IMDB Top-1000 Movies Social Network Analysis

Tariq Ahmed

161134893

Contribution: 25%

Research, data processing, algorithm coding and results analysis. Report writing

M Rizwan Habib

200683421

Contribution:25%

Research, data processing, algorithm coding and results analysis. Poster preparation and Report writing

Cyril D'Cruz

200405351

Contribution:25%

Research, results analysis, and related work. Report writing.

Harry Agyemang

200112345

Contribution:25%

Research, results analysis, and related work. Report writing.

Abstract- This paper explores network analysis of actors' social network. The aim of this paper is to analyse a dataset of movies to unearth communities of actors and recommend actors for a film role that could enter the IDMB top 1000. The network was created using the IDMB top 1000 movies which has films dating back to 1920 at the beginning of motion pictures with sound. As some features were incomplete, using K- nearest neighbour and mode imputation we were able to replace them. Using the Louvain algorithm for community detection and different centrality measures, we analysed the dataset. This was needed to establish if the network followed the small world phenomenon, actors' communities with the dataset, identify influential and central actors. We concluded a list of actors that should be casted in film as they have the best chance of being in a film that would enter the top 1000. We were also able to establish the network is a small world network based on high clustering coefficient and low average path length.

1. Introduction

The global film industry is gigantic market which is estimated to be $136 billion in 2018. In 2019 792 films were realised with the average cost of major studio movie being around $65 million. As such the film industry is both competitive and saturated market with a high barrier of entry. However, when looking at the apex of Hollywood, we see that highest actors are paid around 20-30 million dollars and the highest grossing film was "Avatar" with worldwide gross of $2,846,089,541. As such, there is great incentive for both actors and movie producer to produce a great film that can bring in large audience. While reviews and actor performance are subjective depending on many different factors such as storyline, character development, cinematography, special effects, cast, duration, and pacing. In this report, we mainly aim to focus on both the casting of actors for actor recommendation and creating and analysing a network from a "respectable" and "reputable" dataset. We will achieve this by using the IDMB top 1000 database which contains the highest meta score rated films and spans decades from 1920 to 2020. By mapping the data, we converted a Bipartite graph to One mode weighted projection. Using Gephi and Network X, we applied the Louvain algorithm and different Centrality statistics for community detection and to detect communities and to draw conclusions about who we should cast in a film.

2. Related Work

In the paper titled "An Empirical Study on IMDB and its Communities Based on the Network of Co-Reviewers" researchers M.Fatemi and L.Tokarchuk analysed a dataset of almost 22k movies and 270k reviews for these movies. They used movies as nodes. Nodes were linked if reviewer commented on two movies. Weights were added by number of common reviewers between a movie pair. It was shown that the network is scale free and follows a power law distribution. Furthermore, researchers then applied four different community detection algorithms on this dataset and showed advantages and comparisons of all four algorithms. It was established that Conclude fared better in overall performance of community detection. Finally, the findings of community detection were used to propose a movie recommender system. It was recommended that understanding of community compositions can help to create a better recommender system based upon a cluster of community. This shall also help to

overcome shortcomings of other social networks recommender systems that often suffer from cold start and sparsity of data [1].

In another paper published by Cattani G. and Ferriani S. in IEEE researchers discussed the position of a node is in the network can be used to define its influence within the network. As such much value is put on the different metrics of centrality and their meaning to the complete network. The paper investigates the roles of social networks in individuals and the ability to generate outcomes. Reviewing the Hollywood industry over the period of 1992-2003. For individuals who are in a position which is between the core and periphery their social system is in a favourable position can have a big effect on achieving creative results. It believes that these social networks affect individual creativity. For individuals who are at the core of their social field are provided with a greater exposure as they have access to the relevant sources and support. [2]

Another recent study presents a way of analysing and performing a movie analysis. It goes into depth with roles and their social networks to identify the embedded community structure in movies. The paper elaborates on detecting story lines and using existing methods to achieve that. it proposes in the future to investigate another generalised algorithm that can be applied to all types of movies and not the specific ones chosen in this experiment. It detects relationships between roles and constructs a role's social network based on the label scenes. The social network analysis helps to elaborate and identify leading roles and hidden communities. Using the results from community identification, an experiment is performed using storyline detection that is more flexible for movie analysis. This method proposed shows community identification is highly accurate and robust [3].

