

Project 2

EE232E - Graphs and Network Flows

Team:

Rutuja Ubale (UID: 404558257)

Usha Amrutha N (UID: 204590772)

Pallabi Chakraborty (UID: 404519609)

Question 1:

The following files were used for data analysis: actors_movies.txt and actress_movies.txt.

The actors and actresses list has been combined and the actors/actresses who worked in fewer than 5 movies were removed from the list. Also, the extra spaces in the movie names have been stripped and the data has been cleaned in Python. We have also added an extra parameter called actor ID which is based on the sequence of occurrence which can be used later for mapping. The total number of actors in the entire list after cleaning is 244296.

A screenshot of a part of the file is shown below:

Kapoor, Aditya Roy		69813
Kapoor, Akshay	69814	
Kapoor, Anil (IV)		69815
Kapoor, Annu	69816	
Kapoor, Arjun	69817	
Kapoor, Bharat	69818	
Kapoor, Bhupendra		69819
Kapoor, Col.	69820	
Kapoor, Gaurav	69821	
Kapoor, Gautam	69822	
Kapoor, Goga	69823	
Kapoor, Kamal	69824	
Kapoor, Karan (I)		69825
Kapoor, Kaushal	69826	
Kapoor, Kishore	69827	
Kapoor, Kunal (I)		69828
Kapoor, Kunal (II)		69829
Kapoor, Kunal (III)		69830
Kapoor, Lalit (II)		69831
Kapoor, Master Omkar		69832
Kapoor, Mohan	69833	
Kapoor, Pinchoo	69834	
Kapoor, Prithviraj		69835
Kapoor, R.P.	69836	
Kapoor, Raj (I)	69837	
Kapoor, Raj (II)		69838
Kapoor, Raj (III)		69839
Kapoor, Rajan	69840	
Kapoor, Rajat	69841	
Kapoor, Rajit	69842	
Kapoor, Rajiv	69843	
Kapoor, Ram	69844	
Kapoor, Ramesh	69845	
Kapoor, Ranbir	69846	
Kapoor, Randhir	69847	
Kapoor, Ranjan	69848	
Kapoor, Ravi (I)		69849
Kapoor, Ravindra		69850
Kapoor, Rishi (I)		69851
Kapoor, Rohan (I)		69852
Kapoor, Sanjay (I)		69853
Kapoor, Sashi	69854	
Kapoor, Satyendra		69855
Kapoor, Shahid	69856	
Kapoor, Shakti	69857	
Kapoor, Shammi	69858	
Kapoor, Sharad S.		69859
Kapoor, Shashi (I)		69860
Kapoor, Trilok (I)		69861
Kapoor, Tusshar	69862	
Kapoor, Vikram	69863	

Question 2:

A directed graph of actors was created using igraph. The edges have been calculated using

$$V = \{\text{all actors/actresses in the list}\}$$

$$Si = \{m | i \in V, m \text{ is a movie in which } i \text{ has acted}\}$$

$$E = \{(i, j) | i, j \in V, Si \cap Sj \neq \emptyset\}$$

The number of nodes in this graph is 244296 and the number of edges in the graph is 58026910.

Question 3:

The pagerank algorithm with damping factor = 0.85 was run on the weighted actors' graph and the names of the top 10 actors were determined. The top 10 actors having highest page ranks in order of their pageranks are as follows:

Sr.No	Actor	Page Rank	Comments
1.	Bess Flowers	0.0001580023	Known as the "The Queen of the Hollywood Extras". Her prime run in Hollywood was during the 1940s and 1950s.
2.	Fred Tatasciore	0.0001377731	Voice Actor and stand-up comedian. He is best known for voicing the Hulk.
3.	Steve (IX) Blum	0.0001340366	voice actor of anime, animation and video games known for his distinctive deep voice. He provides voice for the host of Cartoon Network.
4.	Sam (II) Harris	0.0001333384	Most of the time was an uncredited actor in Hollywood.
5.	Ron Jeremy	0.0001270027	Number one adult film star in the U.S during 1980s and 1990s.
6.	Harold (I) Miller	0.0001165205	Mostly an uncredited actor in Hollywood during 1930-60s.
7.	Yuri Lowenthal	0.0001090534	Actor and voice actor. Well-known for his work in voice over in video games and animation.
8.	Robin Atkin Downes	0.000104297	English screen and voice actor, who is known for his work in live action, animation and video games.
9.	Lee (I) Phelps	0.0001012679	He is near the top of Hollywood's most prolific but virtually unknown supporting players
10.	Jeffrey Sayre	0.0001006508	Another prolific but unknown supporting actor.

