ECE232 Project4
IMDb Mining

Juo-Wen Fu 805025223
Alex Hung 705035114
Te-Yuan Liu 805027255
Wesly Tzuo 704946416

June 5, 2018

1 Actor/Actress network

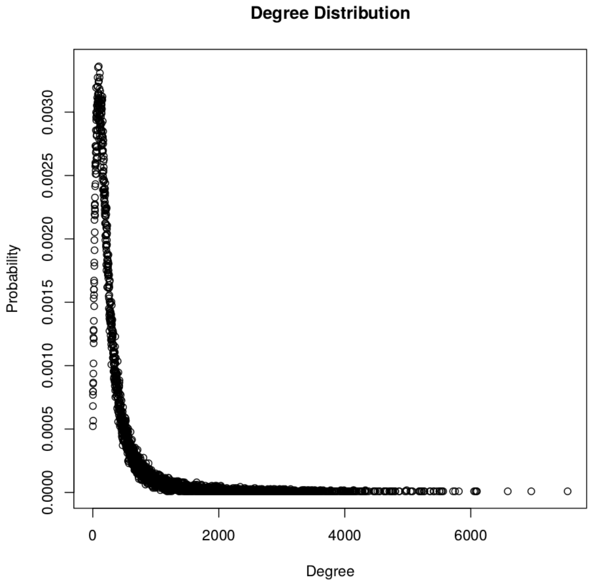
Question 1: Perform the preprocessing on the two text files and report the total number of actors and actresses and total number of unique movies that these actors and actresses have acted in.

After the preprocessing, we have 113129 actors and actresses, and 468244 unique movies.

1.1 Directed actor/actress network creation

Question 2: Create a weighted directed actor/actress network using the processed text file and equation 1. Plot the in-degree distribution of the actor/actress network. Briefly comment on the in-degree distribution.

The in-degree distribution plot is shown below.



From the above graph, we can see the pattern that most of the nodes have low degrees, and this is reasonable because most of the actors and actresses do not have the chance to act in a huge amount of movies and have many connections to other actors and actresses.

1.2 Actor pairings

Question 3: Design a simple algorithm to find the actor pairings. To be specific, your algorithm should take as input one of the actors listed above and should return the name of the actor with whom the input actor prefers to work the most. Run your algorithm for the actors listed above and report the actor names returned by your algorithm. Also for each pair, report the (input actor, output actor) edge weight. Does all the actor pairing make sense?

Our algorithm decides the actor or actress whom the input actor or actress prefers to work the most according to the edge weight. The one with the largest edge weight is the one preferred to work with the most. The concept behind this decision method is that an actor should collaborate with the preferred one in the largest amount of movies among all movies the actor has acted in. Note that ties are broken randomly.
The table of actor pairing is shown below.

Table 1: Actor pairing table

Input actor	**Output actor**	**Edge weight**
Tom Cruise	Nicole Kidman	0.1746
Emma Watson (II)	Daniel Radcliffe	0.52
George Clooney	Damon Matt	0.1194
Tom Hanks	Tim Allen (I)	0.10127
Dwayne Johnson (I)	Steve Austin (IV)	0.20513
Johnny Depp	Helena Bonham Carter	0.08163
Will Smith (I)	Darrell Foster	0.12245
Meryl Streep	Robert De Niro	0.06186
Leonardo DiCaprio	Martin Scorsese	0.10204
Brad Pitt	George Clooney	0.09859

The pairing results come from the collaboration history in movies so all the pairings do make sense. Here are some interesting fact behind the numbers, Tom Cruise and Nicole Kidman were in a marriage relationship and they acted in many movies together. Emma Watson (II) and Daniel Radcliffe were the main characters in the famous Harry Potter movie series and therefore acted in the same movies for many times.

1.3 Actor rankings

Question 4: Use the googles pagerank algorithm to find the top 10 actor/actress in the network. Report the top 10 actor/actress and also the number of movies and the in-degree of each of the actor/actress in the top 10 list. Does the top 10 list have any actor/actress listed in the previous section? If it does not have any of the actor/actress listed in the previous section, please provide an explanation for this phenomenon.

