Network Visualization

Havan Patel  
havangpatel@lewisu.edu  
DATA-53000, Summer 2022

Data Visualization Lewis University

Venkata Malladi

Venkatasmalladi@lewisu.edu

DATA-53000, Summer 2022

Data Visualization

Lewis University

# Introduction

Network visualization, graph visualization or link analysis is the process of visually presenting networks of connected entities as links and nodes. Nodes are represented as data points and links, or edges are represented as the connection between the nodes. There are number of different tools out there now to generate network visualizations and getting even better with the changing technologies. For this report we are going to use Gephi to generate the visual network graphs and make analysis on three different datasets. Also, the purpose of the report is not only to use Gephi but use different algorithms and layouts to find different communities, authoritative nodes, and broadcasting nodes that can show us more related information throughout the network. Network visualization is very informative not only in the fields of data science but in business , and cyber security as well because as data grows, we need to also account for security on it as well.

# Data Description

As described in Section I above, the datasets we are using is an Exploratory Data Analysis (EDA) to get insights from the network visualizations and what information we can conclude from it. The first dataset we are using is the dolphin social network dataset that is publicly available on casos site.[1] There was a study done on this where the dolphins were living off the cost of New Zealand called the doubtful sound. This dataset contains 62 nodes or dolphins in the community and has 159 edges where each represents the association between them.

The second dataset we are using is the ‘US Airport Traffic In 1997’.[2] This dataset has 332 nodes which represents the different airport and 2126 different edges where they represent the connection of that airport with other airports. From this visualization we will see how busy and what is the busiest airport from the year 1997.

The Third dataset we chose was the ‘Comic and Hero Network’ and this comic heroes are coming from the Marvel Universe.[3] This dataset is very large compared to the previous two dataset with 19286 nodes and 96519 edges. For this network visualization, we will try to see the relationship between Comic heroes and highlight some key characters for the entire marvel universe.

# Methodology and results

Chart, line chart

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generated

Figure 1. Degree of Dolphin Network Dataset

Figure 2. Modularity of Dolphin Network Dataset

Chart

Description automatically generated

Figure 3. Betweenness Centrality of Dolphin Network Dataset

For the Dolphin dataset we utilized average degree, network diameter and modularity algorithms or statistics as that is what it’s called in Gephi. The average degree algorithm tells us the average number of edges connected to a node.[4] From figure 1, we can see that the number of degrees that were generates between the dolphins. The statistics that were computed on Gephi for average degree for the dolphins was 5.129, which means that on average the dolphins have 5 edges connected to a node. From figure 1, we see the minimum and maximum number of edges a dolphin had.

Chart, bubble chart

Description automatically generated The second statistic we used was the modularity class which is a community detection algorithm. Running this algorithm, we identified 5 different communities of dolphins as we can see it from the figure 2. The thirds algorithm we utilized was the network diameter, which shows us the shortest path between the nodes within the graph. From this network diameter we used the betweenness centrality distribution that measures all the shortest path between every pair of nodes of the network and then count how many times a node is on a shortest path between two others. The average path length was found to be 3.36 and diameter of 8. (Figure 3).

Fig 4. Dolphins’ dataset with Force Atlas 2 Layout and statistics applied mentioned in Fig 1,2,3

In figure 4, you can see the final out of the dolphin network graph. This network graph was made by using the 3 different algorithms mentioned above in fig 1,2,3. As we see that each node has a different size, and this was generated with using the betweenness centrality. From this figure we see that dolphin SN100 has the biggest size because of how close this dolphin is to the other dolphins. We also see that every dolphin in the network is assigned to a class and each class is represented with different color. This is where the modularity class helps us to distinguish the different dolphins in a network. The layout used for figure 4 was the force atlas which helps us separate out each node and beautifies the network by using different properties like repulsion strength and adjust the graph by size so there is no overlapping. We can see in figure 4 that each edge is curved; this was done in the preview section of Gephi and used default presets to make it look nicer.

