Graph Analysis and Social Networks: A Summary of Individual Practical Work

Ilic Ema (200837)
email: ema.ilic9@gmail.com

Graph Analysis and Social Networks (103000903), Master's Degree EIT Digital in Data Science

Academic Year 2020/2021

1 Introduction

Social network analysis is the process of investigating social structures through the use of networks and graph theory.It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks,memes spread, information circulation, friendship and acquaintance networks, business networks, knowledge networks, difficult working relationships, social networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines. These visualizations provide a means of qualitatively assessing networks by varying the visual representation of their nodes and edges to reflect attributes of interest.

2 Context of the Analysis

Now that the general introduction to the graph analysis is clear, the purpose of this research can be assessed. Namely, the idea of the project was the explorative and community analysis of certain movies in the MovieGalaxies MovieGalaxies Dataset. The purpose was to explore the possible tools that can be used for working with graph structures such as the .gexf format which was the format in which the graphs were provided in the dataset.

Some of the questions answered in the Explorative Analysis were:

- What is the distribution of Ratings of the whole dataset?
- What is the mode of the rankings?
- What are top 10 movies with the largest cast?
- What are the top 10 movies with the smallest cast?
- Who is the most important character in a movie sequel?
- Who are the most important characters according to the betweenness centrality measure?

On the other hand, some of the questions answered in the Community Analysis were:

- Who is the most important in the community?
- Who is the most difficult to reach?
- How many communities are there?
- Who belongs to these communities?

3 Description of the Dataset Software Used

The "Moviegalaxies Dataset. Emilio Serrano. 2015.zip" consists of 808 social networks. Each network is codified in GEXF (Graph Exchange XML Format). These files were advised to be analized using Gephi and/or NetworkX software. While both the softwares were tested, is was soon understood that NetworkX provides more flexibility and a wider scope of possible analysis and visualizations.

4 Explorative Analysis

While many different aspects of the analysis were assessed in the Explorative Analysis, the first step was assesing the movie ratings, and observing the mode.

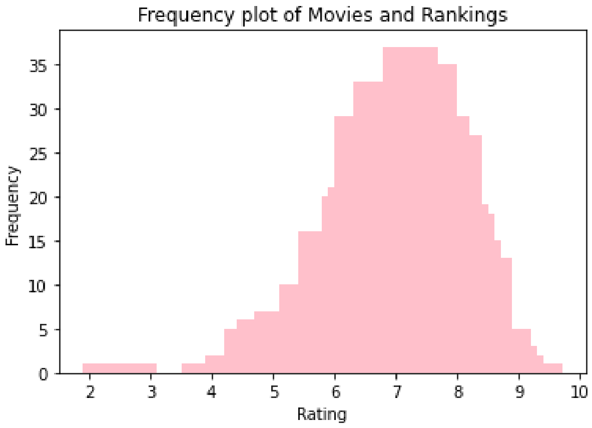
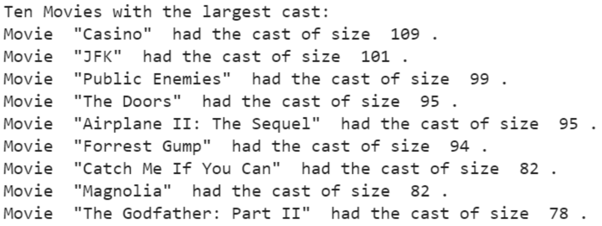