A well-known technique used when analysing social network is the transforming the network into a bipartite network. One study that investigates the analysis of Bipartite networks in movie ratings and catalogue networks. The paper starts off explaining networks and what they represent, node degree and degree distribution, lastly bipartite networks. These have two types of nodes vertices u and vertices m called partite sets and the edges lie between the nodes. The scenario reviews Netflix movies and collects a dataset relating to the movie and their rating helping with the task of predicting the rating for their customers and suggesting movie recommendation. suggesting the Netflix dataset as a bipartite network with the uses and movies as node types and the weights of edges are represented as ratings. The objective is to see how the structure of networks change because of the database being large and taking longer to compute results. For a projected network, two users can be connected if they rated the same movie [4].

3. Dataset

The research in this report is based on a IMDB movies dataset. This dataset has been acquired from Kaggle. There are no ethical or privacy issues that restrict the use of this dataset for this research.

The database is a csv format file. It contains 16 features and 1000 observations. In its original form it contained a few missing values. Namely, the feature “Gross” which contains gross revenue of movies and “Certificate” had missing values between both of them. Subsequently, KNN was used for Gross revenue and Mode for Certificate imputation.

Following is a brief description of dataset headers and data types. This is the dataset in its final form after pre-processing.

Header	Values	Type	Description
IMDB_Rating	1000	float64	Aggregate votes based on voter ratings
Meta_score	1000	float64	Weighted average of ratings given by critics
No_of_Votes	1000	int64	Number of votes by registered IMDB users
Gross	1000	float64	Gross revenue of a film
Poster_Link	1000	object	URL link of movie poster
Series_Title	1000	object	Name of movie
Released_Year	1000	object	Year movie was released
Certificate	1000	object	Age restricted movie certificate
Runtime	1000	object	Total duration of movie
Genre	1000	object	Movie genre type
Overview	1000	object	Brief plot introduction
Director	1000	object	Name of movie director
Star1	1000	object	Main movie character
Star2	1000	object	Second movie character
Star3	1000	object	Third character of movie
Star4	1000	object	Fourth movie character

4. Methodology

We have used both Python and Gephi in this research. Following is a break down for some of the methodologies we have used in this report.

1. Pre-processing

We have used Python to impute missing values. Specifically, we used KNN and Mode imputation to remove missing values. KNN imputes values by considering the values of nearest neighbours of missing value whereas Mode replaces missing value with most frequently occurring value in a feature.

2. Bipartite Graph

We used Bipartite graph of Actors and Movies in this research. In bipartite graph edges occur between two node groups, not within those groups. For example, Consumers and their purchases

3. One Mode Conversion

We converted Bipartite graph to one mode weighted network for analysis. This helped us to have two separate networks. One with movies as nodes and other with actors as nodes. Mathematically if "B" is bipartite network and P` is a one mode network we can write this conversion as.

$$P` = B \times B^{T}$$

4. Small world phenomenon

It is often interesting to establish if a network is small world as a small world exhibits properties of real-world network. It is similar to Gaussian distribution of statistics. The three important metrics for this determination are average path, clustering, and diameter. Further discussion on this is covered in results section of this report.

5. Community detection

Social networks often cluster together to form communities. These community may have common properties that are valuable for analysis. There are various algorithms that are at our disposal for community detection. Here, we have detected communities using Louvain method.

6. Pearson Correlation

We have used Pearson correlation test for Bivariate analysis to determine if different features are related with each other.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

7. NLP

We used NLP to download subtitles and draw word cloud from it.

5. Results

The dataset we obtained consisted of vast number of features. Movie titles in this dataset were present as one feature whereas actors were spread over four different columns. As in a film there is more than one actor involved. Hence, to keep this spread of variables we opted for a **Bipartite network** [5].

