WESTERN SYDNEY
UNIVERSITY

W

Visualisation – 19884657

Assignment -1

Relational Data Visualisation

Rabina Padhy|Master's in Data Science|27-04-2019

Table of Contents

Introduction

Visuals are especially important when you are trying to find out the relationship among hundreds and thousands of variables to discover the insight data. Visualisation can also be defined as "A picture is worth a thousand words". Now every second, trillions of data get generated from multiple industries. When user try to invest in one of the industries, visualisation plays vital role by providing the hidden pattern with sophisticated analysis of data. Visualisation helps to take better decisions, take smarter actions and operate efficiently for top executive directors and departmental decision makers in an organisation (Ward and Grinstein, 2010). Therefore, many people prefer data visualisation to do deep dive into the data and use the insight for faster decision making in an organisation.

To provide meaningful information from a data set, we need to consider their data size, data type and attribute composition. But if you are working in a massive data set like big data, then how to consider important columns among many variables. If you select appropriate attributes, then it helps to drive better visuals and to take better actions.

Benefits of Data visualisation in a real world. (Boost Labs, 2020)

- ❖ Data visualisation is the key to unlock Big Data.
- ❖ Visualization allows for near instant absorption of high amounts of data.
- ❖ It speeds up the decision-making process.
- ❖ It helps to reveals pattern and trends.
- ❖ It quickly shows the relationship between operator and results.
- ❖ Its more interactive then the traditional methods.
- ❖ It enhances the communication.
- ❖ Data visualisation tools show insights that may be missed in traditional reports.

This paper will discuss the technical details, their advantages and disadvantages of Force-directed graph, Degree of Interest Tree Map and Network graph. Gephi tool and DOI Tree Map tool are helps to analysis the dataset and provide better visibility.

Force Directed Graph

Force directed graph method is one of the traditional algorithms widely used for data analytics purpose. The Force directed graph is treated as physical system by applying relevant forces such as gravity force, spring force etc. on each vertex and edges in a graph. In the year 1963, the Force directed graph drawing was done by Tutte, based on method barycentric representations. As per Hooke's Law, there can be two forces that exists between adjacent nodes i.e. repulsive force and attractive force. The Force directed graph is further improvised by applying multiple algorithms which includes FR (Fruchterman and Reingold) algorithm, FA (ForceAtlas) algorithms and Linlog energy model (Hua, Huang and Wang, 2007). In FR algorithm, the force between every two vertices continues changing between nodes, until the energy is minimized. FA algorithms provides additional decipherable shapes to network and options to the optimise the layout. The Force directed graph is further enhanced by applying clustering technique created by Linlog energy model.

Advantages and Disadvantages

The main benefits of Force directed graph is its capability to generate multiple layouts for the same dataset, by applying combinations of input parameters. This makes it quite flexible enough in the class of graph drawing techniques. This drawing technique is very simple and can be done in couple of lines of code. The system is concentrated on minimum cross edges, angle deviations, edge deviations and angular resolution. However, this

graph technique provides some disadvantages as it becomes unstable for large data sets. Firstly, the issue of weak local minima turns out to be increasingly essential as the number of vertices of the graph increments. Secondly, it has running time comparable to O ($|V|^2$), where v is the number of vertices in the graph because it must calculate the repulsive force for all the vertices (Kit,2016).

Degree of Interest Tree Map

Maximum Tree maps achieves space utilization of a single geometrical area like rectangle, hexagon, polygon etc. By limiting the data visualization to rectangle or polygon, it limits human capability on graph recognition. To overcome from this, we come up with flexible connected approach which is a combination of Degree of interest method and the tree map paradigm. DOI tree map emphasizes large hierarchical structure within limited area with better interpretability (Tu and Shen, 2007). It decomposes a problem into two or more sub-problems of the same or related type until the sub-problem becomes simple enough to be solved directly. It helps to make the data more readable and more structured. This method is very simple and easily understand by non-expert of visualisation users.

