# Network analysis on IMDB dataset

Marinopoulou Theoktisti and Pantelidou Kyriaki-Nektaria

*Abstract*— **In this paper we present the results of IMDB dataset analysis about actors' collaborations in the 1000 most popular movies. For this purpose we exploit some graph properties and metrics. The questions that are to be answered refer to: the top influencers, the most isolated actors from Hollywood and the ones that hold actors' groups together in our network. Moreover, we examine how successful actors' collaborations are and finally which future collaborations we would suggest.**

## I. INTRODUCTION

Network analysis is the process of investigating structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes v (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. In our approach, each edge is weighted and shows how many times two nodes interacted. Some basic metrics of graph theory that we use in our investigation, are: i) degree that gives the number of edges which touch node v ii) closeness centrality which is the sum of the length of the shortest paths between the node and all other nodes in the graph and iii) betweenness centrality which calculates the sum of shortest paths that pass over node v. Additionally, another concept that we focus on, is the detection of communities in our network. A community is defined as the group of nodes that are densely connected internally. For this purpose, the measure that we make use of is modularity, which calculates the strength of division of a network into modules (also called groups, clusters or communities).

## II. DATASET PREPROCESSING

Firstly, we use the Kaggle dataset ”IMDB-movie-data” that contains data of 1,000 most popular movies from 2006 to 2016. The attributes included are: Title, Genre, Description, Director, Actors, Year, Runtime, Rating, Votes, Revenue, Metascore. We continue with some preprocessing steps. The columns we keep are Actors and Rating. Each row in Actors contains 4 actors. We split the row and keep every actor separately, in order to create nodes.csv file. Furthermore, when two actors are contained in the same row, we create an edge that links their nodes. This way, edges.csv file is created. The above process is implemented in Python and Java.

## III. NETWORK ANALYSIS AND VISUALIZATION

The platform we use to visualize and analyze our network is Gephi platform. We feed Data Laboratory with nodes.csv and edges.csv files and as a result we see the graph in Overview tab. The graph is undirected and initially the nodes are overlapping. For an optimized visualization, we run Noverlap layout and then we set nodes' size ranking by Degree, where the ones with smallest degree get a size of 30 and, increasingly, those with the biggest get a size of 100. Also, by default, the biggest weight an edge has, the bolder it appears. The initial graph is represented in Fig.1. It contains 1916 nodes and 5819 edges. The Network Diameter (biggest shortest path) is 9 and its Density is 0.003.

## IV. QUESTIONS ON OUR GRAPH

### *A. WHICH ACTORS ARE THE TOP INFLUENCERS?*

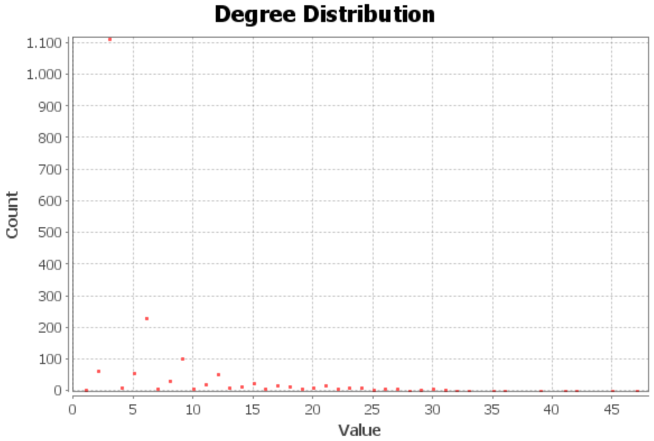
The first thing we want to answer, and the one that can be derived easily from Gephi platform is the actors with the most collaborations, and thus which are the ones who influence our network the most. For this question we are based on degree metric, we run the ”Average Degree” from Gephi's Network Overview and we sort our nodes from the ones with the highest degree to those with the lowest. The greater the degree, the more collaborations an actor has. The result is that the top 3 influencers are:

Fig. 1. Initial Graph Visualization

1) Hugh Jackman (47 collaborations)
2) Brad Pitt (45 collaborations)
3) Christian Bale (42 collaborations)

When running ”Average Degree”, a graph with degree’s distribution is produced as seen in Fig.2. Nodes’ degree follows the power law distribution, as there are many nodes with low degree and fewer with high.