Table

Description automatically generatedApplication, table

Description automatically generated

Figure 5. Degree Distribution of US 1997 Air Traffic Network Dataset

Figure 6. Size Distribution of US 1997 Air Traffic Network Dataset

Table

Description automatically generated

Figure 7. Betweenness Centrality of US 1997 Air Traffic Network Dataset

Again, for the US 1997 air traffic dataset we are utilizing the same algorithms and dataset for figure 5,6,7 that were mentioned in previous figures 1,2,3. From the graph above we see that on average each node or airport has about an average 6.4 degree to other nodes. From figure 6 we can distinguish that there are 4 unique class where each of this airport nodes belong to.

Chart, bubble chart

Description automatically generated Fig 8. US 97 Airport Traffic dataset with Force Atlas Layout and statistics applied mentioned in Fig 5,6,7

The network visualization for figure 8 was made by using the force atlas layout and 3 different statistics mentioned in figure 5,6,7. By adding the colors, labels and adjusting the node sizes by degree, it shows a clear picture of the US airports traffics pattern and popularity. Based on the size of the node we can also tell which airports were the busiest. From the visualization we can see that Chicago’s O’Hare airport, followed by Dallas Fort Worth International Airport were the busiest. Due to so many edges and nodes it is hard to track and see the patterns. The reason that the Chicago airport was the busiest because it is connected to other 139 node which is the highest. Because of that there will be a lot of traffic going back and forth which is why the size of that node is the biggest. Therefore, it could be the central point for the network due to high degree.

Table

Description automatically generatedGraphical user interface, chart, application, table

Description automatically generated

Figure 10. Modularity of Comic Hero Network Dataset

Figure 9. Degree Distribution of Comic Hero Network Dataset

Table

Description automatically generated

Figure 11. Avg. Weighted Degree Distribution of Comic Hero Network Dataset

As you can see for figure 9, 10, and 11 that for the comic hero dataset, we have looked at the average degree, avg. weighted degree, and modularity class algorithms. When running the statistics, we found out that the number of communities that were generated were 55 and that was to be expected because of how big of the dataset is and how many comic characters there are. Then we used the average degree calculation for determining the size of the node. The average degree was 10.009 and we filter on the node size with minimum degree of .5 to maximum degree of 400 to only focus on the relevant comic heroes.

![A picture containing outdoor object, colorful

Description automatically generated]()

Figure 12. Comic Hero Network Visualization

Figure 12. is a comic hero network visualization generated using the force atlas 2 layout which allow the connected nodes to be drawn together and other connected nodes to be separated from other nodes. This will amplify the relation between the nodes and make a clear visualization for analysis. As mentioned earlier that we ran different statistics on the graph to create different measurements to filter on the graph. From the running the statistic we see that spiderman had 1625 connection between different class and nodes, which is the most and then followed by Captain America with 1352 and Iron man with 1168 connections. In figure 12 you can see that each nodes are colored and put into different class by using the modularity and average degree filters calculated from the algorithms mentioned in figures 9, 10, and 11.

# Conclusions

Gephi is very powerful and useful data processing software that is opensource. In this report we showed you how Gephi provides us with different layouts and algorithms/statistics which can be used to discover factors, communities, and other interesting relationships on different types of networks. However, there are few flaws that we encountered when using Gephi. Gephi is slow and cannot handle large type of dataset because it is very hardware hungry software. Also, it is missing some features like undo, sudden lag, glitch, and other bugs than can definitely hinder your visualization or analysis.

##### References

[1] “Dolphins.” Homepage, <http://www.casos.cs.cmu.edu/computational_tools/datasets/external/dolphins/index11.php>.

[2] “US Air97” [Online] available, <https://github.com/gephi/gephi/wiki/Datasets>

[3] “Comic and Hero Network” [Online] available, <https://github.com/gephi/gephi/wiki/Datasets>

[4] Introduction to Network Analysis, <http://thomaspadilla.org/na2014/>.