A Bipartite network is a graph whose vertices can be split to form two disjoint independent sets U and V. The split occurs in such a way that every edge connects a vertex in U to one in V.

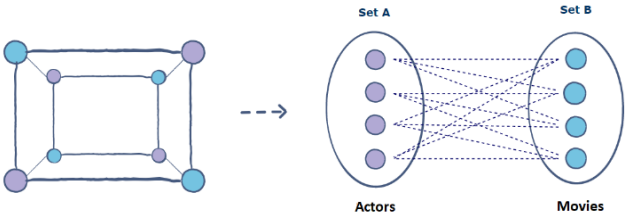


Figure 1 - A depiction of Bipartite graph [6]

The use of bipartite network can help us to split bipartite in two separate one mode networks. This will help us to run two separate network analysis. In one of the two one mode graph we used actors as nodes and on the other second one mode graph we used movies as nodes. As such, we can create network of movies based on common actors or create a network of actors based on common movies. We chose to report the network statistics using "actors" (features named as stars in dataset) as nodes. However, for community detection we have used both actors and movies as nodes turn by turn to obtain results.

Only in community detection we have used both actors as nodes and movies as nodes turn by turn to obtain results. Their reasoning is further discussed in community detection part.

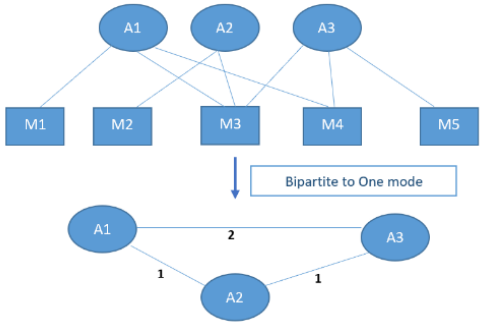


Figure 2 - One Mode conversion. "M" nodes represent Movies whereas "A" nodes represent "Stars or Actors".

One mode network Actors as nodes basic information:

Number of nodes	2706
Number of edges	5827
Average degree	4.30
Density	0.002
Clustering coefficient	0.84*
Connected components	193
Shortest path length in largest connected component	6.60
Diameter in the largest connected component	22

Nodes and Edges

We used actors as nodes, and they are 2706 in total whereas edges formed are 5827.

Average Degree

This measures connectedness of a graph. The average degree for this dataset is 4.30. In simple terms this is the average number of nodes connected to any node is graph.

Density

From the results of density, we can see that this network is sparsely connected as the value of density is very low. Values near 1 indicate dense structure.

Clustering Coefficient

In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. Values close to 1 indicate that nodes are tightly connected in clusters.

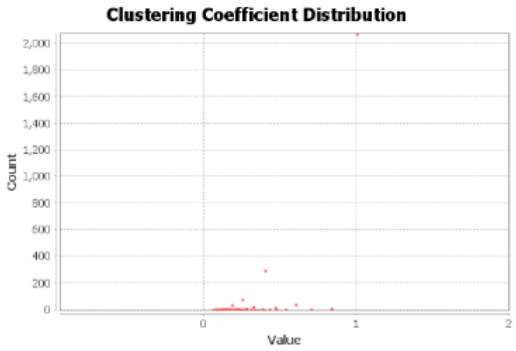


Figure 3 - Clustering coefficient Gephi

The average clustering coefficient we obtained is 0.84*. This is indicative that nodes are strongly clustered together. Figure 3 shows that the clustering coefficient is evenly spread with most nodes have a value closer to 0 than 1. This indicate that there are many weak links within our network which should be expected due to shape of our network which has centre cluster and periphery of actors which many have no link to the centre cluster. Also, as our graph dates back to 1920's we have some actors who were relevant to specific era of film such as Charlie Chaplin who is unlikely to linked to many actors as he retired with be last casted in a film in 1967.However, as we have some nodes with high clustering coefficient this has likely skewed the average network clustering coefficient of Gephi. A more likely network clustering coefficient is around 0.5 as we see in the distribution in figure 3.