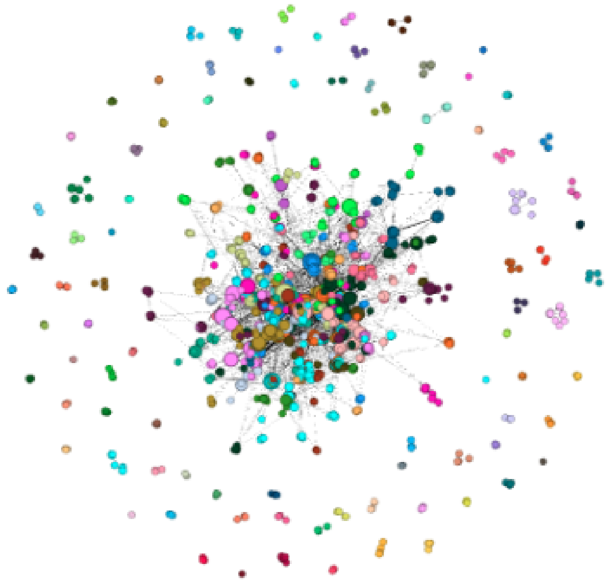
Fig. 2. Nodes’ Degree Distribution



### B. WHICH ACTORS WORK ISOLATED FROM HOLLYWOOD?

Another question we answer through graph metrics is which actors are far away from the center of our graph, that is translated into which of them work isolated from the Hollywood. We are based on closeness centrality to give the answer, by running the ”Diameter” from Gephi’s Network Overview. Then, we run OpenOrd Layout, and based on this visualization we can easily observe that are numerous groups (more specifically cliques, where all nodes are connected with each other) not connected to the main component of the graph. The graph is represented in Fig.3. If we look closer, these groups represent movies with non-American actors, isolated from the center. One worth mentioning example, is the existence of ”The Lobster”, a movie with greek cast (Fig.4).

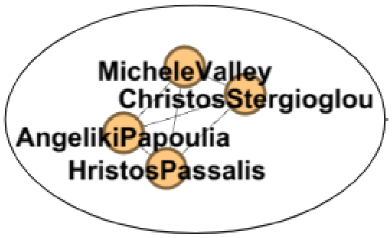
Fig. 3. OpenOrd Layout after running ”Diameter”



### C. WHICH ACTORS ARE IMPORTANT FOR THE COMMUNITIES’ COHESION?

Moreover, another question we answer and is quite important for our further research, is which actors are those who keep groups together. In other words which actors are those with the highest betweenness centrality degree. The meaning of this will be understood later, as some future collaborations wouldn't be suggested if these actors didn't exist. To measure actors' betweenness centrality we run "Diameter" from Gephi's Network Overview. The result is that the top 3 actors which are useful for groups' cohesion are:

Fig. 4. Cast-clique from movie "Lobster"



1) Mark Wahlberg (0.026)
2) Hugh Jackman (0.025)
3) Christian Bale (0.023)

### *D. HOW SUCCESSFUL ARE ACTORS' COLLABORATIONS?*

In this section, we want to examine how efficient actors' collaborations are. There are actors that tend to collaborate more than one time with each other and we examine if these collaborations lead to highly or lowly rated movies. In order to locate more-than-one time collaborations, we use Gephi's filter "Edge Weight", for edges with weight over 1. After filtering, there are 404 connected actors, which means 202 collaborations, as seen in Fig.5. Generally, we observe that collaborations follow power law distribution, since there are many nodes with a few collaborations and fewer nodes with many. Also, the maximum collaboration number is four. Fig.6 depicts the distribution of collaborations that carried out one, two, three and four times.

The next step is to split our initial dataset into highly rated movies ($>7$) and lowly rated movies ($= < 7$). So, we create two new graphs, the first one with the collaborations ($>1$) for highly rated movies and the second one with the collaborations ($>1$) for lowly rated. First, we present our observations in lowly rated graph, the representation of which is depicted in Fig.7.

