IMDB Actor Network Analysis

By: Andy Dotter, Chris Nobblitt, Jake Lasky, Kyuhyun Jeong

Data: We used a dataset containing information pulled from IMDb in 2018. Our dataset consisted of 5000 movies and the 3 biggest actors in each of those movies. The data also included valuable information such as each movie's IMDB score and the gross revenue from the movie along with other various attributes.

Problem Statement: Our goal was to analyze IMDb data to find anything interesting about the relationship between actors who played in the same movies. Particularly, how did these actors group together in terms of revenue and IMDB score, which actors were the most influential, and if there were clusters of main actors and their supporting cast. We broke our project down into three main sections described below:

1. Show how actors interact with one another and the connections between them. Actors are nodes and number of times they act with each other are edges

2. See the interactions between a movie's IMDb reviews, gross revenue, and actors. Actors are nodes and ratings or revenue are edges.

3. See which actors / movies bridge groups of movies. Movies are nodes and number of common actors are edges.

Section 1: Actor Interaction Top 250

Data Prep: First we pulled our IMDB dataset into R. We sorted our movies in descending order based on IMDB rating (0 – 10 scale, 10 being the highest). We then kept the top 250 movies and the three biggest actors associated with those movies. We then exported this in the form of an adjacency list into Gephi. We then analyzed the resulting network graph with actors as nodes and edge weights as the number of times the actors played together in movies. The resulting network graph can be seen below:

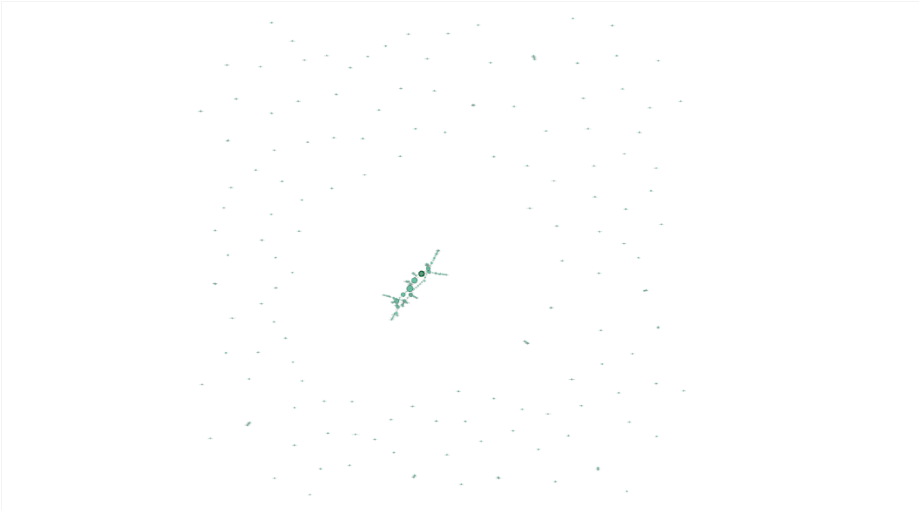

Figure 1. Social Network Graph for top 250 actors by IMDB score. Nodes are actors, weight of edges is number of times they have acted with each other.

Since we took the top 250 movies, we find a huge ring of unconnected actors surrounding a large group of actors who act in many of the top 250 movies. Because of this we isolated for the largest connected component and focused on actors who frequently acted with each other. Below is the graph for the largest connected component.