Advantages and Disadvantages

DOI Tree map advantage is that, it can be easily scalable without needing much changes in the structure. It provides multiple foci views with context interaction. It uses proper usage of the display area so that can easily view all the open nodes. Data overlapping is less in the DOI tree map because if user open multiple trees at one time then it adjusts the display view. However, it includes some disadvantages as it's hard to visualise the total structure of a graph. If the dataset is not hierarchical structure, then it's difficult to visualise the data by DOI tree map (Nguyen, Simoff and Huang, 2014).

Network Graph

Networks are models of relationships, showing who is connected to what, and each connection may show a weight associated to the relationship. Graph is combination of set of vertices and set of edges showing how the vertices are connected. In the network visualisation we can have several elements to show variety graphs like colours, shapes, position and size. Better Network graph will be considered as minimum overlapping of nodes and less edge crossing.

Advantages and Disadvantages

The main advantages of Network graph are it can show nodes for large data set. However, it contains more overlapping and cross edges. Task performance can be solved by Network graph. Since Network graphs connects all the nodes for large data set, it is expensive than other techniques. It consumes more time than other methods. When the network graph generates multiple memberships for single node then there might be an occurrence of misinterpretation and becomes hard to conclude.

Data Set using Multiple Technique

Insight about Olympic Games Data Set

The modern Olympic games are playing vital role in global sporting events, this attracts worldwide interest from people of all races, ages and social positions, transforming them in absolute global and democratic event. Since decades, the modern Olympic Games were part of world culture and modern history, reflecting political tensions, economic crisis, cultural flowering, and technological development. To know more about Olympic games won by countries and their medal counts let's analyse using 15317 records from year 1976 to 2008 (Olympic Games Data Analysis, 2020). The datasets comprise of attributes such as Country, City, Year, Medal and Sports. Before proceeding further, we would perform data cleansing.

Force Directed Technique

To visualise Olympic Data set in a Force directed graph format, I have chosen Gephi Tool (Quick Start, 2017). This tool is very interactive with user and have better representation of dataset by using their advance visualisation. Before applying the Dataset, we have split the data set into two different files, one contains Nodes and another one contains Edges. In the Node sheet, we have ID and Labels as columns and in Edge Table, we have ID, Target, Source and Weight as columns. The Labels represents various countries, Target represents Medal names, Source is a reference to the Countries and Weight represents number of medals won by the country.