The result demonstration table of the top 10 actor/actress is shown below.

Table 2: Actor ranking table 1

Actor	**Pagerank socre**	**Number of movies acted in**	**Network node in-degree**
Bess Flowers	0.000235	828	7537
Fred Tatasciore	0.00019897	353	3954
Sam Harris (II)	0.00019721	600	6960
Steve Blum (IX)	0.00019551	373	3316
Harold Miller (I)	0.00017272	561	6587
Ron Jeremy	0.00015858	637	2905
Lee Phelps (I)	0.00015732	647	5563
Yuri Lowenthal	0.00015675	317	2662
Atkin Downes Robin	0.00015179	267	2953
Frank O'Connor (I)	0.00014695	623	5502

From the above table, we know that the top 10 list has no actor or actress listed in the previous section. According to our analysis, most of the actors and actresses listed in the previous section are famous and therefore most of the movies they act in should be huge and take longer time to complete in order to correspond to their fame. Besides, these famous actors and actresses have to spend a considerable amount of time attending social activities and have fewer time to act in movies. Thus, it is reasonable that they act in fewer movies and have less chance to collaborate with other actors and actresses, and they have fewer in-degree connections in the network, so their pagerank scores are not the highest.

Question 5: Report the pagerank scores of the actor/actress listed in the previous section. Also, report the number of movies each of these actor/actress have acted in and also their in-degree.

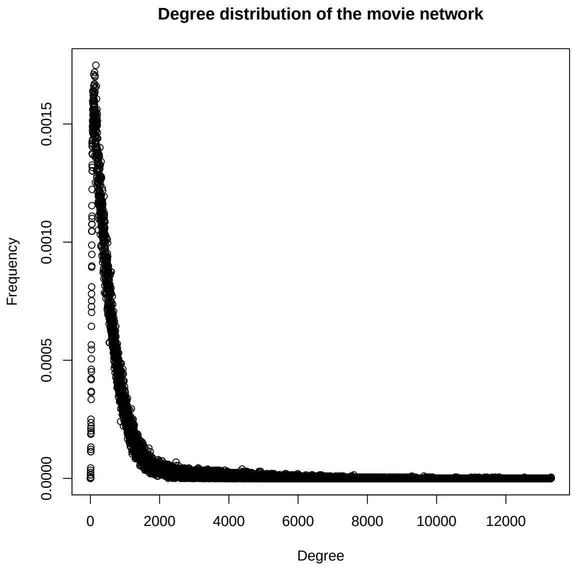
The result demonstration table of the listed 10 actor/actress is shown below.

Table 3: Actor ranking table 2

Actor	**Pagerank score**	**Number of movies acted in**	**Network node in-degree**
Tom Cruise	3.975 x $10^{-}5$	63	1651
Emma Watson (II)	1.749 x $10^{-}5$	25	453
George Clooney	4.004 x $10^{-}5$	67	1573
Tom Hanks	5.107 x $10^{-}5$	79	2064
Dwayne Johnson (I)	4.203 x $10^{-}5$	78	1357
Johnny Depp	5.383 x $10^{-}5$	98	2144
Will Smith (I)	3.202 x $10^{-}5$	49	1319
Meryl Streep	3.962 x $10^{-}5$	97	1594
Leonardo DiCaprio	3.169 x $10^{-}5$	49	1301
Brad Pitt	4.299 x $10^{-}5$	71	1739

From the above table, we have verified what we reasoned in the previous analysis and these famous actors and actresses indeed have fewer in-degree connections and thus lower pagerank scores.

Question 6: Create a weighted undirected movie network using equation 2. Plot the degree distribution of the movie network. Briefly comment on the degree distribution.