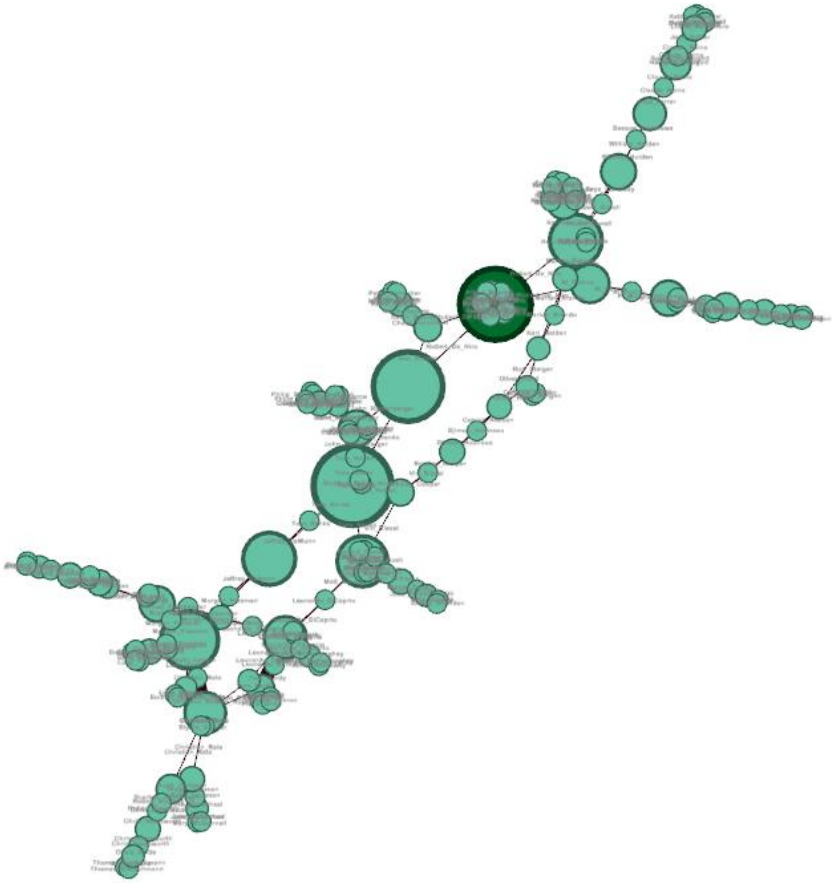

Figure 2. Largest connected component for top 250 movies by IMDB score. Actors are the nodes, number of times they acted with each other are the edge weights.

Interesting Findings:

Actors in the main connection include Morgan Freeman, Leonardo DiCaprio, Tom Hardy, Christian Bale, Hugh Jackman, Scarlett Johansson, Chris Hemsworth, Brad Pitt, Robert De Niro, and Tom Hanks. This is most likely due to the superhero movies (The Avengers and its supporting films) as well as Batman and then the odd film here and there that allow them to connect to one another.

Section 2: Actor groups based on IMDB score and Gross Revenue

Data Prep: Utilizing our list of 5000 movies from IMDB we isolated the data to the top 250 movies based on gross revenue. We converted the data frame to an igraph in R and then calculated the number of connected components and isolated our dataset for the largest single connected component. We then exported the largest single connected component data as a gexf file. Then using Gephi we looked at clusters of actors and the influence the actors had on the overall network.

Graph 1 (revenue): Actors relationship for the largest component of a social network graph for the top 250 movies ranked by gross revenue. Actors are nodes, edges are weighted based on gross revenue from the movies the actors were in.