Technical Details

- ❖ Upload this Edge dataset into edge spreadsheet and Node dataset into Node spreadsheet in the Gephi tool.
- ❖ We can visualise the data in the graph window after uploading. Refer Figure 1.
- ❖ To improvise, we run Network Diameter, Modularity, Avg. Clustering Coefficient and Avg. Path Length from Statistics window.
- ❖ To change the colour code of Node, Go to Appearance window then Nodes then Partition. Select the Modularity Class and change the colour code as per modularity wise.
- ❖ Select the Layout as Force Atlas and change the repulsive strength from 100 to 2000. Check the Adjust by Size option. Change the gravity to 30.
- ❖ Select the top 10 countries from the data laboratory and change their colour code, size and appearance. Change the Edge colour code as per Medals and thickness as per medal count. Refer the Figure 2.
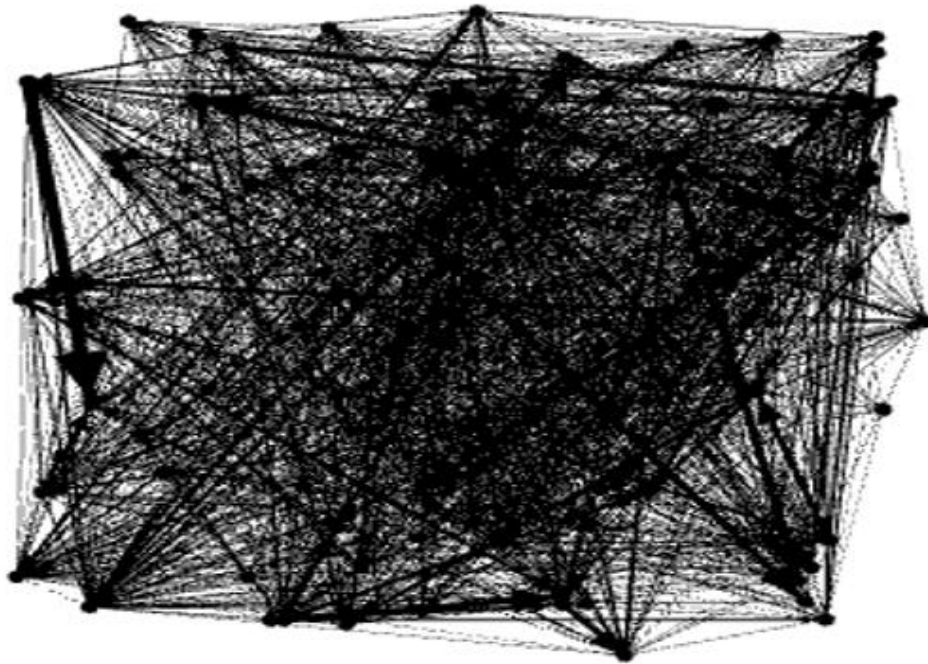
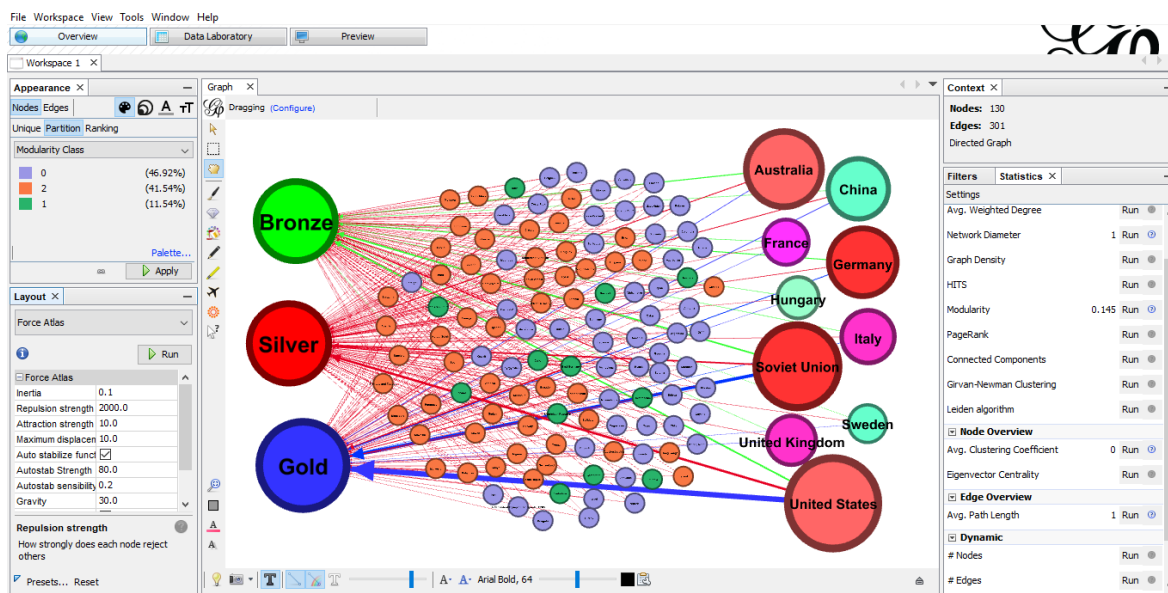Figure 1 No Layout applied on Olympic Dataset

Figure 2 Force-directed Graph on Olympic Dataset

After the transformation, we can clearly see the data in different shapes and colours which easily concludes that, United States has more Medals in the Olympic games and maximum medals are Gold (represented by the thickest edge and a bigger circumference for the country). Second position is awarded to Soviet Union with more Gold Medals. Followed by Australia, Germany and China with slighter difference in the medal count and their circumference. Sweden stands at the last of the top 10 list. All other counties are in a smaller size and different colour to distinguish from top 10 counties.

DOI Tree Map Technique

Let's visualize the Olympic game datasets in a hierarchical structure, from year 1976 to year 2008. I have re-structured the data into simple text format so that we can visualise the hierarchical structure. We have grouped the 15317 records by medals against countries, to reduce it to 302 records. Olympic games were conducted in different cities every 4 years. Following is the visualization in the DOI Tree map.