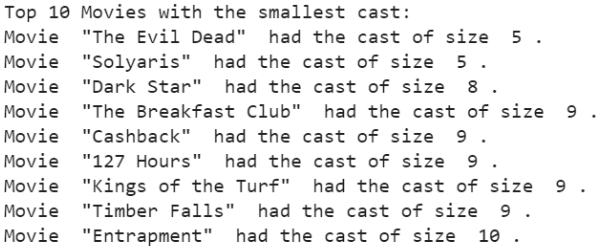
Figure 1: Modal Ranking of Movies

As can be observed from the plot, the IMDB rankings have a unimodal distribution. They take on values between 2 and 9.6, with the modal ranking between 7 and 8. Anything above 8.00 can be considered a ’good ranking’.
On the other hand, further analysis was done to analyse the cast size per each movie. Namely, on the pictures below, one can observe the two movies with the largest and smallest cast. Namely, the movie Casino has the largest cast of 109, while the movie The Evil Dead has the smallest cast of 5 members.



```
Ten Movies with the largest cast:
Movie  "Casino"  had the cast of size  109 .
Movie  "JFK"  had the cast of size  101 .
Movie  "Public Enemies"  had the cast of size  99 .
Movie  "The Doors"  had the cast of size  95 .
Movie  "Airplane II: The Sequel"  had the cast of size  95 .
Movie  "Forrest Gump"  had the cast of size  94 .
Movie  "Catch Me If You Can"  had the cast of size  82 .
Movie  "Magnolia"  had the cast of size  82 .
Movie  "The Godfather: Part II"  had the cast of size  78 .
```

Figure 2: Movies with the largest cast



```
Top 10 Movies with the smallest cast:
Movie  "The Evil Dead"  had the cast of size  5 .
Movie  "Solyaris"  had the cast of size  5 .
Movie  "Dark Star"  had the cast of size  8 .
Movie  "The Breakfast Club"  had the cast of size  9 .
Movie  "Cashback"  had the cast of size  9 .
Movie  "127 Hours"  had the cast of size  9 .
Movie  "Kings of the Turf"  had the cast of size  9 .
Movie  "Timber Falls"  had the cast of size  9 .
Movie  "Entrapment"  had the cast of size  10 .
```

Figure 3: Movies with the smallest cast

On Figure 4, on the other hand, a frequency distribution of the movies with a certain cast size can be observed. Namely, it is clear that the mode of this unimodal distribution is between 25 and 40, which means that most movies have approximately that number of cast members. On the other hand, some outliers can be observed at values of 0,1, and 2. It is assumed that they errors in the dataset as these movies didn’t show up during the analysis of the smallest cast which was done by merging the IMDB rankings table with the .gexf files by leveraging on the movie_id.

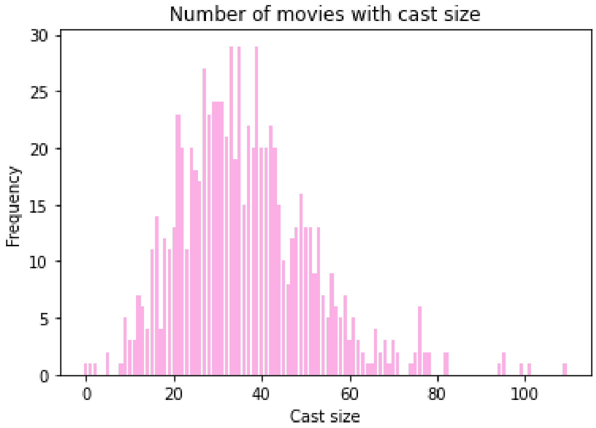


Figure 4: Modal Ranking of Movies

The idea was to create a multigraph of a movie sequence, such as the Alien sequence (five movies). When the multigraph was created using the appropriate algorithm, and then visualized, the outcome can be observed on the figure below.

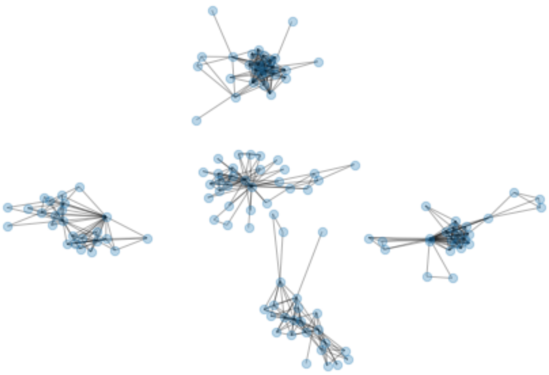


Figure 5: Alien multigraph

Even though the intention was to observe the relationships between the same characters through different movies, the problem encountered here was the fact that the same characters have different IDs in different movies. Thus, the possible alternative would've been to relabel node IDs or to use the node attribute 'label' (name of a movie character) instead of the node ID to establish the connections, however, it was decided that this was outside the scope of this course.