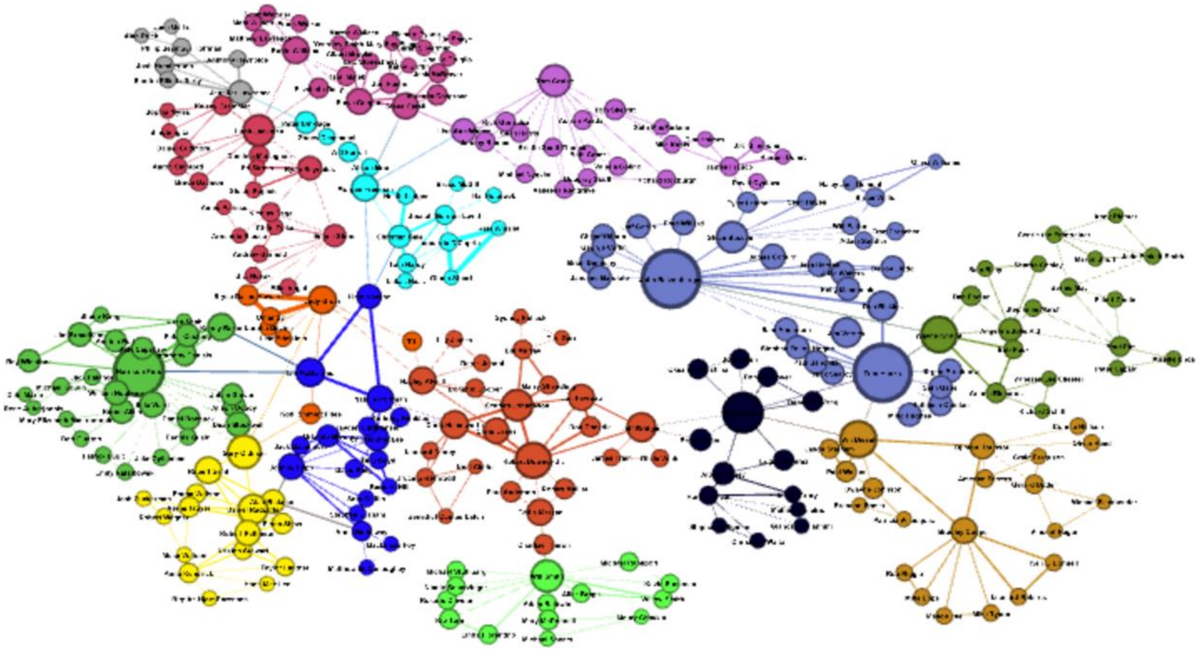

Figure 3: Social Network of top 250 movies based on gross revenue (edge weights = revenue). The Graph was isolated for the largest connected component. Size of the nodes is the Eigenvector centrality and color is based on modality clustering.

Clusters using Modularity (revenue)

When calculating the modularity maximization with edges weighted by gross revenue for movies and actors. When we clustered we ended up with 15 clusters. Each of these clusters had a main actor in each group, the main actors for each group can be seen in the network graph by the largest node in each cluster (color). These actors are:

1. Matt Damon
2. Gary Oldman
3. Ian McDiarmid
4. John Ratzenberger
5. Hugh Jackman
6. Robert Downey Jr.
7. Morgan Freeman
8. Jennifer Lawrence
9. Vin Diesel
10. Judy Greer
11. Robin Williams
12. Tom Cruise
13. Will Smith
14. Harrison Ford
15. Wayne Knight

Most influential by revenue

We ran an Eigenvector Centrality on my network analysis with the weight of the edges as the gross revenue from the associated movies and actors. The actors that are most influential from a gross revenue standpoint are:

- i. John Ratzenburger (1.0)
- ii. Tom Hanks (0.993)
- iii. Harrison Ford (0.828)

a. The most influential actor that is part of only one triangle is Don Rickles who is part of both John Ratzenburg and Tom Hanks triangle and no other actors. This movie is Toy Story 3.

Graph 2 (IMDB score): Actors relationship for the largest component of a social network graph for the top 250 movies ranked by Gross Revenue. Actors are Nodes, Edges are weighted based on the IMDB rating of the movie that the actors were in.

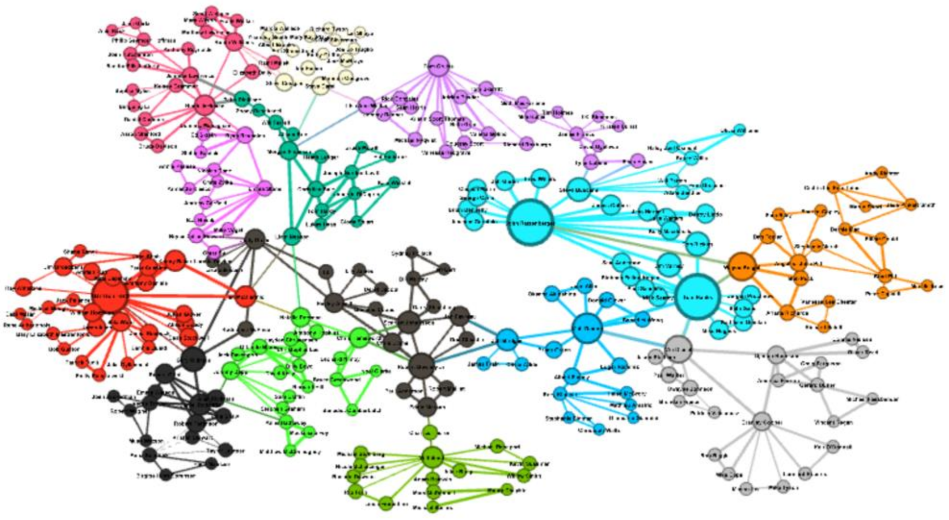

Figure 4: Social Network of top 250 movies based on gross revenue (edge weights = average IMDB score in this graph). The Graph was isolated for the largest connected component. Size of the nodes is the Eigenvector centrality and color is based on modality clustering.

