Exploring Cinematic Franchises Viewed as Network Architectures using the IMDB Database

HNRS 304: Networks in the Info Age
Final Report Submission

Researcher: Cameron J. Calv
Professor: Prof. Steve Weber

ABSTRACT

This project explores how cinematic franchises look like plotted as network graphs. Nodes in the presented graphs represent performers and connections represented two performers sharing the same film. The properties of the movie franchises are theorized based on the appealing network graphs' appearances. Patterns in the graphs are explored, and tabulations of the interconnectedness of each graph represented by the calculation of the clustering coefficient are presented. The variably separated bubble structures give details of the interconnectedness of the graph from a qualitative point of view while the clustering coefficient gives a mathematical representation of how performers relate to the franchise both on average as maximum.

INTRODUCTION

As the information age continues to evolve, many aspects of the world are being placed in the perspective of data structures. Quantizing this information in a way that makes it possible for a computer to manipulate is an important step in analyzing this data. One such data structure of interest, and the main topic of the course is the network which is essentially a collection of nodes connected in various ways by edges. Analyzing data represented as a network can have a multitude of benefits for various fields. In biology, the complex interactions of protein structures can be represented as a network of proteins connected by metabolic processes. The internet is naturally a network of connected end-hosts that use intermediate nodes to exchange information in a distributed fashion. In this project, the network of performers in cinematic franchises is explored and presented in a visually appealing manner to gain insight to the interconnectedness of these franchises.

For this project the network is constructed under the following rules. Every node in the network represents a performer that is listed under the main page cast list for a cinematic entry in the Internet Movie Database (IMDB). For a particular IMDB keyword search, the displayed selection of movies, tv shows, minifilms, etc. are referred to as a cinematic franchise for the purposes of this project. Each node in the network is colored in a way that represents how many appearances the corresponding performer has had within the franchise. If a node is shown with a color representing the number three, then this performer has appeared in three movies out of the total number of films within the franchise. A connection is drawn between two nodes if the two performers represented by those nodes have appeared in the same film.

The process by which the data was gathered for this report involved three main steps: web scraping, data cleanup, and data plotting. The web scraping part of the project consisted of querying the IMDB website for all films that fall under a particular keyword search. Each movie entry is further explored to see which cast members are displayed on the main page for a title. The program stores every movie and its associated cast members. The data cleanup section creates the network representation of the film data. Every cast member is looked up in reverse within the list of films to determine how many appearances they have. A list of every movie a performer has been in is tabulated and checked against those for all other performers. Naturally, if a film appears in two performer's appearances list, they share a film and the connection is added to the network representation. Finally, the program takes all the connection data and plots them according to a particular plotting rule that creates a visually appealing network plot. The function provided by the *networkx* Python package allows the use of the Kamada-Kawai algorithm to create a force-directed representation of the network plot which causes various visual effects of the graph that will be discussed following the presentation of the most interesting plots.

EXPERIMENTAL RESULTS

Twenty-three film franchises were plotted in order to get idea of what type of structures appear within the network graphs. Figure 1 shows six of these plots that will be used to describe the various structures that seem to manifest when the network graphs are plotted according to a force-directed algorithm. The plotting function uses the Kamada-Kawai algorithm which produces these network structures.