It can be seen that 5 out of the top 10 actors with highest page ranks are supporting actors who have been mostly uncredited for their roles. Since supporting actors work in a large number of films compared to the famous movie stars and thus have very high connections with other actors/actresses this could be possible. Also, Ron Jeremy (5th highest pagerank) is an adult movie star who worked in a large number of movies and this can be supported by the fact that male adult movie stars are lesser in number and tend to act in most of the movies made. Another observation here is that 5 of the highest page ranked actors were from the early or mid 20th century. This might have been because there were fewer number of actors then and they acted as extras in almost all the movies that were made. Well-known actors may not appear in the list because they may have had significant roles in a few movies, thus justifying their low page rank scores.

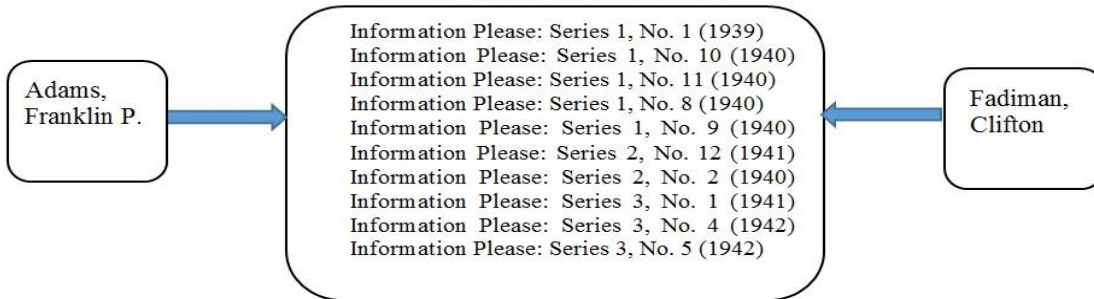
The top 10 famous celebrities in our opinion and their page rank values are listed below:

Sr.No	Actor	Page Rank
1.	Tom Cruise	1.172932e-05
2.	Natalie Portman	4.400885e-06
3.	George Clooney	2.195821e-06
4.	Kate Winslet	1.020989e-05
5.	Robert Downey Jr.	7.395434e-06
6.	Leonardo DiCaprio	1.960902e-06
7.	Brad Pitt	1.474613e-06
8.	Julia Roberts	2.101313e-06
9.	Johnny Depp	4.085536e-06
10.	Sandra Bullock	1.200901e-06

We note that the Page Ranks of the famous actors are significantly lesser than the highest Page Rank scores (these are probabilities of visiting the actor among 244296 actors. So, they are significantly small). This is because the famous actors mainly act in a lead role and thus tend to act in a fewer number of films compared to the supporting actors who work in majority of the movies as extras.

About 372 pairs of actors have both directed edges between them with a weight of 1 indicating that all their movies are common with the other actor. All such actors mainly act in a TV series/ movie series and they have only worked in that TV/ movie series (all seasons). In a smaller film industry (languages other than English which are spoken less, e.g: Bosnian and Herzegovinian) this is also quite possible where there are fewer number of actors. Another case which is possible is that both the actors might have worked in a few short films which are common to both of them. Examples of each type mentioned above are given below:

Radio Quiz Show Series



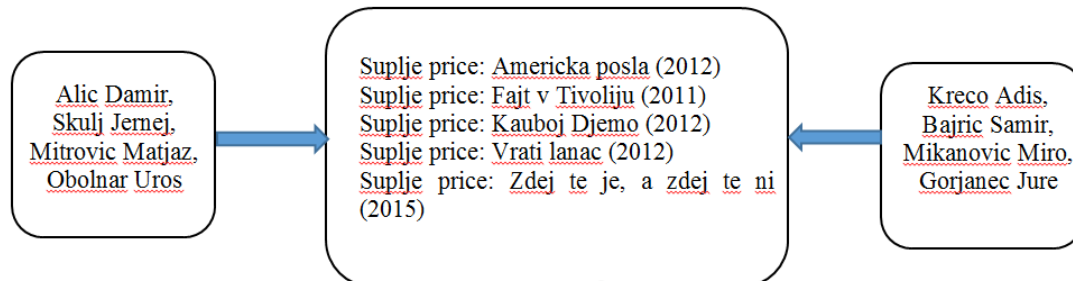
Both the actors were moderator and panelists respectively

Short Films



Both the actors worked in short films and worked together in all their films.

Bosnian and Herzegovinian Movie Series

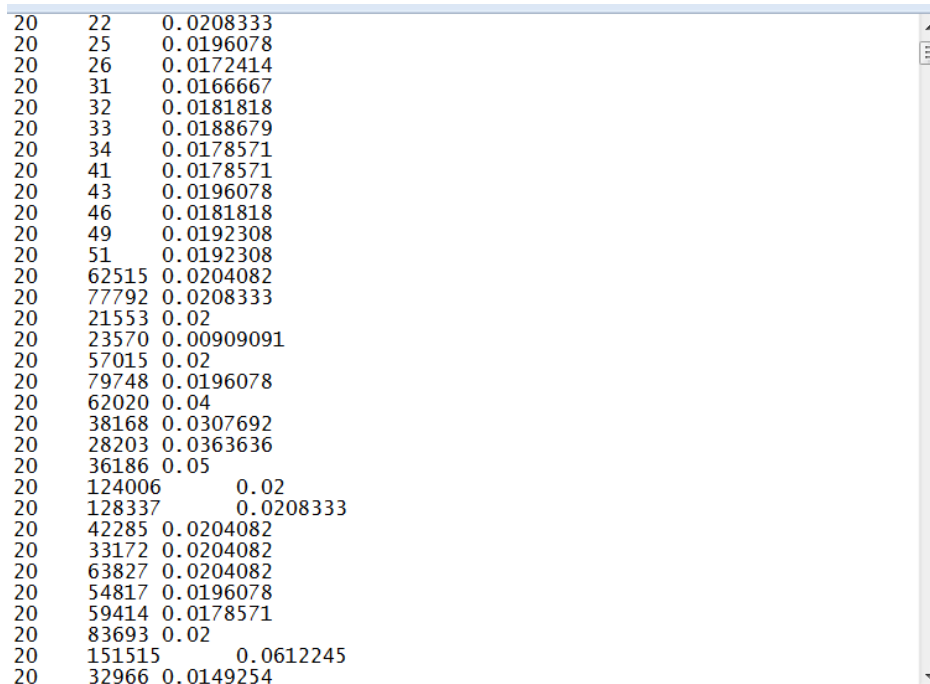


All the actors were characters in the series and this was the only series they worked in

Major surprises were that famous actors were not amongst the highest page ranked actors. This has already been explained above that the famous actors mainly act in lead roles and gain a star status and fame. Thus they tend to act in a lesser number of movies compared to the supporting actors who work in majority of the films as extras. So, their page ranks by using our method would be less than the supporting actors but would not definitely be the least as generally a many actors act in the films of these famous actors as the movies are huge. Hence, they lie in the middle in terms of page ranks.

Question 4.