Clusters using Modularity (IMDB score)

When calculating the modularity maximization with edges weighted by IMDB score for movies and actors. We ended up with 14 clusters where each of these clusters had a main actor in each group, the main actors for each group can be seen in the network graph by the largest node in each cluster (color). These actors are:

1. Chris Hemsworth
2. Scarlett Johansson
3. Matt Damon
4. Ryan Reynolds
5. Wayne Knight
6. Hugh Jackman
7. Vin Diesel
8. Morgan Freeman
9. Gary Oldman
10. Steve Carell
11. Harrison Ford
12. Robert Downey Jr.
13. Tom Cruise
14. John Ratzenberger

Most influential by IMDB score

We ran an Eigenvector Centrality on my network analysis with the weight of the edges as the IMDB rating from the associated movies and actors. The actors that are most influential from a IMDB standpoint are:

i. John Ratzenburg (1.0)
ii. Tom Hanks (0.987)
iii. Harrison Ford (0.73)

Again, the most influential actor that is part of only one triangle is Don Rickles who is part of both John Ratzenburger and Tom Hanks triangle and no other actors. This movie is Toy Story 3.

Conclusion & Next Steps:

As you will see both graphs are similar when we analyzed them based on Gross Revenue or by IMDB rating. Some of the similarities of these graphs is due to isolating for the largest component from the top 250 movies based on revenue. If we expanded the number of movies we included and then focused on the largest component we may see a larger difference when we weight the edges by gross revenue or IMDB score. One thing that is interesting to note is that when we look at the individual with the highest eigenvector centrality in each group, the overlap between the graphs is approximately 60%, i.e. 60% of the actors that had the highest eigenvector centrality based on gross revue also had it the highest for IMDB rating. One thing is clear, there is a small number of actors that have a large influence on the success of a movie.

Section 3: Actors that bridge movies

The objective of this section was to see which actors bridge groups of movies. Movies are nodes and number of common actors are edges. Below is an overview of the network.