Steps Performed

- ❖ Convert the data set from csv file to text format so that we can generate the xml file.
- ❖ Change the colour of the label by using Colour Property.
- ❖ By using Draw Property, we can set to Tree Visual Cue to show the data in different format.
- ❖ In the Layout property, selected the Collapsable Stack Layout for better visualisation. Please refer Figure 3.
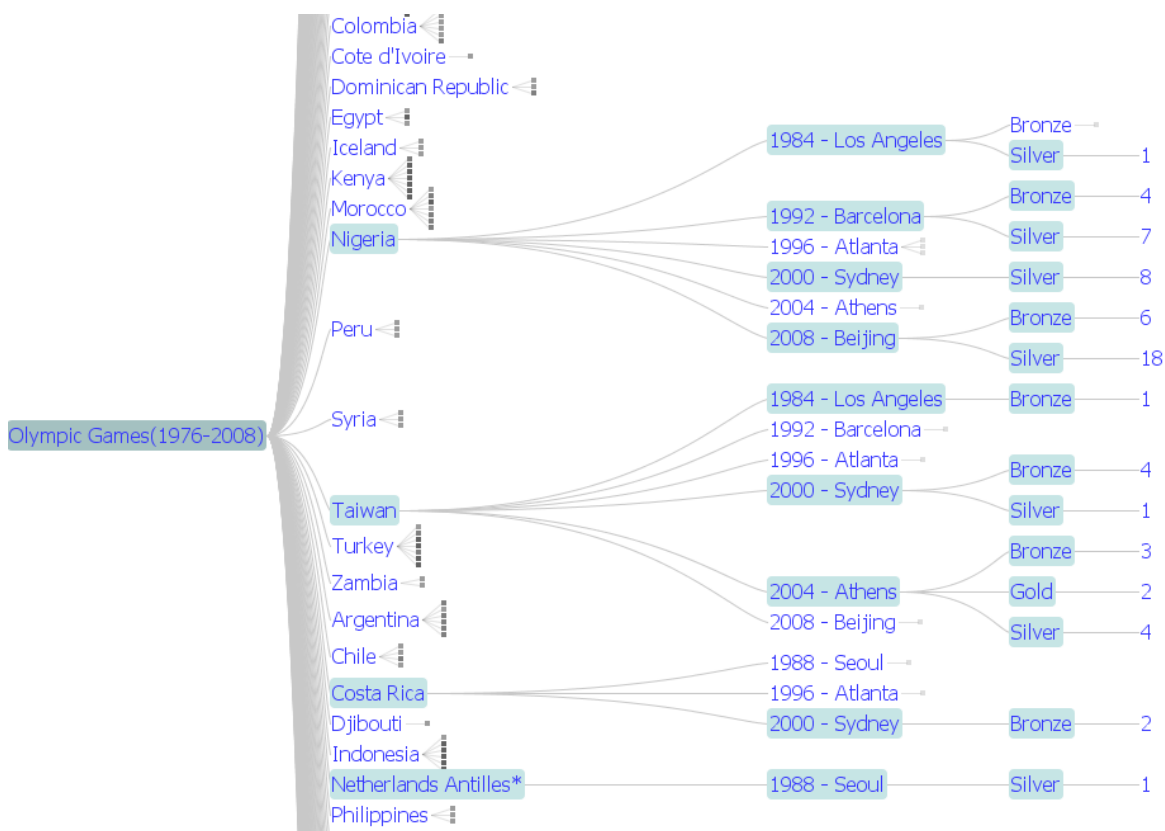

Figure 3 DOI Tree Map Graph on Olympic Dataset

Performing above steps, we were able to generate the DOI tree map. But due to the space constraints, the map did not confine into a single page. To resolve the space issue, the datasets were reduced into top 10 countries won Olympic medals during 1976 to 2008. In figure: 2, we have data set for top 10 counties and their medal count in year wise format.