Degree centrality

Degree centrality measures how many node ties in a network. In our exercise we obtained a list of top nodes (actors) based on their degree centrality. Following is a list of Top 5 actors with highest degree measure.

Robert De Niro	45
Tom Hanks	38
Brad Pitt	36
Al Pacino	35
Clint Eastwood	33

Figure 4 - Top 5 actors based on degree centrality.

Betweenness centrality

The goal of betweenness centrality is to lookout for node that is acting as a bridge in a network.

Dev Patel	*0.065*
Saurabh Shukla	*0.064*
Ranbir Kapoor	*0.063*

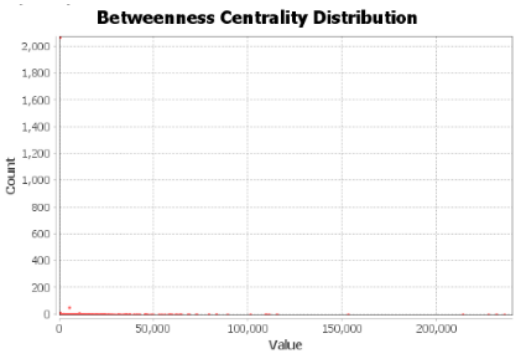


Figure 5 - Betweenness centrality.

Dev Patel has the highest betweenness centrality which means that Dev Patel is connected with multiple groups in this network belonging to different genres or audiences. Figure 5 shows that there is a spread of betweenness values with most falling between 0- 50,000 in value and a count of less than 200. This is because many actors only had 3 connected edges which could be films that are produced independently or high rated films with actors that only had one large role that gained success.

Closeness centrality

The sum of shortest path distances of a node from other nodes. It is the measurement of node's capacity to affect all other elements in network. Figure 6 shows that evenly spread graph with most actors having a count less than 50 and density packed value sections. This means due to the network shape, where an actor in the centre cluster is likely to have worked in Hollywood and the periphery are actors who films are likely to have other audiences such Japan and India.

Robert De Niro	0.148
Christian Bale	0.147
Christopher Plummer	0.147

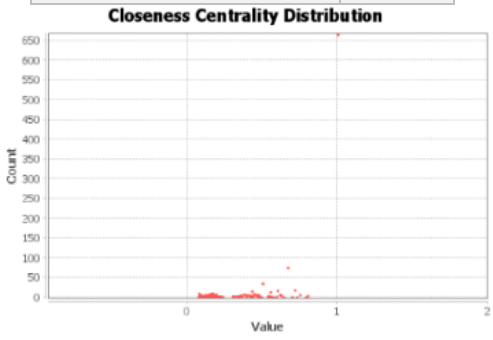


Figure 6 - Closeness centrality and top 3 actors in ascending order

Eigenvector Centrality

Eigenvector centrality is a measure of the influence of a node in a network. It is the measure of extent to which a node is connected to influential other nodes.

Brad Pitt	0.26

Matt Damon	0.22
Leonardo DiCaprio	0.22

Figure 7 - Top 3 actors based on eigenvector centrality.

6.1 Small World Phenomenon

Social network graphs are expressed in terms of their degree distribution, average path length, clustering coefficient and other metrics. Together this information can be utilized to establish that whether a network is following small world, scale free or random phenomenon.

Small world networks resemble real world graphs, and they are often associated with six degree of separation experiment. Often researchers, try to establish if a social network is small world or not. A network is a small world if most of the nodes who are not neighbours can be reached from every other node by a small number of steps [3]. Small world networks exhibit properties such as small diameter, high clustering coefficient and short average path length. For IMDB data clustering coefficient = 0.84, average path length is = 6.60 and diameter = 22.

From above we can establish that this network is following a small work phenomenon as most of the nodes can be reached from under 7 hops and high clustering coefficient depicts tight knit groups. It is expected as group of actors collaborating would like to work in their favourite genre. For example, Robert De Niro and Al Pacino like to work in Action movies.