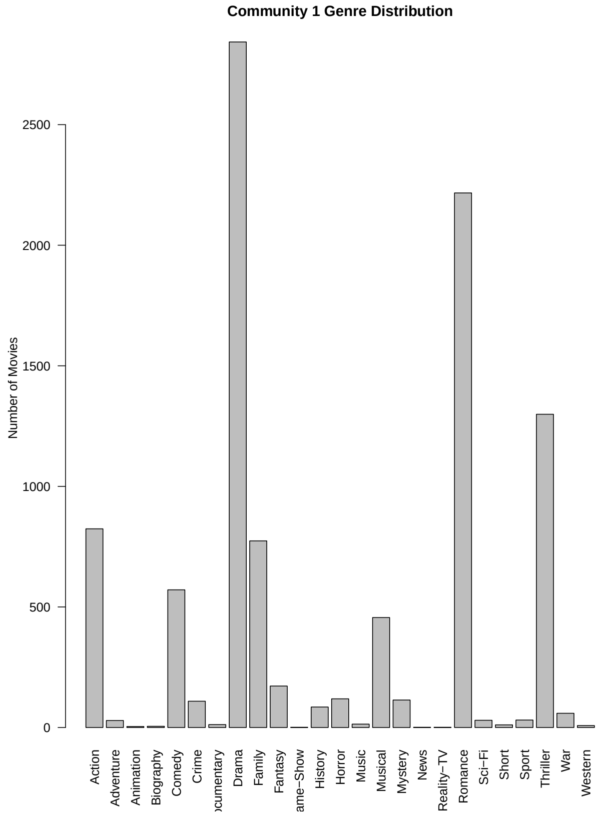


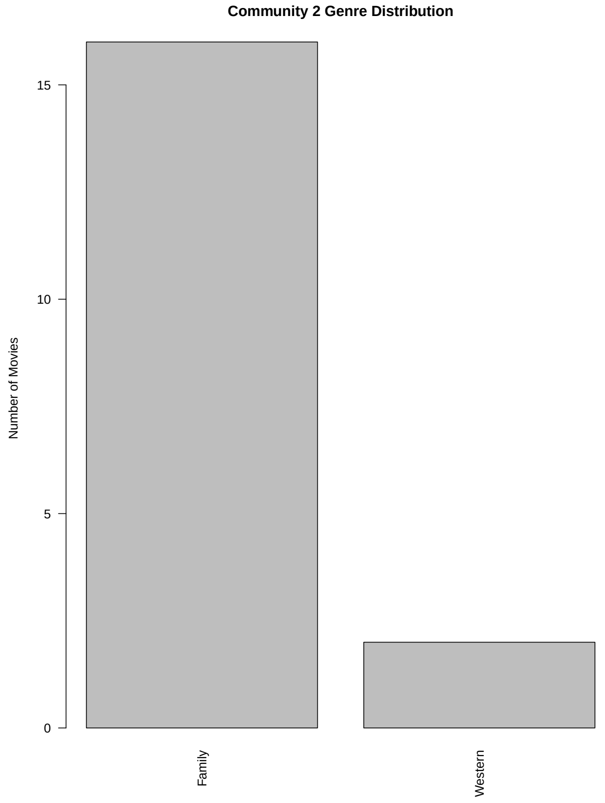
In the graph, two movies are connected if they have at least one same actor. When the numbers of degree reaches around 300 the frequency decreases. Most of the movies are connected with under 2000 movies. This means that only a few movies have duplicate actors with 2000 more movies.