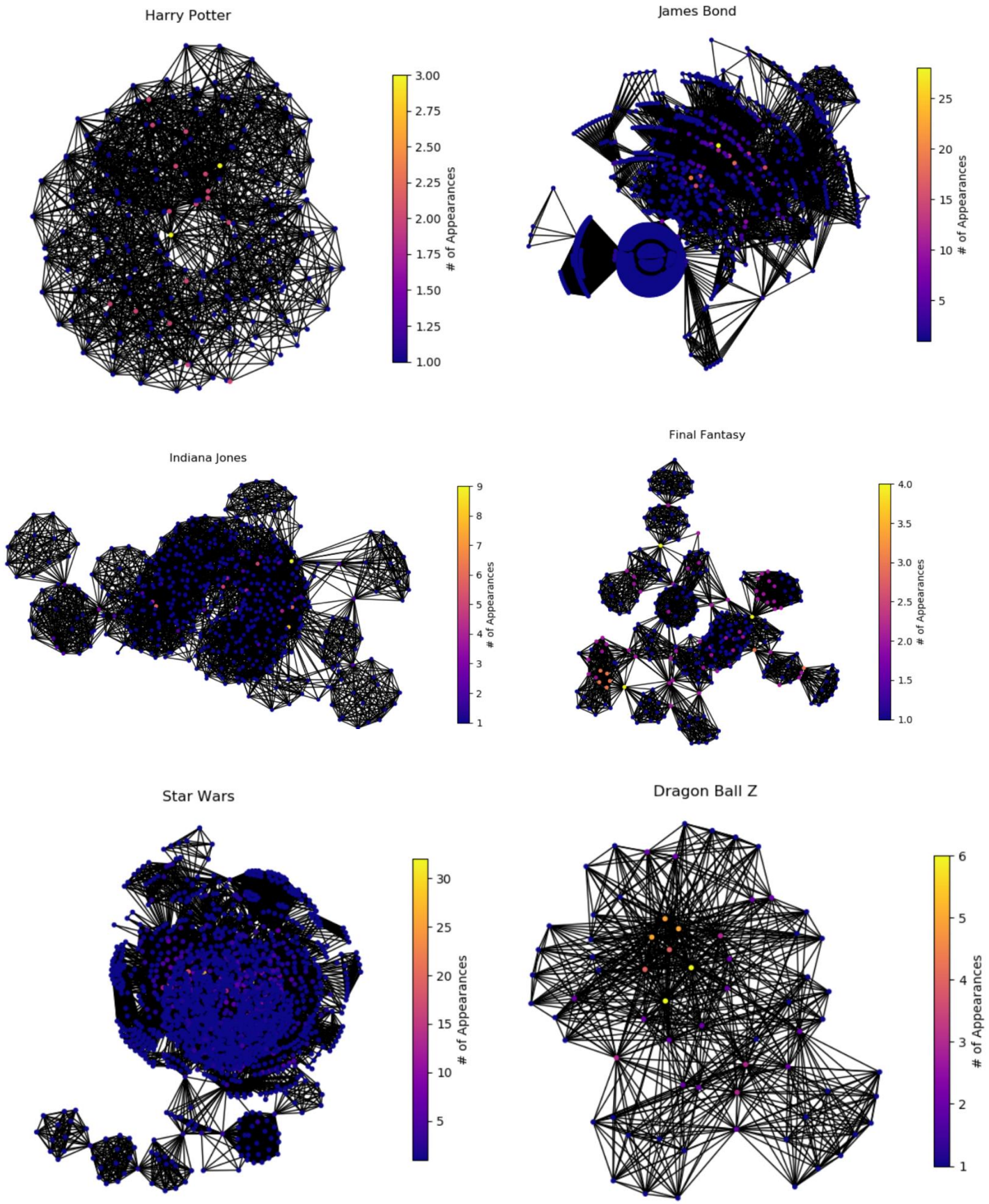


Figure 1: A collection of figures showcasing six cinematic franchise plots. The selected plots show certain distinct structures that will be explored and discussed. Plots shown are plotted using Python's matplotlib library with a force-directed graph structure.

Below in Table 1 every franchise that was explored is tabulated along with clustering characteristics of the network graphs they produce. The table contains the total number of nodes which corresponds to the number of performers that appear on the front page of the titles that make up the franchise. The maximum and average clustering coefficient are also provided for each franchise. The clustering coefficient is a measure of how interconnected the nodes in the graph are. It is calculated by taking the number of connections a node has to its local neighbors and dividing it by the total number of connections it could have with its local neighbors. It is intended to be used to see how close each performer is to every other performer with closeness thought of in the same way as the 'Six Degrees of Kevin Bacon' thought.

Table 1: Tabulated results of the clustering characteristics of all cinematic franchises explored using the code for this project. Each franchise name is the keywords used to find all the title.