Technical Details

❖ Convert top 10 counties data set from csv to text format.
❖ Change the Draw property to Simple Visual Cue.
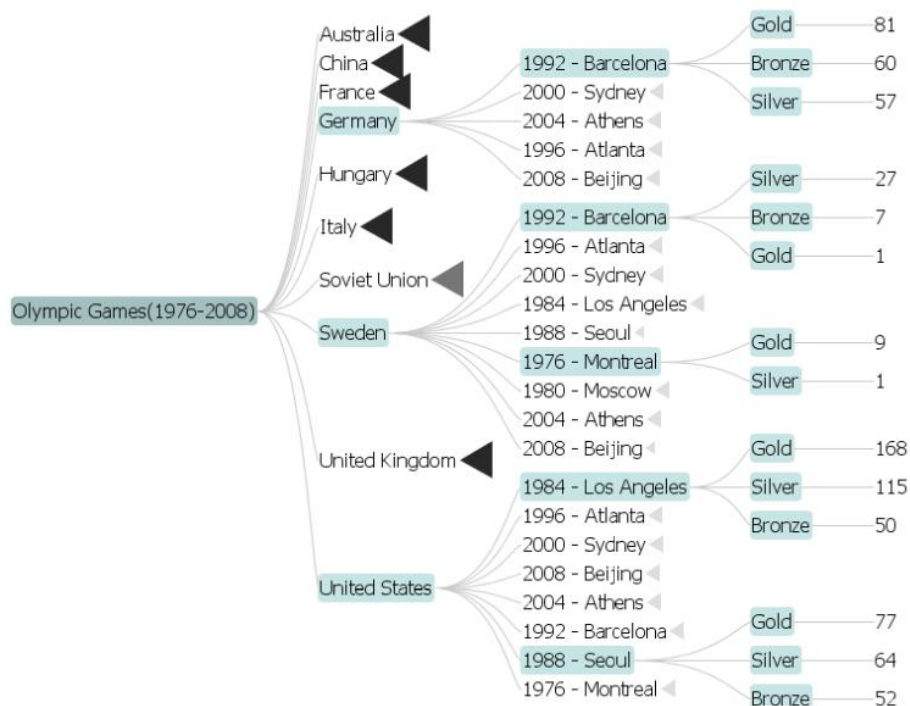❖ Change the Colour Property to Label colour change to light blue colour. Refer Figure:4

Figure 4 DOI Tree Map Graph for Top 10 Country in the Dataset

After comparing Figure:3 and 4, We conclude that, DOI Tree Map visualisation technique is not good for large data set. It's provided clear structure with better understanding of data for smaller data set. By using DOI Tree map, we can visualise that data in a hierarchical structure.

Network Graph Technique

To squeeze the datasets into single window, we adopt network graph technique. The network graph is built using Gephi tool. After uploading the raw data into Gephi Tool we get graph as the Figure 1.

Technical Details

❖ After running multiple statistics and change the Layout to Force Atlas with repulsive force of 100000. Refer left side of the Figure 5.
❖ Further improvement, we applied Layout of OpenOrd with expansion 50% and Liquid 50%. Refer right side of the Figure 5.
❖ Enhance further, Change the Layout to Fruchterman Reingold with area of 200000. Refer left side of the Figure 6.