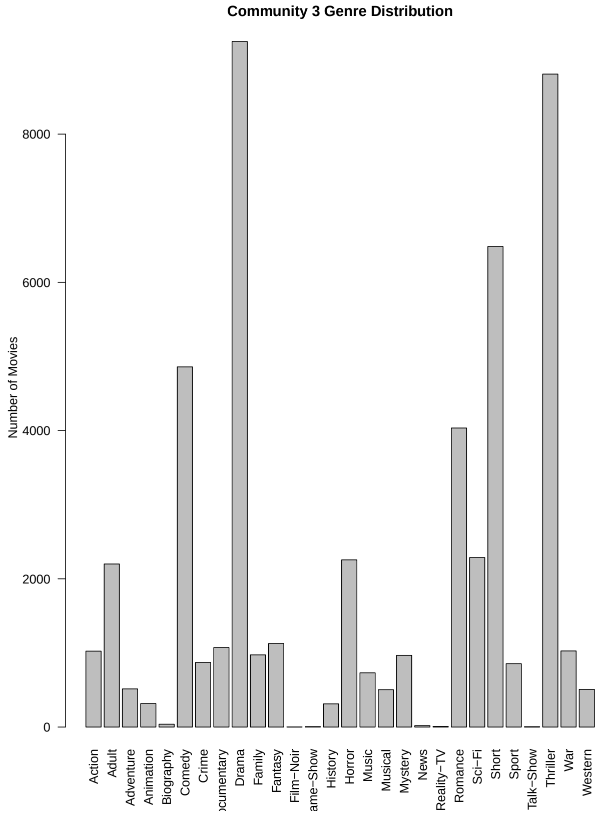
2.2 Communities in the movie network

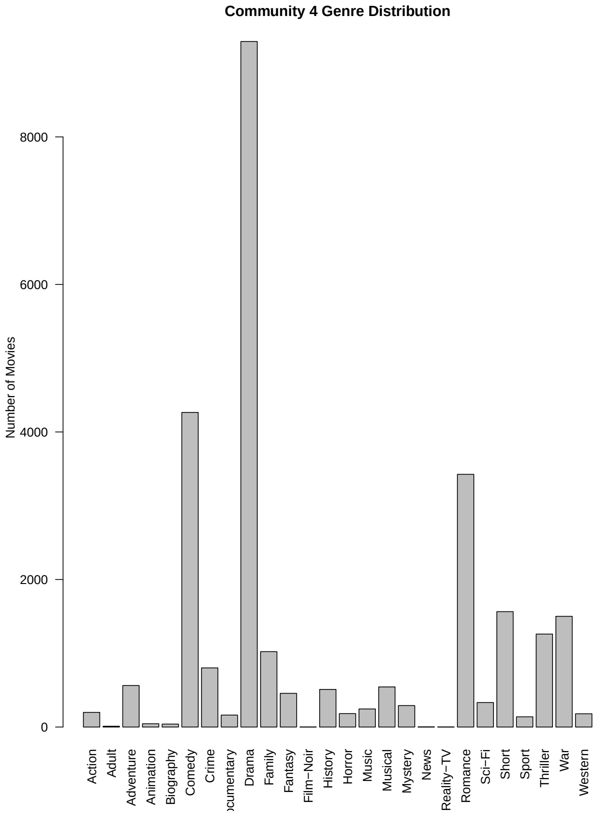
In this part, we will extract the communities in the movie network and explore their relationship with the movie genre. For this part you will need to load the movie_genre.txt file.

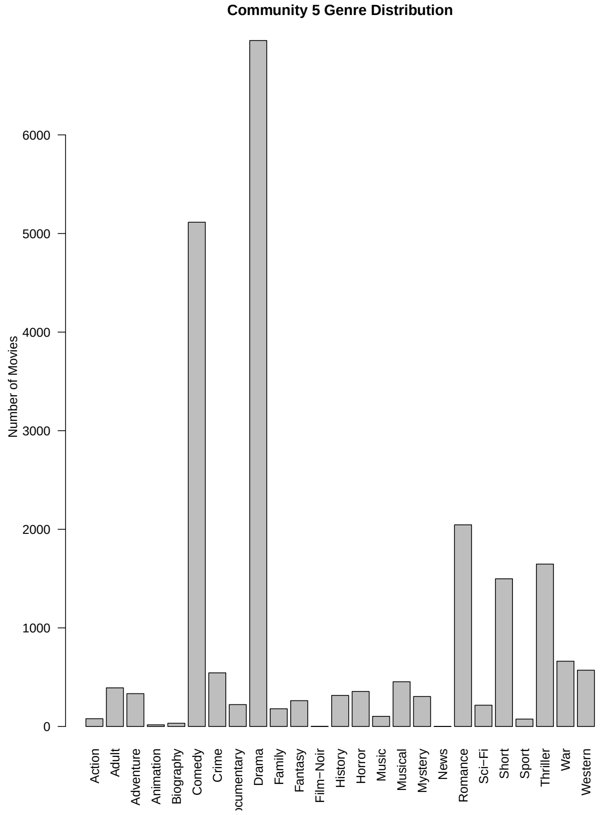
Question 7: Use the Fast Greedy community detection algorithm to find the communities in the movie network. Pick 10 communities and for each community plot the distribution of the genres of the movies in the community.