In addition, the Final Destination movie was plotted as a graph, and character names were set to be dispalyed. It is interesting to observe that Alex and Clear are the main characters with the most connections, while some secondary characters such as ludworth and John Denver have very few connections.

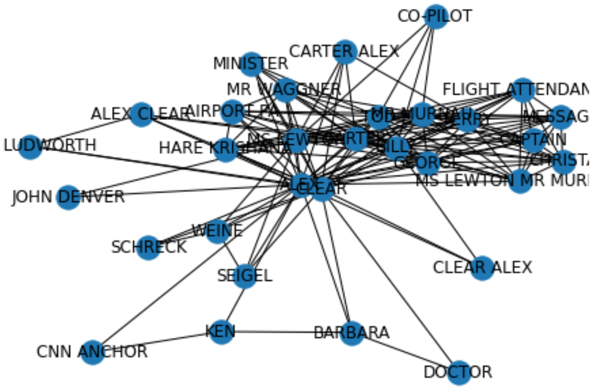


Figure 6: Final Destination Graph Network

To deepen the analysis of the main vs. secondary characters, another interesting and helpful visualization was achieved, called the **Ego Graph**.

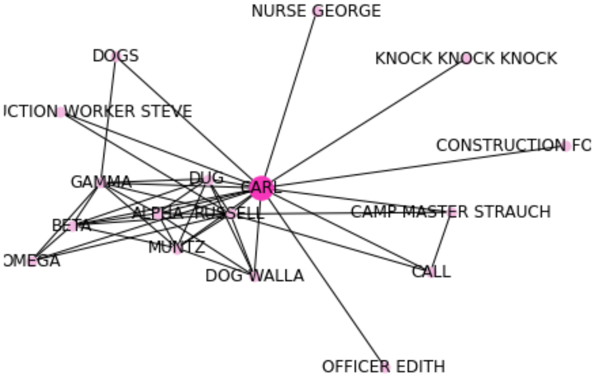


Figure 7: UP movie Ego Graph

With the Bright pink, big node in the middle, it is easy to observe that the main character of the movie is Carl. The secondary main character is Russel while there are also some completely secondary characters with very few connections, such as the Dogs, Nurse George, etc.
Finally, in network theory, a giant component is a connected component of a given random graph that contains a finite fraction of the entire graph's vertices. Thus, at different thresholds

(defined as p_value, we can observe the giant component growing, and we can also observe other quite large connected components.

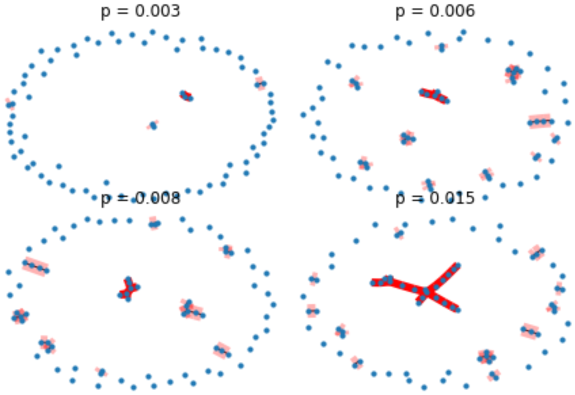


Figure 8: UP movie Giant Component Network

Moreover, to wrap up the explorative analysis, the a graph for the godfather movie was made taking into consideration the betweeness centrality measure. In graph theory, betweenness centrality (or "betweeness centrality") is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. For the sake of visualizations, only the first fifteen nodes with the largest betweenness centrality were considered.