❖ At last to highlight the top countries, I choose Layout of Noverlap with speed 3 and select specific nodes and try to change their setting so that it can easily identified. Refer right side of the Figure 6.
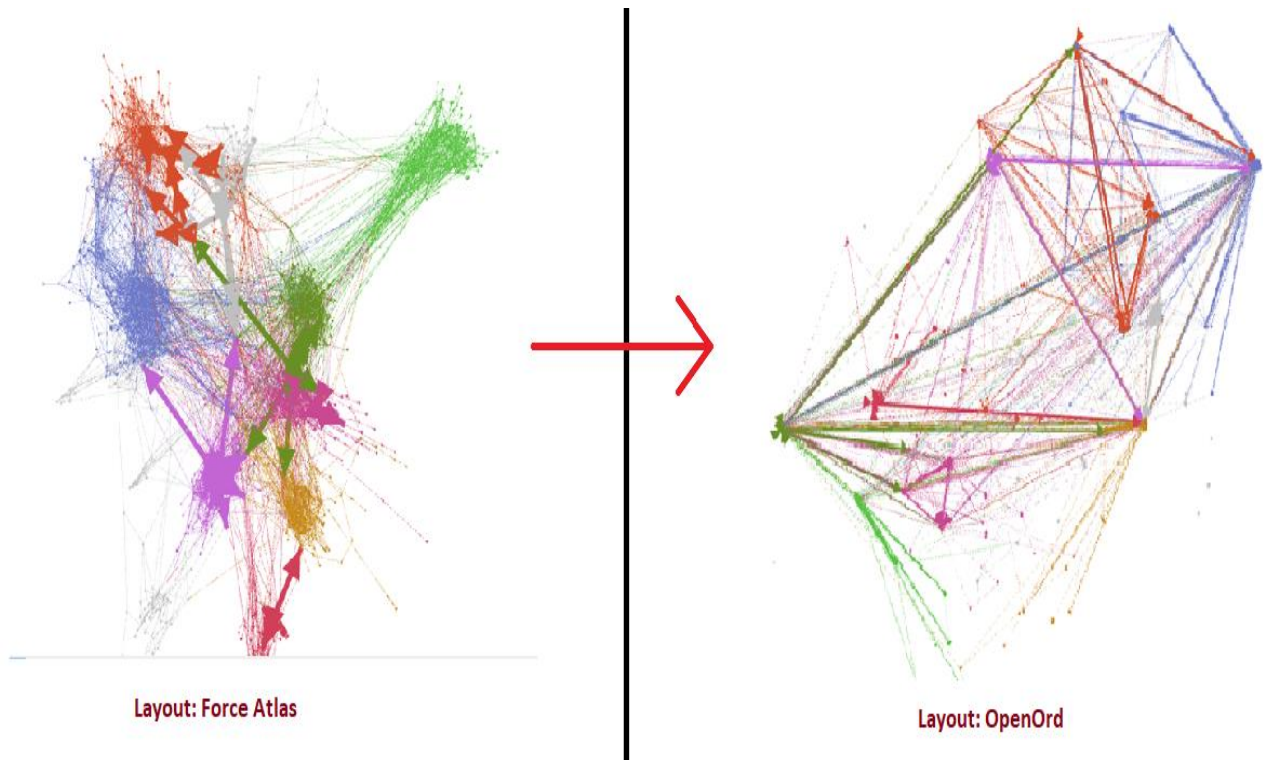

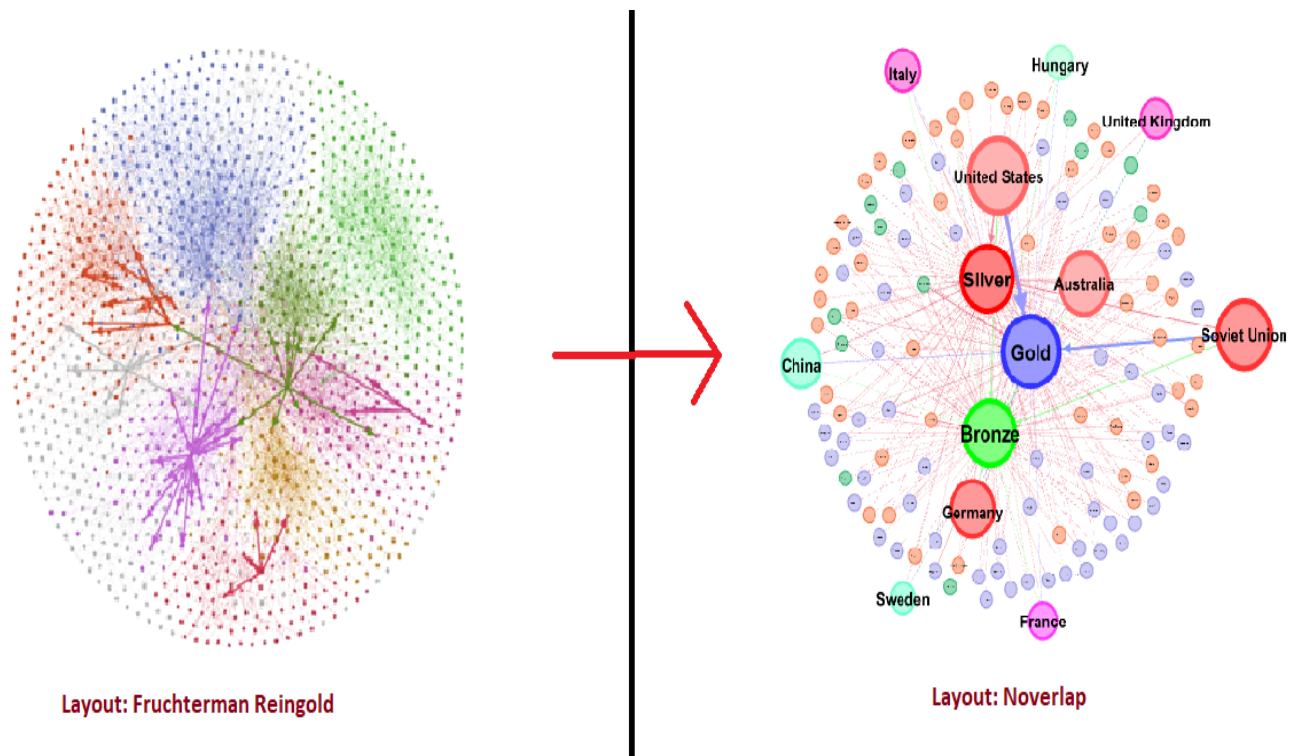Figure 5 Network Graph with Different Layout on Olympic Dataset