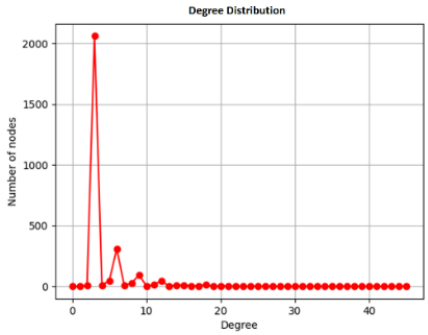


Figure 8 - IMDB Dataset degree distribution

6.2 Community Detection

Community detection is important aspect of social networks and graph theory. It has been seen that nodes who are in same community tend to have similar behaviour and properties. In fact, community detection is being used more often in building recommender systems. It can also be said that community detection in network theory is somewhat similar to what we do in unsupervised learning such as clustering.

There are number of algorithms that can be used to detect communities. Below are most used five community detection algorithms:

1. Girvan-Newman
2. Conclude
3. Lancichinetti
4. Louvain
5. Palla

Whilst carrying out community detection we framed two problems.

i. Looking for pair of actors who appeared together in movies.
ii. Most frequent actors appeared in movies dataset. In other words, counting the number of movies they did individually.

To undertake these two tasks, we split community detection in two parts. First, actors were used as nodes to obtain results for part (i) above and after words we used Movies as nodes to obtain results for part (ii).

In both instances we used **Louvain method**.

Louvain find communities by local optimization of modularity. It then aggregates nodes of the same community and builds a new network out of communities. These steps are repeated iteratively until the maximum modularity is attained [1].

6.2.1 Community Detection with Actors as nodes

In our first analysis we detected 213 communities with modularity score for best partition at 0.85.

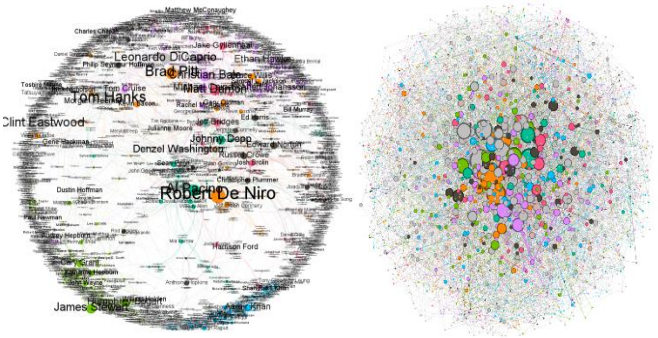


Figure 9 - Community visualisation using Force Atlas-2 (left) and Fruchterman Reingold (right)

We were able to find actors who are working together. When communities are portioned, we can clearly see from the weight of the edges how many times actors have appeared together in a movie. The first two pictures are for Hollywood movies and the third visualisation below show Indian actor communities. Each community contain more than 100 actors, but to get better visualisation only 20 actors are displayed below.

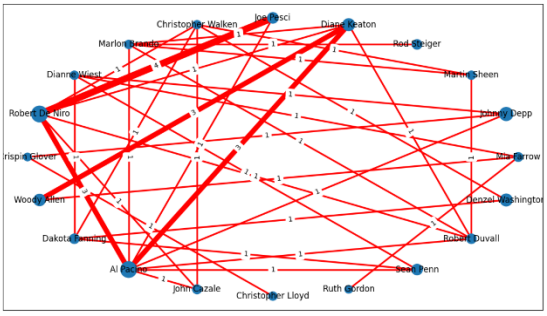


Figure 10 - Actor as nodes community-1

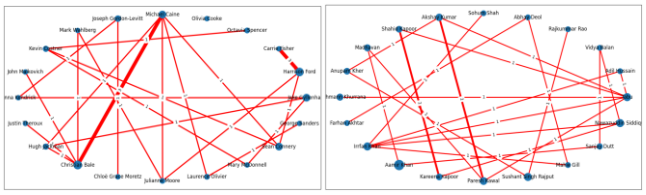


Figure 11 – Community-2 Hollywood (left) and community-3 Bollywood (Indian) actors community (right)