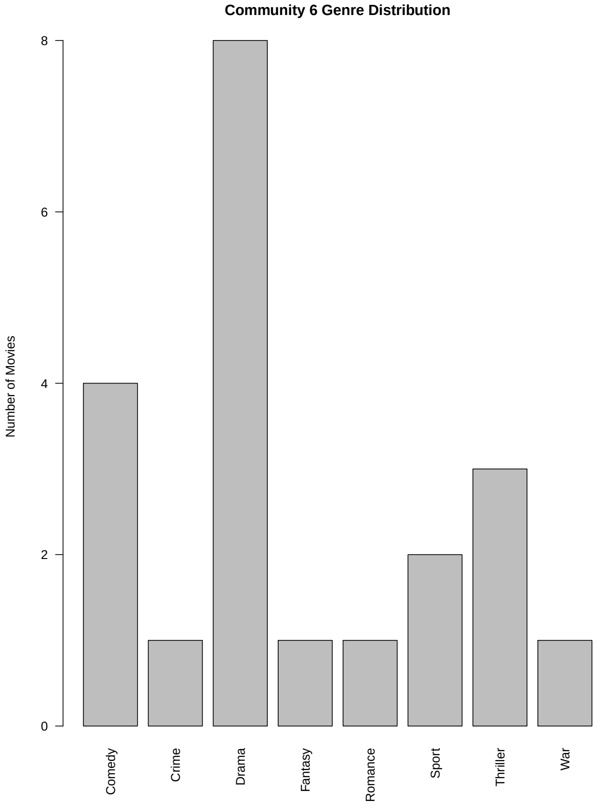


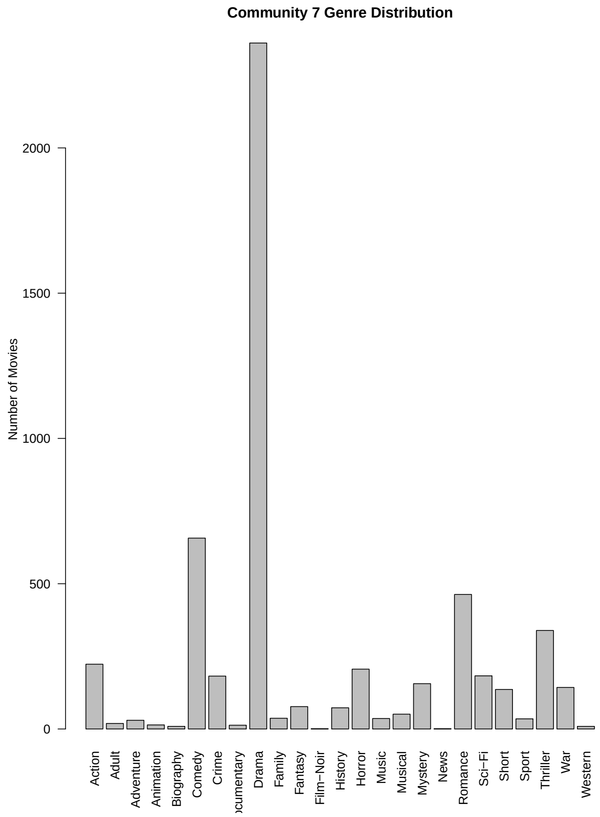


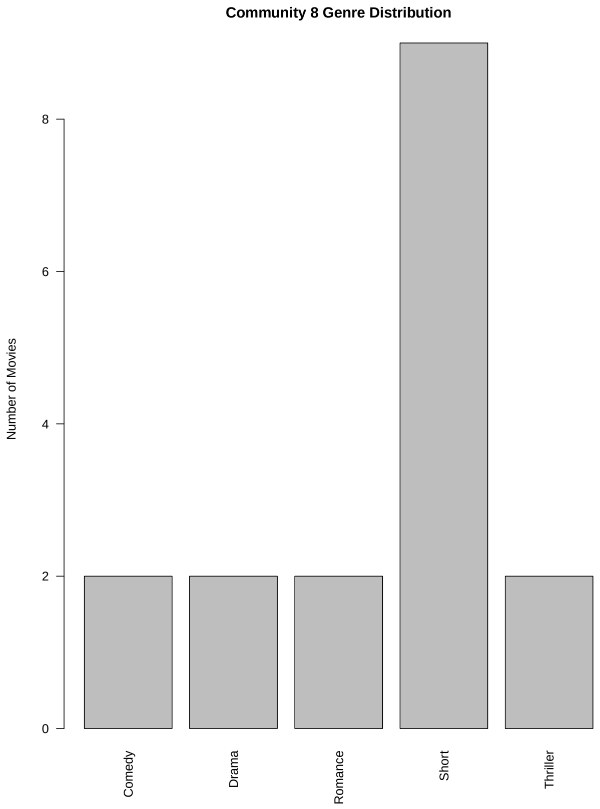

Community 3 Genre Distribution
Number of Movies
0
2000
4000
6000
8000
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
ocumentary
Drama
Family
Fantasy
Film−Noir
ame−Show
History
Horror
Music
Musical
Mystery
News
Reality−TV
Romance