Figure 6 Network Graph with Different Layout on Olympic Dataset

Conclude

By analysing three techniques, Force directed graph, DOI Tree map and Network graph I can conclude saying that the Force-directed graph is very interactive and easily understood while visualising. Comparatively in Network graph the number of nodes connected to the entities could be identified effectively. However, in the DOI Tree map, we can only see a threshold of data on a single screen. In the Force directed graph, the relationship with nodes is indicated by their colours, thickness and shapes. Whereas in Network graph, its hard to see the relationship between nodes. Only by highlighting the specific portions of graph we can see the clear picture of network. However, in the Degree of Interest Tree map, I can clearly see the count of specific country won Olympic medal but not able to cover total countries due to limited space issue. Force-directed graph and Network graphs are more flexible than DOI Tree map as it covers entire dataset. Therefore, all three methods have their own merits and demerits. Force-directed graph and Network graph seems to be more transparent and provide more benefit as it has many options to change their layout structures as compare to DOI Tree Map.

Reference

1. Ward, M. and Grinstein, G., 2010. Interactive data visualization: foundations, techniques, and applications.
2. Boost Labs. 2020. 9 Amazing Benefits Of Data Visualization Every Business Needs To Know - Boost Labs. [online] Available at: <https://www.boostlabs.com/9-benefits-of-data-visualization/> [Accessed 26 April 2020].
3. Gephi.org. 2017. Quick Start. [online] Available at: <https://gephi.org/users/quick-start/> [Accessed 26 April 2020].
4. M. Ghoniem, J.-D. Fekete, and P. Castagliola, "A comparison of the readability of graphs using node-link and matrix-based representations," in Information Visualization, 2004. INFOVIS 2004. IEEE Symposium, 2005.
5. Hua, J., Huang, M. and Wang, G., 2007. Graph Layout Performance Comparisons of Force-Directed Algorithms. International Journal of Performability Engineering.
6. iti.kit.edu.2016. [online] available at: <https://i11www.iti.kit.edu/_media/teaching/winter2016/graphvis/graphvis-ws16-v6.pdf> .
7. Tu, Y. and Shen, H., 2007. Visualizing Changes of Hierarchical Data using Treemaps. IEEE Transactions on Visualization and Computer Graphics, 13(6), pp.1286-1293.
8. Q. V. Nguyen, S. Simoff and M. L. Huang, "Using Visual Cues on DOITree for Visualizing Large Hierarchical Data," 2014 18th International Conference on Information Visualisation, Paris, 2014, pp. 1-6.
9. Kaggle.com. 2020. Olympic Games Data Analysis. [online] Available at: <https://www.kaggle.com/arpitsolanki14/olympic-games-data-analysis> .