The movie network was created based on the Jaccard index of two movies. Jaccard index was used as weights to create an undirected graph. This measure tells us the similarity and diversity of given sample sets based on their union and intersection values. The graph network thus created was undirected. The data cleaning was done in C++ in such a way that the movies which did not have the genre information listed were removed. Additionally, we ensured that no extra character spaces we included within/between movie names. Though the question asks for filtering movies for less than 5 actors, the processed file was too large to run Fast Greedy Newman Algorithm. So, we increased the threshold to 20. This increased the processing speed of the file for the following questions. A part of the file with movie ids and Jaccard indices is as shown below:



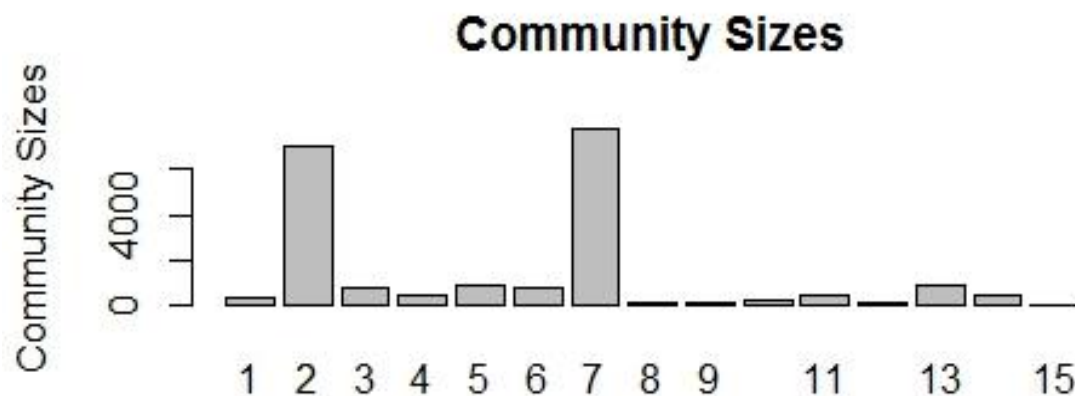
20	22	0.0208333
20	25	0.0196078
20	26	0.0172414
20	31	0.0166667
20	32	0.0181818
20	33	0.0188679
20	34	0.0178571
20	41	0.0178571
20	43	0.0196078
20	46	0.0181818
20	49	0.0192308
20	51	0.0192308
20	62515	0.0204082
20	77792	0.0208333
20	21553	0.02
20	23570	0.00909091
20	57015	0.02
20	79748	0.0196078
20	62020	0.04
20	38168	0.0307692
20	28203	0.0363636
20	36186	0.05
20	124006	0.02
20	128337	0.0208333
20	42285	0.0204082
20	33172	0.0204082
20	63827	0.0204082
20	54817	0.0196078
20	59414	0.0178571
20	83693	0.02
20	151515	0.0612245
20	32966	0.0149254

Column 1 and 2 represent the movie ids of movies in the network with corresponding Jaccard weights. Here Jaccard index is the number of common actors in given two movies to the total number of actors in both the movies. The code for cleaning is done in "Q4_clean.cpp" to

construct an undirected network. Each star accounts for an increment to the total number of common stars between movies. A table was constructed to display the connections between two movies because of the common stars. Since the two movies can have more than one actor in common, we note that the movie IDs repeat several times across the actor's column. This table was used to construct the undirected movie network.

Question 5

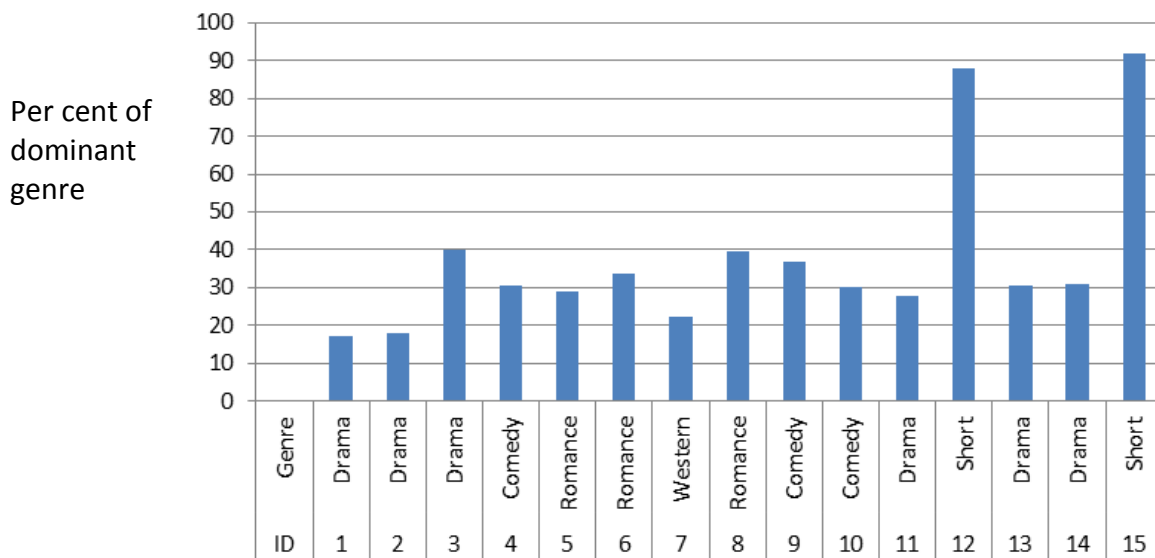
Since the movie network was truncated with threshold 20 for actor/actress, the movie network represents only a subset of communities that can be formed. In this case, number of communities was found to be 15 after running Fast Greedy Newman algorithm. The histogram shown