Fig. 5. Collaborations ($>1$) in initial network

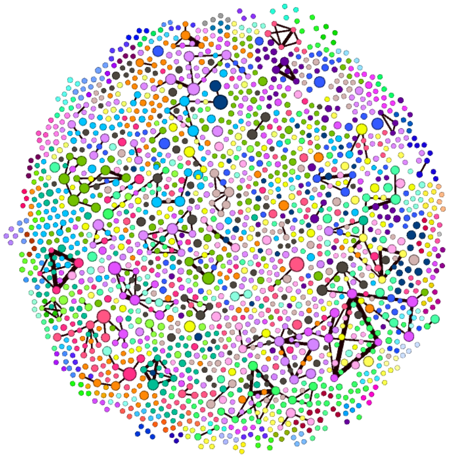
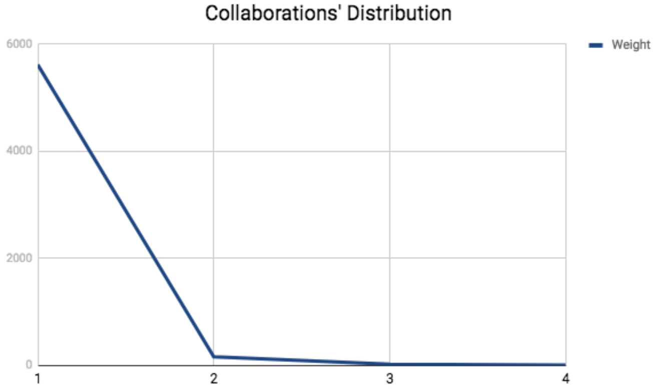


Fig. 6. Collaborations' Distribution



In this graph, we locate 103 collaborations, which are the 50% of initial graph's collaborations. The cliques here represent movie sequels. Thus, 42% of these collaborations are sequels. Some of them are The Twilight Saga, Horrible Bosses, Transformers, Alice in Wonderland, Sex and the City etc. The genres of sequels found in lowly rated movies are adventure, comedy, action, sci-fi, romance, drama. In addition, some worth mentioning bad collaboration pairs are:

1) Marion Cotillard - Michael Fassbender
2) Jennifer Aniston - Jason Sudeikis
3) Natalie Portman - Chris Hemsworth

Furthermore, we examine how collaborations are formed in highly rated movies. First of all in this graph there are 60 collaborations (many fewer than the lowly ones) which constitute the 30% of the initial graph. At this point we need to say that the rest 20% of the collaborations are movies that one time the collaborations lead to highly rated movie and the other to lowly. That's why these

Fig. 7. Collaborations (>1) in lowly rated movies

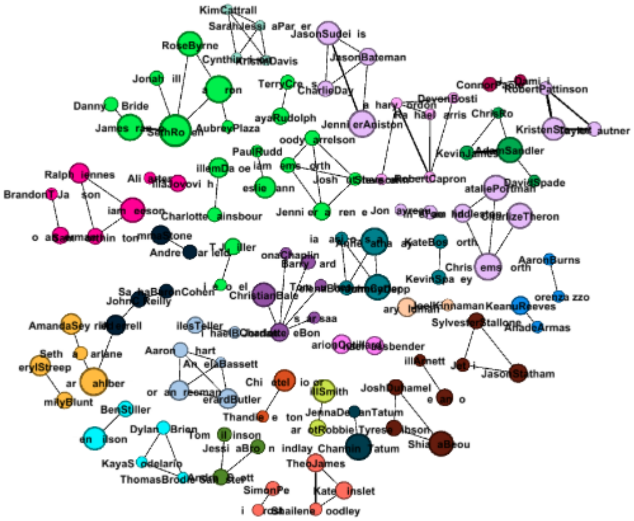
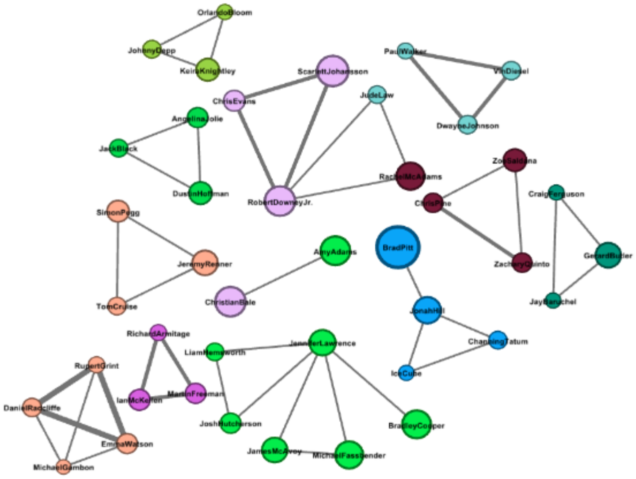


