structure and weaker connection with the rest of the network[2]. In the context of the population data, higher modularity shows stronger interrelationship within the cluster.

In figure 4(b), the dense portion of the cluster majorly represents countries from north and central Africa, some of them even share their borders. The demographics, healthcare and economic conditions of the countries are similar in nature. This suggests the hypothesis that the population growth of a specific country can be predicted using the projected growth expectation of other equivalent countries.

Furthermore, the dense clusters of figure 4(c) (Modularity = 4) also represents African countries, aggregates of African countries and some countries from the middle east. Larger value of modularity indicates that the interdependency among countries in figure 4(c) is high. Again, these countries seem to have economic growth and geographical proximity in common. However, this type of analysis cannot be extended to all the countries in figure 4(c). Sparsely located nodes (bottom right) indicates the presence of Iceland, Canada, Kyrgyz Republic, these countries are significantly different from the former countries discussed above and are positioned farther from the dense cluster.

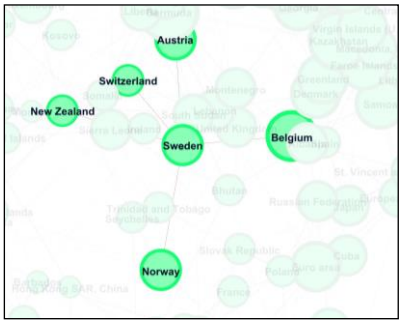

Figure 5 Relationship Between Sweden & Other Countries

In some cases, we cannot get specific details to explain the relationship of one country with other predictor countries. There are some cases, in which we can surely say that the country which is highly related with other countries in terms of geographical proximity, economic level, health care, cultural influences. Considering a specific example of Sweden in figure 5, we can say that the countries which represent Sweden has geographically proximity with Sweden except New Zealand.

III. Conclusion

Network analysis using Gephi is very useful to explore the structural properties of best predictors. We can say that relationship among population of some of the countries are influenced by geographical proximity, economic level among many other factors. In the larger context, network analysis can be extended to complicated real problems to find inherent relationships and patterns among representative nodes.

IV. References

[1] Kobourov, Stephen G. *Spring Embedders and Force-Directed Graph Drawing Algorithms* (2012), , *arXiv:1201.3011*, *Bibcode:2012arXiv1201.3011K*

[2] En.wikipedia.org.*Modularity (networks)*. [online] Available at: https://en.wikipedia.org/wiki/Modularity_(networks) [Accessed 2 Oct. 2018].

.