Franchise Name	Total Nodes/Performers	Maximum Clustering Coefficient	Average Clustering Coefficient
Dragon Ball Z	69	0.555	0.448
Final Fantasy	298	0.703	0.489
Harry Potter	203	0.527	0.477
High School Musical	446	0.500	0.492
Indiana Jones	531	0.750	0.493
James Bond	1897	0.833	0.490
Kingdom Hearts	109	0.595	0.481
Legend of Zelda	134	0.783	0.468
Lord of the Rings	298	0.582	0.471
Mad Max	133	0.500	0.492
Marvel Cinematic Universe	477	0.833	0.477
Mission Impossible	101	0.500	0.485
Nintendo	1223	1.000	0.489
Planet of the Apes	240	0.577	0.470
Pokemon	521	1.000	0.513
Rocky	209	0.833	0.484
Sega	296	0.681	0.485
Star Trek	1503	0.818	0.480
Star Wars	1880	1.000	0.474
Stargate	199	0.590	0.499
Terminator	397	0.833	0.483
The Muppets	847	1.000	0.519
X-Men	699	0.750	0.474

DISCUSSION

Observing the plots that were generated based on the movie database, there are some common shapes that manifest within the force-directed network representations. Most, if not all the shapes are some form of bubble or deviation from a complete one. The bubble-like structures manifest because of the property that every cast member in a movie is connected to each other cast member of the same movie. To that point, whenever a movie has a sizeable number of cast members on the front page, the connections representing that movie tend to group together into a fully connected mesh. What occurs in the Harry Potter plot is that many of these bubbles are also connected in a partially connected mesh, grouping the entire franchise into one larger bubble. To this extreme, movies that share similar casts tend to create plots that are very circular, with the perimeter of each bubble being made up of inner bubbles that correspond to the individual movies within the franchise.

A slight deviation occurs when a portion of the movie franchise fits this category and another portion does not. Depending on the proportion of movies within the franchise that share a similar cast to those movies that do not share a similar cast, the shape of the plot can be starkly different. In the case of the James Bond plot, the intense interconnected-ness of the large portion of the franchise creates a dark and dense core of

what seems to be two concentric bubbles with connections to multiple incomplete out-perimeter bubbles. The level of connectedness between the outer arcs and the inner core varies greatly, but it appears that each perimeter itself is fully connected. I believe that this manifests because a subset of movies shares similar cast members with the movies of the core, but also have their own set of cast members not present in the core movies.

Another deviation occurs in the Indiana Jones plot, where there is a clearer picture of the core movies that all share various cast members. There are also other titles in the franchise that clearly share a very small number of cast members to that of the core creating this mesh of bubbles connected by anywhere between one and a great many nodes. It is interesting to note that within this plot to the far left there is a clear connection between two bubbles with a small number of nodes connecting the two. The interconnecting node is of a lighter color than those around it since it corresponds to a cast member making more than a single appearance. The observation that it is the performers with the most appearances that are most likely to be connected to other movie bubbles is not surprising and is further shown in the next plot.

The Final Fantasy plot shows a collection of titles that do not seem to share many cast members are appeared to be clearly distinct from one another. Exceptions arise when there are one or more cast members that make multiple appearances. These performers are highlighted with a lighter color than the performers surrounding them. Since each movie is distinct with little cast-similarity with the other titles, the plot manifests as a collection of bubbles that share various degrees of connectedness with its surrounding bubbles.

A plot that shows what I believe to be a culmination of all the structures shown in the generated plots is that of the Star Wars franchise. The plot shows the core structure with what appear to be outer and inner perimeter bubbles as well as a chain of bubbles extending from the bottom. These may be titles of a different type of media than the majority for the franchise such as video games, tv shows, or internet docuseries and reviews. On the other hand, a plot that does not seem to show too much structure is that of the Dragon Ball Z plot which seems to show a high degree of interconnectivity but without enough cast members to create the very obvious bubble structures present in the other graphs.

I attempted to quantify the interconnectivity of the graphs by computing the average and maximum clustering coefficients for each of the plots. Though there are quite a few franchises to go over, the points of the measurements are summarized here. Some franchises show to have a maximum clustering coefficient of unity. This means that a performer is connected to each and every other one surrounding it. From a global standpoint, that means that a performer has been in every single title. Though this is not often the case due to the vast mediums of the film titles, it does appear. I original thought that perhaps the higher the clustering coefficient was for a particular plot, the more interconnected it would appear to be. Despite that being the case mathematically, I share the thought that the plots that visually appear to be the most connected are the ones that do not sport very high clustering coefficients such as the James Bond plot with its intensely connected core.