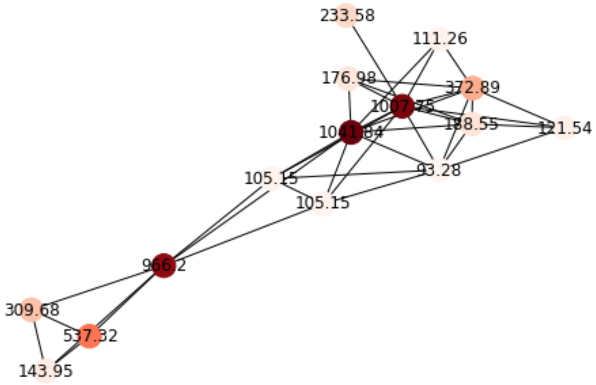


Figure 9: Godfather Movie Betweeness Centrality Network Visualization (First 15 nodes with the largest Betweenness Centrality)

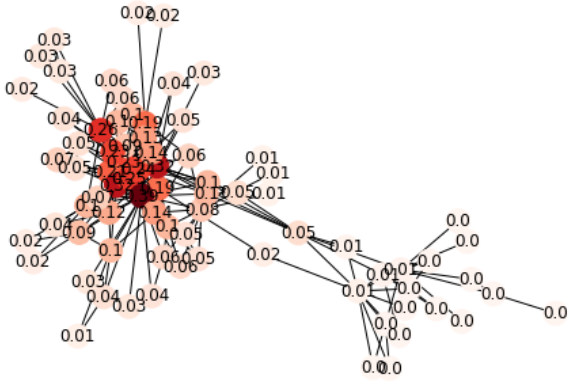


Figure 10: Godfather Movie Eigenvector Centrality Network Visualization

For comparison, the same graph was plotted for the very same movie (Figure 10), this time considering eigenvector centrality measure instead. In graph theory, eigenvector centrality (also called eigencentrality or prestige score) is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

5 Community Analysis

In the study of complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally. In the particular case of non-overlapping community finding, this implies that the network divides naturally into groups of nodes with dense connections internally and sparser connections between groups. But overlapping communities are also allowed. The more general definition is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same community(ies), and less likely to be connected if they do not share communities. A related but different problem is community search, where the goal is to find a community that a certain vertex belongs to.

On the plot below, observe two community networks of the Goodfatrer II movie: the first one is created out of dendrogram cut at level 0 (12 communities), and the second one created out of communities which came out when the dendrogram was cut at level 1 (6 communities).

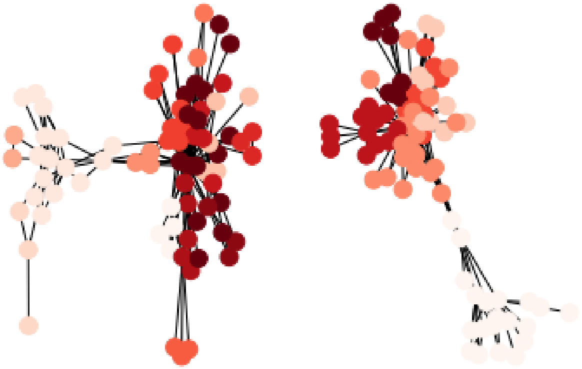


Figure 11: Godfather II Community Networks Cut at layer 0 and 1, respectively

For the reference, the best partition was calculated at 0.434

6 Open Questions and Remarks

Even though it was ultimately decided to use NetworkX, the Gephi software was tried and tested as well. Ideally, with more resources provided (time most importantly), an interesting combination could be the creation of multigraphs in NetworkX, and importing them in Gephi in order to visualize them in different and useful ways.

For the reference, Before it was understood that NetworkX can take the data provided as input, it was attempted to write a function for accessing different node and edge attributes in the .gexf file. The code is left at the end of the python notebook to be assesed together with the rest of the work.

References

[1] Fletcher, Jack McKay and Wennekers, Thomas. *"From Structure to Activity: Using Centrality Measures to Predict Neuronal Activity"*. Neural networks, 12(1):145–151,

2017.

[2] David Austin *"How Google Finds Your Needle in the Web's Haystack"*. nature, 323(6088):533, 2019.

[3] Yan, Xiaoran; Jacob E. Jensen; Florent Krzakala; Cristopher Moore; Cosma Rohilla Shaliz *"Parsimonious Module Inference in Large Networks"*. Journal of Statistical Mechanics: Theory and Experiment, 323(6088):533, 2019.

[4] Hamdaqa, Mohammad; Tahvildari, Ladan; LaChapelle, Neil; Campbell, Brian *"Cultural Scene Detection Using Reverse Louvain Optimization"*. Neural Information Processing Systems, 25:1097–1105, 2014.