World Population Visualization

Sayeed Alvi, Anirudh Shaktawat, Amrita Sundari V

October 2, 2018

Abstract

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

1 Introduction

The aim is to draw some insights about the population dependency of a country by visualizing 16 years of predicted population data of each country using a L1 regularized regression model trained on the past 40 years population data. In this challenge we had a limitation of finding atmost 5 predictors for each country only.

Thus 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 multi-collinearity. Hence to circumvent this problem, L1 regularization 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.

2 Network Analysis Using Gephi

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 this case, 0 to 257 are Ids, and names of the countries are specified as labels in the node.csv. 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.

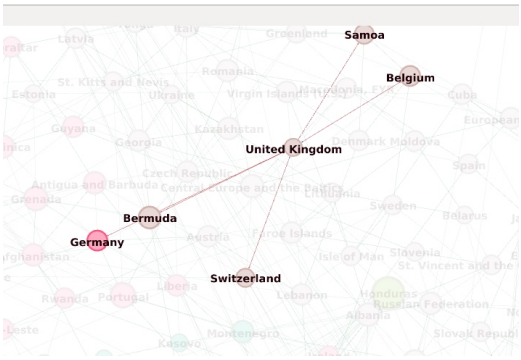

Figure 1: Estimation of United Kingdom population by neighboring countries

In Fig 1, We can see the edges that are coming in and going out from the node United Kingdom, highlighted by red and blue color respectively. It is also important to note that there will only be five edges that goes in for each country, while the edges that are coming out can be zero or more.

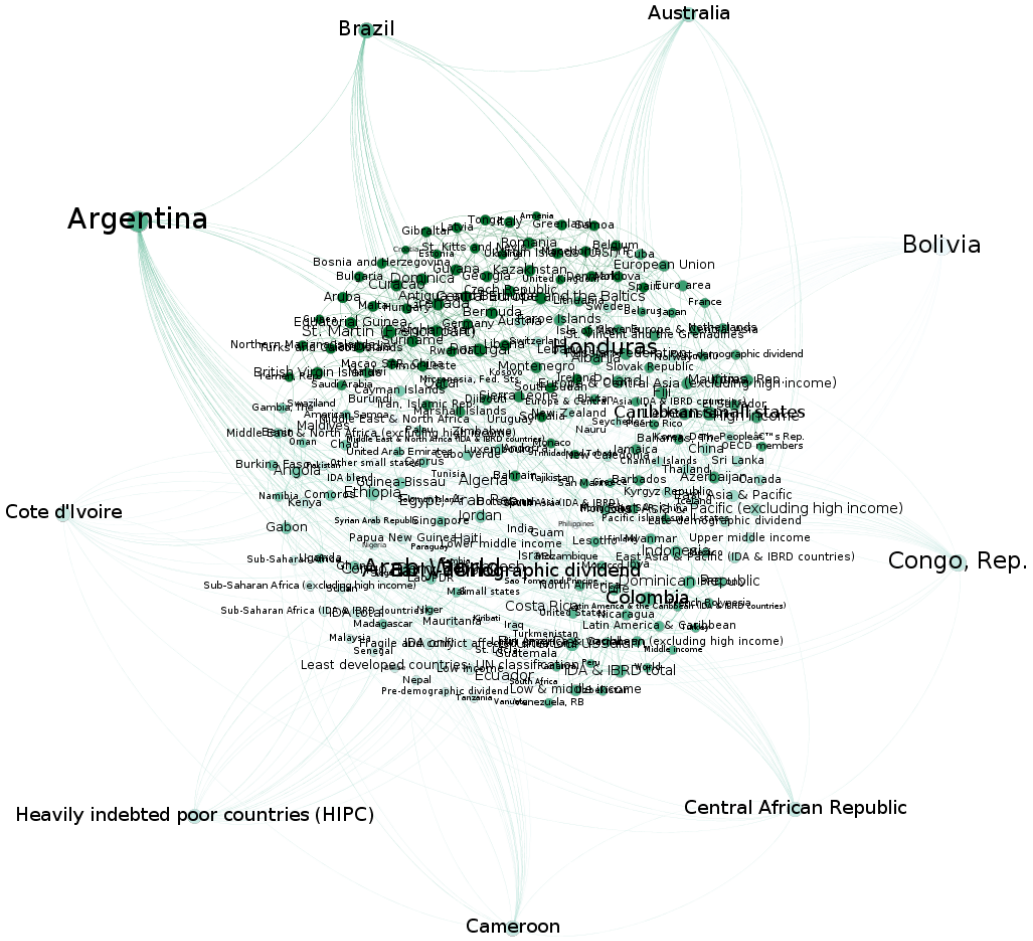

Figure 2: Network graph to find about the countries with most total number of edges.