CONCLUSIONS

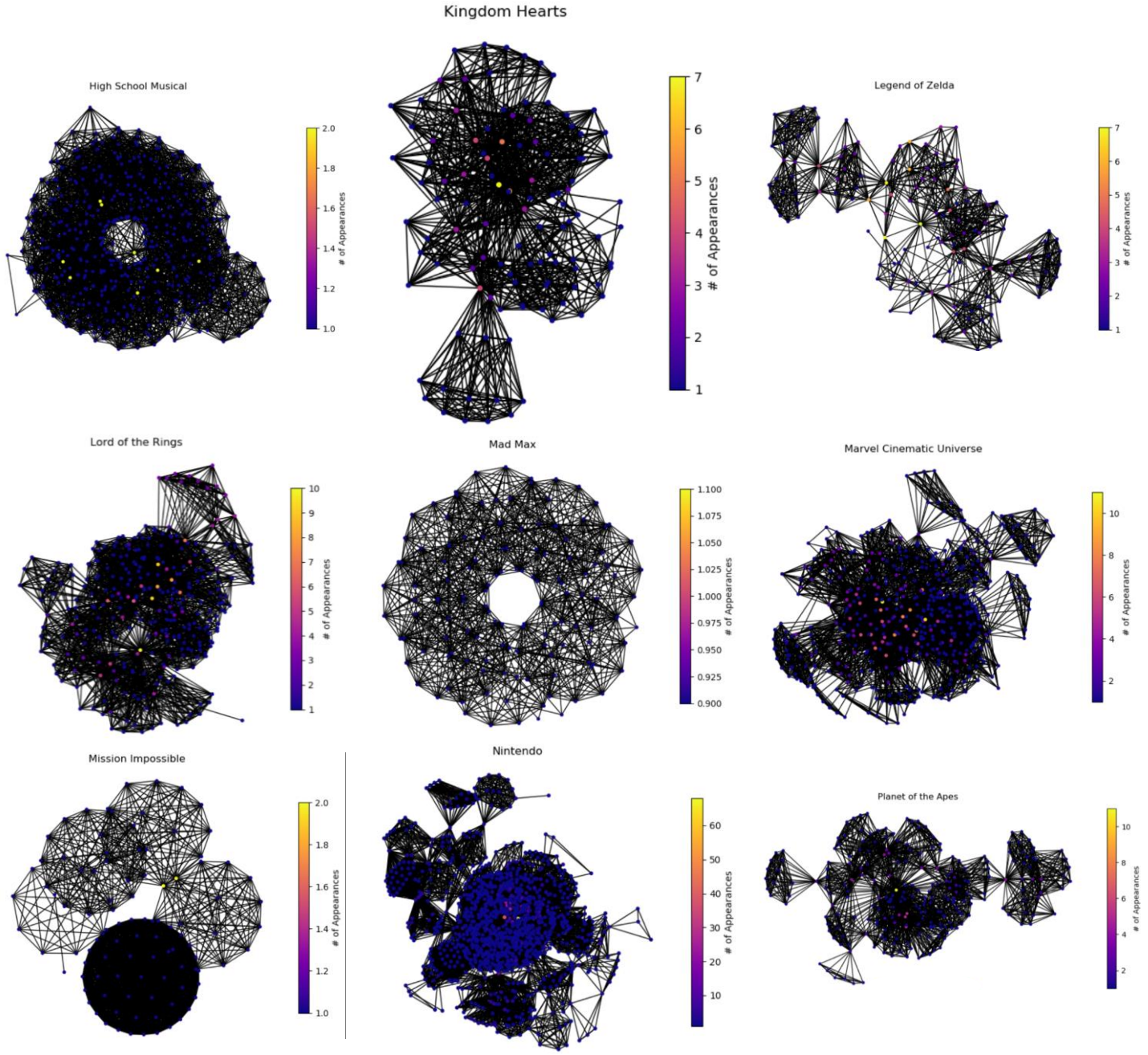
The plots generated for this project show varying degrees of interconnectivity that may be represented quantitatively or (by my own personal preference) qualitatively. The use of computer aided technology to scrape the Internet Movie Database for casting information allowed the creation of complex network graphs representing our favorite cinematic franchises. Plot shown not only give a visual representation of interconnectivity within movie franchises, but also key insights about various attributes of the franchise such as the total number of performers, or how many appearances any performer has made within the franchise. The end result was a collection of wonderfully abstract plots that created yet another visual perspective of cinema.