Modularity of this network is 0.22. Since, the threshold was to set 20 the network on which we perform Fast greedy and Community tagging is relatively small. The tagged communities for the network are as given below. 15 communities were tagged with genres that appeared in 20% or more of the movies in the community. There are as show below.

Community ID	Genre
1	Drama
2	Drama
3	Drama
4	Comedy
5	Romance
6	Romance

7	Western
8	Romance
9	Comedy
10	Comedy
11	Drama
12	Short
13	Drama
14	Drama
15	Short



We see from the histogram plot that the communities 2 and 7 have majority portion of the movies in the network. There for the percentage of the portion of majority community for 2 and 7 is less as compared to communities say, 12 or 15. This is because; from the histogram plot we can verify that 2 and 7 have a lot of movies, which could be from different genres. So it is difficult to find a genre which has major number of movies due to diversity. On the other hand, we spotted quite a few Indian movies, which usually belong Romantic/Drama in community 8 and 3. This could be the reason that 3 and 8 are tagged as Drama, and romance respectively.

Question 6.

The Top 5 neighbours of three given movies are marked by highest Jaccard indices (or weights). Also, the community to which each neighbour belongs to is also returned. A community map was created to map the movie neighbours to its corresponding communities.

The top 5 nearest neighbours of Batman v Superman: Dawn of Justice (2016)

Movie(Node) ID	Jaccard Index	Name	Community
3029	0.0600	Man of Steel (2013)	2
23122	0.0442	Secretariat (2010)	2
64017	0.0441	The Networker (2015)	2
103772	0.0434	Promised Land (2012))	2
28273	0.0432	Won't Back Down (2012)	2

The top 5 nearest neighbours of Mission: Impossible - Rogue Nation (2015)

Movie(Node) ID	Jaccard Index	Name	Community
67150	0.1428	Breaking the Bank (2014)	2
59842	0.1400	Cuban Fury (2014)	2
138465	0.1020	Mission: Impossible - Ghost Protocol (2011)	2
81948	0.9600	Kingsman: The Secret Service (2014)	2
67163	0.0816	Walking on Sunshine (2014)	2

The top 5 nearest neighbours of Minions (2015)

Movie(Node) ID	Jaccard Index	Name	Community
87443	0.2187	WALL·E (2008)	2
205183	0.2000	Tokyo Mater (2008)	1
87442	0.1780	The Emperor's New Groove (2000)	1
87427	0.1110	Brother Bear (2003)	2
163500	0.1110	Crash Tag Team Racing (2005)	2

Rogue Nation (2015) and Minions (2015) belong to community 2. Therefore, all the neighbours belong to community 2. This can be verified from the tables illustrated above. Also, it is noted that the neighbours need not necessarily belong to same genre. For instance, Batman vs Superman: Dawn of Justice (2016) is a sci-fi movie while Secretariat belongs to Family drama. This can be attributed to the fact that network was created based on the actors/actress, and not on the movie genre. Also, the edges have been weighted based on the number of common actors. However, it is interesting to note that for Minions (2015), and neighbours all belong to animated movies.