Fig. 8. Collaborations (>1) in highly rated movies



collaborations are absent from our graphs. One obvious observation that derives from this graph is that the biggest percentage of collaborations consists of movies that are sequels. We can see collaborations in movies such as Harry Potter, Hunger Games, Pirates of the Caribbean, Avengers etc. Just 35% of them are collaborations of actors(specifically pairs) that collaborate in movies that aren't sequel. This shows very good chemistry between these pairs, not depending only on the movie. Pairs like these are:

1) Jennifer Lawrence - Bradley Cooper
2) Amy Adams - Christian Bale
3) Brad Pitt - Jonah Hill

Finally, another observation has to do with the movies' genre. The genres of movies in highly rated movies are adventure and fantasy.

The answer to our initial question is that 50% of the cases where actors collaborate more than once, end up in movies with low rating. Apart from that, we can easily conclude that sequels tend to result in highly rated movies. Especially, when their genre is fantasy/adventure. Movies that create a fantastic world with fictional characters are quite high in audience's liking. Besides, sequels of comedy, drama, romance don't have the same reflection and are more likely to have low ratings.

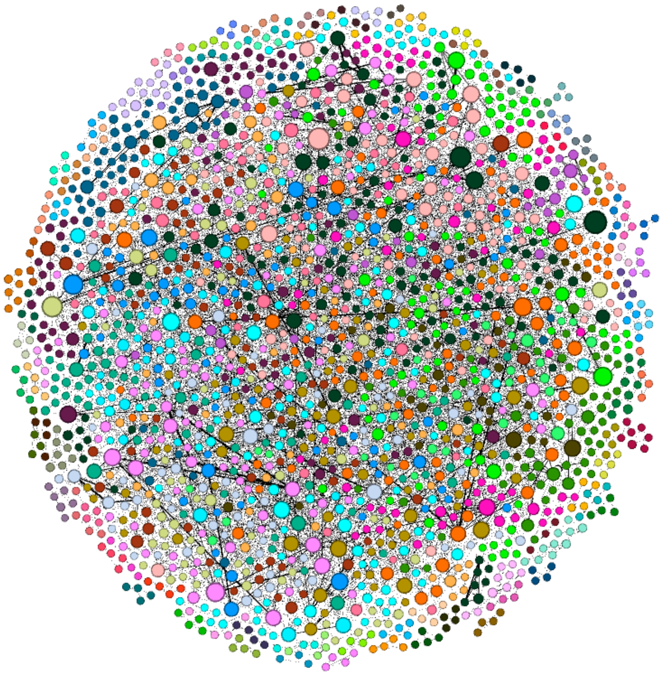
### *E. WHICH COLLABORATIONS WOULD WE SUGGEST?*

In this section we tried to answer a question that has a great importance for the cinema space, but there has been little research on this in the literature. The question regards to which collaborations that have not been done so far would be the most appropriate to be carried out in the future. For this question we need to examine actors by communities. From Gephi's statistics, we run Modularity with the default projection value of 1.0. This gives us a result of 172 communities. The biggest ones (22 groups) include more than 1% of network's nodes, since the rest have less than that (most of them are cliques). Next we run OpenOrd and Noverlap layouts for a better visualization as one can see in Fig.9. Same colors represent nodes in the same community.

The next step is to catch out the proper pairs in order to suggest a collaboration between these actors. The way we choose to do the recommendations is by detecting actors that are in the same community but have never collaborated in the past. Within the communities not all nodes are connected with each other as seen in Fig.10. This means that we can recommend nodes that there is no edge connected them, but they have the same color (since they join in the same community).