Community	Actor-1	Actor-2	Together
Community 1	Robert De Nero	Joe Pesci	4
Community 1	Robert De Nero	Al Pacino	3
Community 1	Woody Allen	Diane Keaton	3
Community 1	Al Pacino	Diane Keaton	3
Community 2	Christian Bale	Michael Caine	3
Community 2	Carrie Fisher	Harrison Ford	3
Community 3	Karena Kapoor	Shahid Kapoor	2
Community 3	Akshay Kumar	Paresh Rawal	2

6.2.2 Community Detection with Movies as nodes

In the second part of community detection, we used Movies as nodes. This enabled us to detect total appearances of actors in dataset and individual communities. Here, we again used Louvain. A summary statistic is as follows.

Number of nodes	1000
Number of edges	2798
Average degree	5.60
Clustering coefficient	0.52
Modularity	209
Total communities	209

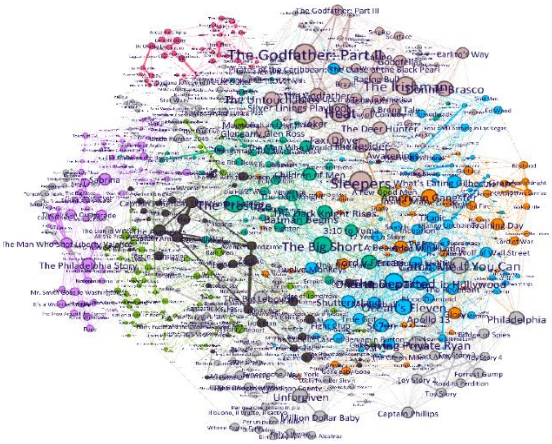
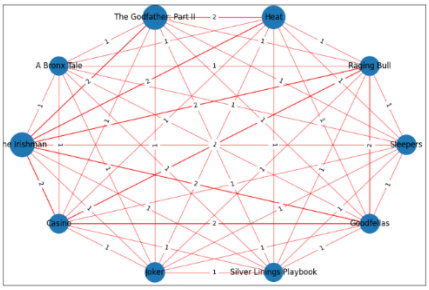


Figure 12 - Force atlas 2 community visualisation

Below we have lists of two communities in the dataset. We are displaying top 5 actors from these communities.



Community number	*1*
Movies in community	*60*
% of movies in network	*6.0%*
Robert De Niro	*17*
Al Pacino	*13*
Diane Keaton	*6*
Joe Pesci	*6*
Sean Penn	*6*
Woody Allen	

Community number	*2*
Movies in community	*75*
% of movies in network	*7.5%*
Christian Bale	*11*
Brad Pitt	*11*
Matt Damon	*10*
Michael Caine	*9*
Edward Norton	*7*
Jack Nicholson	*6*

It can be observed that Robert De Niro, Al Pacino, Christian Bale and Brad Pitt have highest number of movies in IMDB top 1000 movies database.

6.3 Further Statistical analysis

The dataset we used contain several features. We carried out further analysis on these features for further insights. The

visualization below depicts a time series graph of revenue generated by movies based upon their year of release.

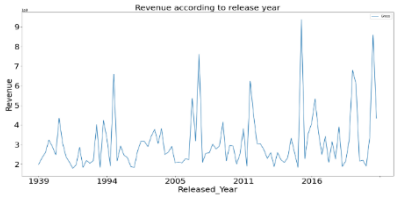


Figure 13 - Released year vs Revenue.

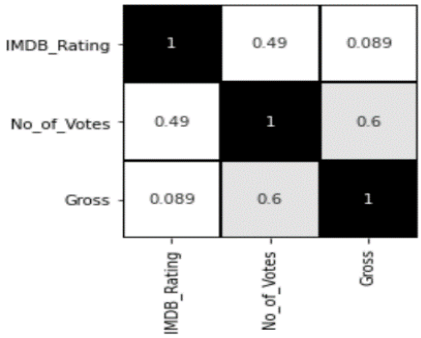
We could see that five years between 2015-20 have a big chunk top grossing movies.