Question 7:

We downloaded the ratings list and derived a function to predict the ratings of the following three movies using the movie network:

Batman vs Superman: Dawn of Justice (2015)

Mission Impossible – Rogue Nation (2015)

Minions (2015)

We have used ratings of neighbour movies and movies in the same community.

The function we used for this is the weighted average of the ratings of the nearest neighbours. We decide the nearest neighbours on the basis of which neighbours have the highest edge weights with the target movie. 10 such neighbours have been selected to use as predictors for the regression. The ratings we get are as follows:

Movies	Predicted Ratings
Mission Impossible: Rogue Nation (2015)	7.4
Minions (2015)	6.3
Batman vs Superman: Dawn of Justice (2016)	7.1

Question 8:

In this question, we train a regression model to predict the ratings of the same three movies as above. The features we use are as follows:

A) Top 5 pageranks of the actors (five floating point main values) in each movie

B) if the director is one of the top 100 directors or not (101 boolean values). These are the directors of the top 100 movies from the IMDb top 250.

This would mean that we will use the top five page ranks of the actors in the movies and a boolean vector of length 101 where the value of 1 would indicate that the director has directed that particular movie. If not, the value will be 0.

The ratings we predicted are as follows:

Movies	Predicted Ratings
Mission Impossible: Rogue Nation (2015)	7.7
Minions (2015)	6.5
Batman vs Superman: Dawn of Justice (2016)	7.6

The R squared value we got is **0.9497** which is very high. R-squared tells us what proportion of the dependent variable has been explained by the independent variable.

R-squared = Explained variation / Total variation

94.9% is a very high percentage, which means that the independent variables we chose are doing a good job of explaining the dependent variable

Question 9

In this question, we used a different approach to predict movie ratings. We construct a bipartite graph with actors and actresses representing the vertices of one part and movies representing the vertices of the other part. An actor/ actress are connected to all the movies he or she has been a part of. A bipartite graph, is essentially decomposed into two disjoint sets such that no two within the same set are adjacent

The following steps have been used to predict the movie ratings:

1. Firstly, we assign score to each of them mapping every movie to corresponding actor as given in the movie_actor.txt. The calculation of scores is as given below:

$$\text{Score of an actor} = \text{Ratings of all his movies} / \text{Total number of movies.}$$

2. The bipartite graph can be completely established when we use the information presented in previous step such a way that for each movie, we can calculate the average rating of all its actors present in the movies.

In order to save on the computation time we can choose only top-N actors for a given movie. The idea behind our approach is that since the actors have enacted in quite a few movies throughout their career, we can use their popularity measure (here the average ratings of their movies in the past) to predict whether or not a given movie will have high rating. The regression model was fit to perform the same and the results of aforementioned movies are as follows:

Movie	Original	Predicted Rating	Absolute Error Rate
Batman vs Superman: Dawn of Justice (2016)	7.1	6.192	12.7%
Minions (2015)	6.4	6.507	1.67 %
Mission: Impossible - Rogue Nation (2015)	7.5	6.141	18 %

The predicted movie ratings were fairly close to the original. However the slightly high error rate in Batman and Mission impossible is due to the fact that these movies had featured actors who are fairly new to the career in film industry. Hence, their average ratings may not be consistent in this computation. However, despite this fact, the lead roles have incredible success in Film industry (say, Tom Cruise). It is because of the ratings of these lead actors the

results are fairly accurate. A more robust model can be designed with some more interesting metrics such as the popularity of **an actor over time** to see if his/her movie will have high rating in the upcoming releases. The idea is to take a **moving/running average** sort of method to check the recent popularity of a famous star. In some cases, it may happen that the popularity of an actor decreases over time, or he/she may not be active anymore. In such a case, movie featuring the actor need not necessarily have a high predicted rating.

Google drive link for processed text files as given below:

<https://drive.google.com/open?id=0Bzjewn0CM3GLUXJ4b21paVIVdVk>