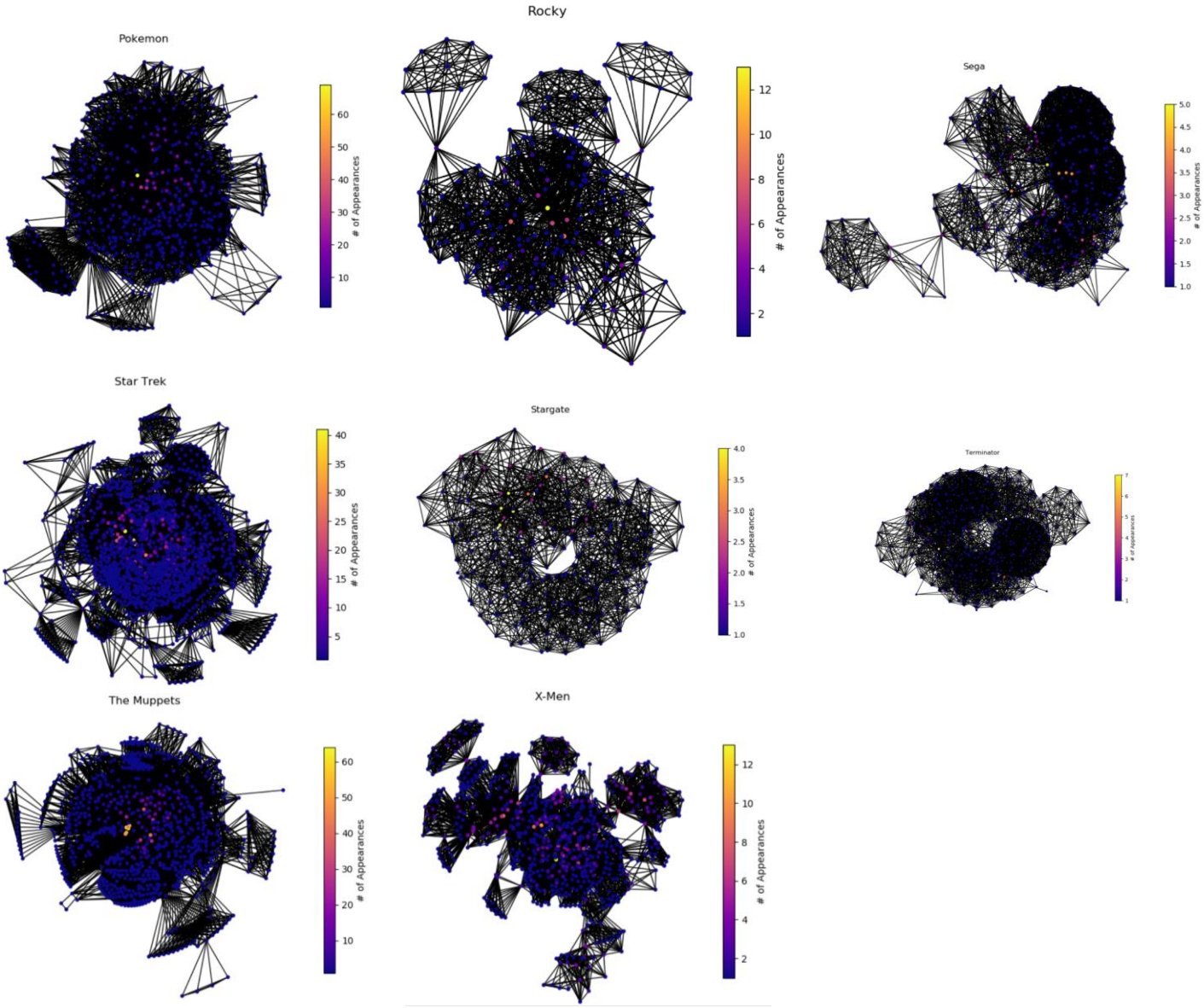
DATA AVAILABILITY

All data gathered for the use of this report is publicly available from the IMDB website linked at *imdb.com*. The source code utilized to generate these plots can be found online in the author's public GitHub repository at *https://github.com/cjawesomest/imdb_keyword_network_plotter.*

APPENDIX

This section shows the other network graphs that were not included in the upper half of the report. For higher resolution plots of the networks, refer to the ‘doc’ folder included with the GitHub repository.




Pokemon
# of Appearances
Rocky
# of Appearances
Sega
# of Appearances
Star Trek
# of Appearances
Stargate
# of Appearances
Terminator
# of Appearances
The Muppets
# of Appearances
X-Men
# of Appearances