We export the nodes table from Data Laboratory to a csv file and we keep the Label, Modularity Class and Degree columns. From this file, we keep

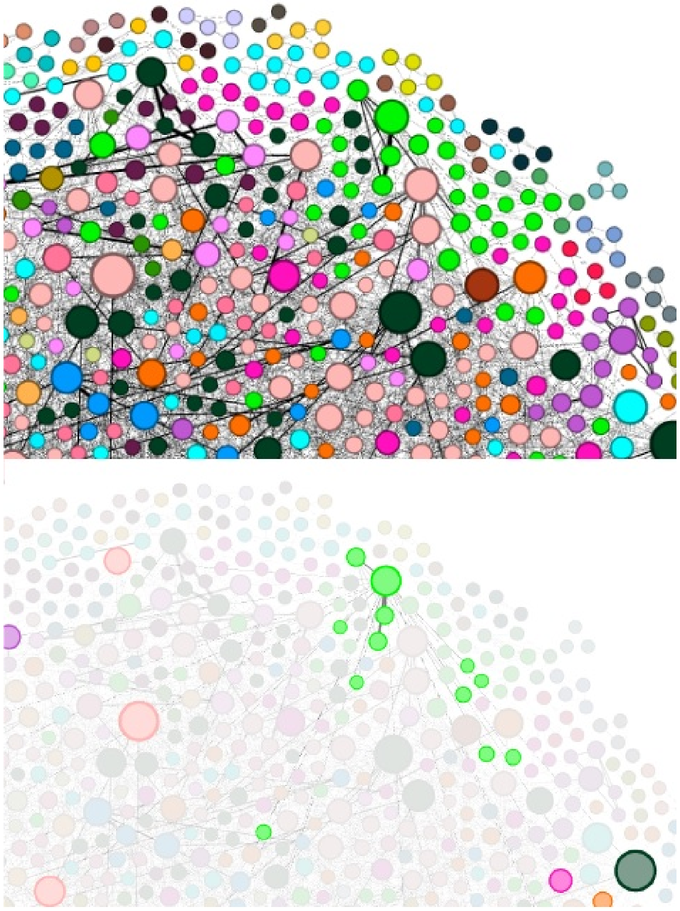
Fig. 9. Actor communities in the network



only the communities that include >1% of nodes from the graph. Next up, we keep the actors with degree >10, since actors with numerous collaborations would not easily collaborate with those with 1 or 2 experiences in their career. After this process, we end up with 285 actors overall. We use python code to manage the clusters and movies csv files. We create two lists, the first one has each row representing a cluster and containing a list with all the actors in this cluster and the second one with each row representing a movie and containing a list with the actors who participated in it. Following up, we produce all combinations in pairs for the actors that are in the same cluster and, respectively, for the actors in the same movie. Finally, we check if a pair of actors in the same cluster already exists as a pair in the same movie. If it doesn't, then this is a recommendation for collaboration between those two. Our data of 285 actors and 22 clusters result in 1965 new collaboration pairs. Some of the most feature collaboration pairs we distinguished are:

1) Chris Hemsworth & Charlize Theron
2) Brad Pitt & Gerald Butler
3) Jake Gyllenhaal & Matthew McConaughey
4) Ryan Reynols & Jennifer Aniston

Fig. 10. Actors that connect with nodes out of their community



5) Anne Hathaway & Keira Knightley

## V. CONCLUSIONS

In conclusion, we can punctuate that despite the large size of our network, there are few nodes (actors) who have many collaborations, are important for network's cohesion and seem to influence it more. This seems to be verified by the fact that our network follows power law distribution, where there are few nodes with many edges and many nodes with few edges. Moreover, most of the actors participate in Hollywood movies, while the nodes who are isolated from it form cliques of non-American actors. Next up, when examining actors' collaborations, we notice that there are only a few of them with >1 collaborations, regarding the size of initial network. Small percentage of these collaborations end up in highly rated movies, while most of them are sequels. Finally, when we discover communities in the network, we see that only 22 out of 172 groups are relatively big.

## VI. FUTURE EXTENSIONS

In this section we suggest some future extensions that could be applied for better and more accurate results. The future work regards to the research on future collaborations and more specifically to the parameters that can be taken into account. In our research we are based on communities and the extra collaborations that result through them. In our study, 1965 recommended pairs came out, which is a quite large number. We have not taken under consideration actors' tastes in some particular movie genres. So, in a future work, we can produce pairs that tend to prefer the same genre and have never participated in the same movie before. As a result, the recommendations are going to be more targeted and realistic.