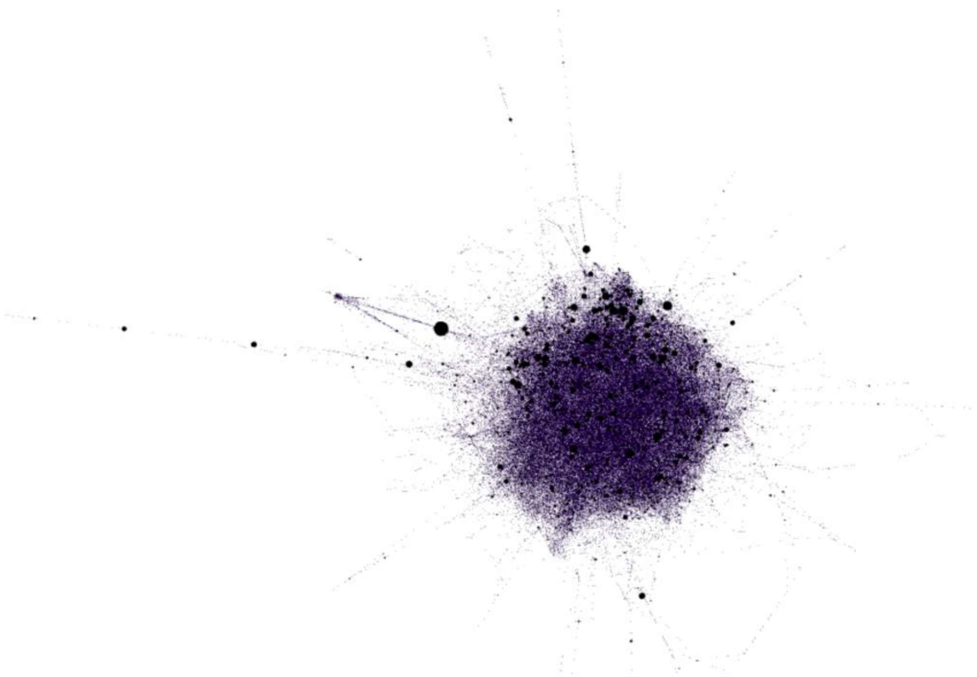

Figure 5. Network graph of all 5000 movies. Movies are nodes and edges are the number of common actors between the movies. Size of nodes are determined by that movies weighted betweenness centrality.

Interesting findings:

Unleashed bridges many movies together because of its stars. Morgan Freeman, Jet Li, and Bob Hoskins never appear in films with each other other than this feature. These three actors work alongside a wide variety of other actors, which connects them to actors and films from a huge spectrum of genres.

Lucky Number Slevin is an interesting film because it has the highest weighted degree, or rather it shares the greatest number of actors with other films. Bruce Willis, Morgan Freeman, and Dorian Missick star in this movie. Bruce and Morgan both star in many movies, and in a few together, which probably accounts for *Lucky Number Slevin*'s high degree. However, this movie also has an average betweenness centrality, suggesting that while these actors appear in many movies, the diversity of their co-actors is surprisingly low. In fact, looking through Bruce's and Morgan's filmography, you see the same names appear commonly for both men. Dorian is an exception to this, but does not appear in many films in comparison.

On the outskirts of this network, we see *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb* connecting most of the movies to an offshoot of movies. Keenan Wynn is responsible for this, taking on more esoteric films like *Annie Get Your Gun* and George C. Scott, who played in some films with bigger names like David Keith.

Overall Conclusion:

When we isolate for the top 250 movies either based on IMDB score or by gross revenue it tends to bring up the same main actors. If we look at all the movies in our dataset, we still see that there are some actors that are very influential in the network. Overall, we see two different types of components in our networks. One is a large component comprised of the big actors in Hollywood, your Tom Hanks, Morgan Freemans, etc. and their connections with each other and other less known actors that play in movies with them. We also see many small components that are not connected to the main actor component. These are undoubtably lesser known movies where all three of those actors played in one particular movie and did not play in any other movies in our dataset. The largest takeaway from our analysis is somewhat expected, Hollywood has a group of influential actors that dominate the big movie scene and many fewer known actors that play in lesser known movies. Our data is influenced by only having a list of the biggest three actors in a movie making it harder to identify all the fringe actors. In addition, the fact that we isolated for the top 250 actors by IMDB rating and gross revenue also inflate the importance of a few actors. Even with this the big actors are still a big deal in Hollywood.

Appendix:

Code Files:

R_Social_Network_Project – R script used to create section 1 dataset for Gephi.

Actor_IMDB_and_Revenue_Network_graphs.R – R script used to create top 250 dataset by gross revenue. This was used for section 2.

Movie_nodes – Python script used to create section 3 dataset for Gephi.

Gephi Files:

Section1_Actor_Centrality – Gephi file with Network for section 1 actor centrality.

Section2_Actor_IMDB_Rating_Modularity_Centrality – Gephi file with the network for section 2 for top 250 movies by gross revenue with edge weights as IMDB ratings.

Section2_Actor_Revenue_Modularity_Centrality – Gephi file with the network for section 2 for top 250 movies by gross revenue with edge weights as gross revenue.

Section3_Actor_Moie_Bridge – Gephi file with network for section 3 where we have movies as nodes and edge weights as the number of common actors between the movies.

Data Files:

Section2_Actor_Gross_Revenue_Analysis – Excel analysis of exported data from section 2 gross revenue network.

Section2_Actor_IMDB_Analysis – Excel analysis of exported data from section 2 IMDB network.

Original_Raw_Data_movie_metadata – This is the original IMDB dataset with all 5000 movies.

Section3_Actor_Movie_Bridge_Data – This is the data fed into Gephi for section 3.

Section1_manual_csv – This is the data fed into Gephi for section 1.