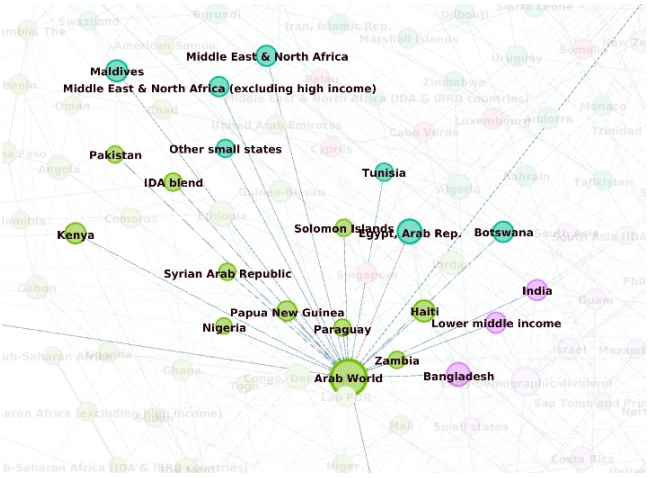

Figure 3: Arab World and its association with other countries

The Network Graph in Fig 2, was made using the Fruchterman ReinGold Layout, which

makes the nodes equally spaced in the three dimensional space resulting in a sphere like figure as shown. We have used a filter, to modify the topology of the graph and made the countries with 20 and more edges to project out from the rest of the countries. We found out that the 8 countries had more than 20 edges, which effectively means that most of the other countries population number were dependent on these 8 countries' population.

In Fig 3, We can see that the Arab World is connected to more than 20 countries. From this we can infer that, in addition to the other 8 countries, Arab World also plays a critical country on which a significant number of countries depend. This was not directly evident from the network graph, but we can independently find it by looking in the main cluster of Fig 2

In Fig 4, We have tried to cluster the countries by finding the modularity for the network, and applying them to the nodes and edges. Higher modularity indicates stronger connection within a cluster and weaker connection with the rest of the network.

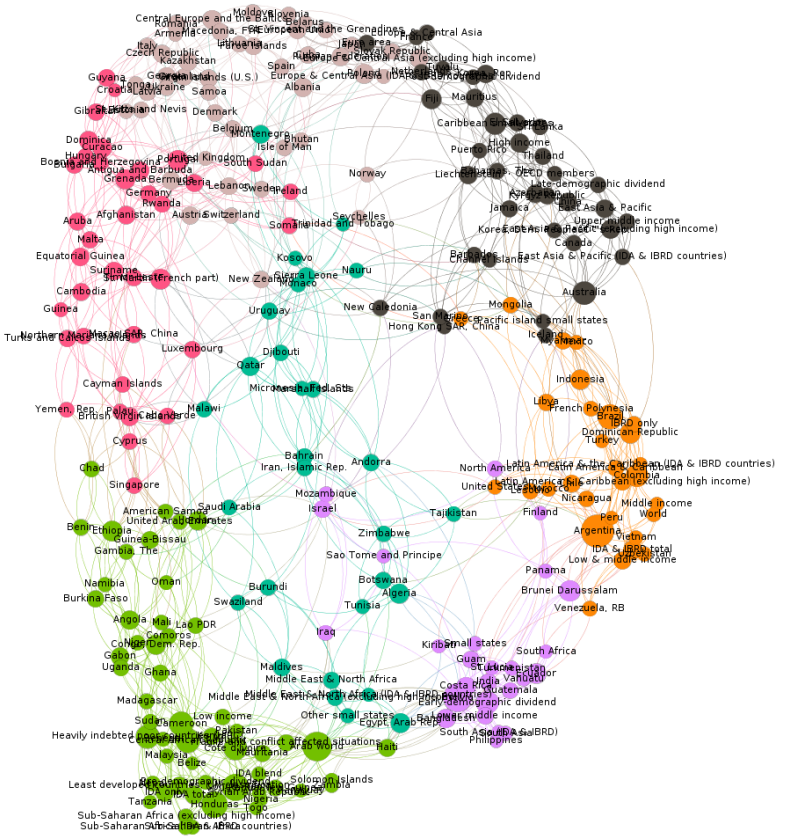

Figure 4: Network Graph Using Force Atlas 2 Layout (Modularity = 7)

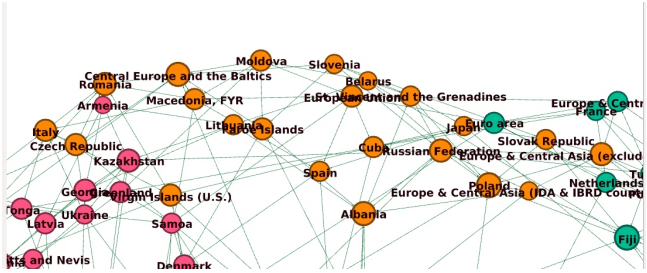

Figure 5: Cluster representing Spain, Russia, Poland and other European countries

As we can see, the countries were grouped into 7 clusters. Interestingly, the countries in the

clusters are mostly representing the neighboring countries. For example, we can see in Fig 5, the orange colored cluster represents most of the European countries.

3 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 in addition to 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.

References

[1] Introduction To Machine Learning with Python: A Guide for Data Scientists, Andrea C. Muller and Sarah Guido.

[2] Gephi Tutorial Quick Sort - version0.7alpha2

[3] Gephi Tutorial Layout and Visualization