6.4 Pearson Correlation

We further explored the possibility of correlation between two features. For this purpose, we used bivariate correlation matrix also known as Pearson correlation denoted by 'r'.

This gives us a score between +1 to -1. A score towards 1 represent a strong correlation with positive and negative denoting the direction of relationship and 0 indication no correlation.

We could only observe a moderate relationship between number of votes and Gross revenue at a score of "0.6". It generally translates that higher the number of votes there is higher probability of movie grossing more.



6.5 Natural Language Processing

Finally, we used NLP techniques to draw word cloud for highlighting prominent words spoken in a Genre.

We used Python's Subliminal framework to fetch subtitles of movies and then afterwards used word cloud to display them.

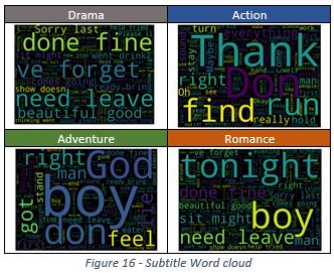


Figure 16 - Subtitle Word cloud

Above picture contain word cloud of only four genres.

6. Conclusion

We have successfully applied community detection to extract top actors who have frequently featured in this movie dataset. We have showed that Robert de Nero with 17 movies have the highest number of films in top 1000. We have also shown that the greatest number of collaborations between two actors are 4 and these were between Robert de Nero and Joe Pesci.

Using centrality as the basis of actor recommendation, we will use the highest eigenvector, betweenness and degree centrality to recommend actors. As discussed in the results, the actors with the highest centrality are Robert de Niro, Dev Patel, Brad Pitt, and Christian Bale. Therefore, we recommend that casting a film with these 4 actors in, is most likely to both attain a spot in the IDMB top 1000 movies and attract a larger audience of both India and western audience that are more likely to watch Hollywood films.

7. Further work and evaluation

Social network theory is a vast subject, and this dataset has given us an excellent opportunity to apply various aspects of network theory. However, there is still room for improvement and due to limited time and skills few applications were left open ended. It will be interesting to see further improvements such as:

1. Considering Directors as well as movies and actors.
2. Undertaking study of communities based on reviews.
3. Use of NLP to analyse script of movies. To check if they meet movie certificate. For example, to verify if kids' movies contain obscene words etc.
4. Analysing a dataset that contains more demographics as the IDMB dataset tends to have English speaking films on it list which does not fully consider other films in different languages like French and Spanish.
5. The dataset does not consider the cost of producing the film such as the marketing spent to promote the film to broader public.
6. Also using the complete dataset affecting the results of producing a film today or in the future as some films on the dataset may be considered obscene and indecent if shown to an audience in 1920. While some films maybe considered outdated and old fashion when seen by audience in 2021 or the future.

References

[1] L. T. Maryam Fatemi, “An Empirical Study on IMDb and its Communities Based on the”.

[2] G. a. F. S. Cattani, “A Core/Periphery Perspective on Individual Creative Performance: Social Networks and Cinematic Achievements in the Hollywood Film Industry.,” *Organization Science,* vol. 19, no. 6, pp. 824-844, 2008.

[3] C. C. W. a. W. J. Weng, “Movie Analysis Based on Roles’ Social Network,” in *IEEE*, Beijing, 2007.

[4] M. B. Díaz, “Analysis of a Bipartite Network of Movie Ratings and Catalogue Network Growth Models,” University of Oxford, Oxford, 2008.

[5] P. I. K. A. P. C. B. E. M. P. G. B. Georgios A Pavlopoulos, “Bipartite graphs in systems biology and medicine: a survey of methods and applications,” *Giga Science,* vol. 7, no. 4, 2018.

[6] E. Team, “Educative,” 04 04 2021. [Online]. Available: https://www.educative.io/edpresso/what-is-a-bipartite-graph.