Sci−Fi
Short
Sport
Talk−Show
Thriller
War
Western

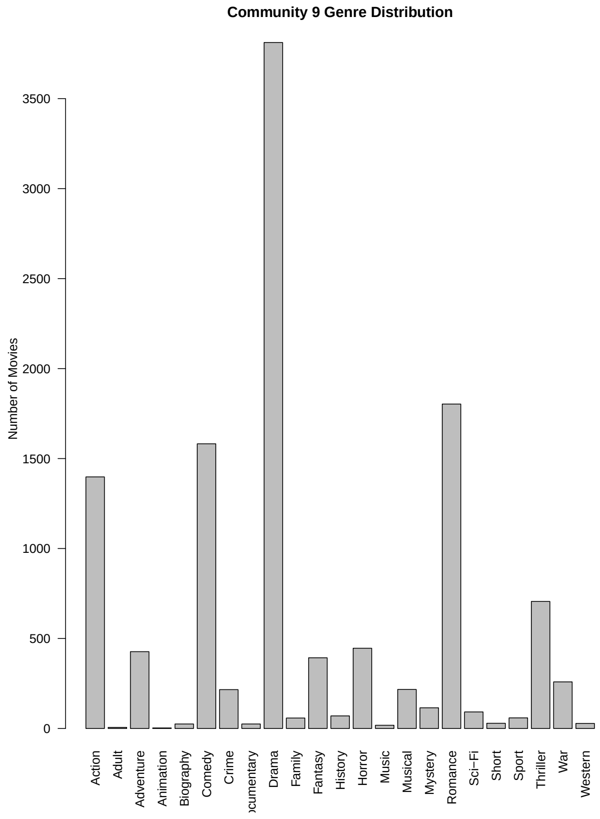

Community 4 Genre Distribution
Number of Movies
0
2000
4000
6000
8000
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
ocumentary
Drama
Family
Fantasy
Film−Noir
History
Horror
Music
Musical
Mystery
News
Reality−TV
Romance
Sci−Fi
Short
Sport
Thriller
War
Western

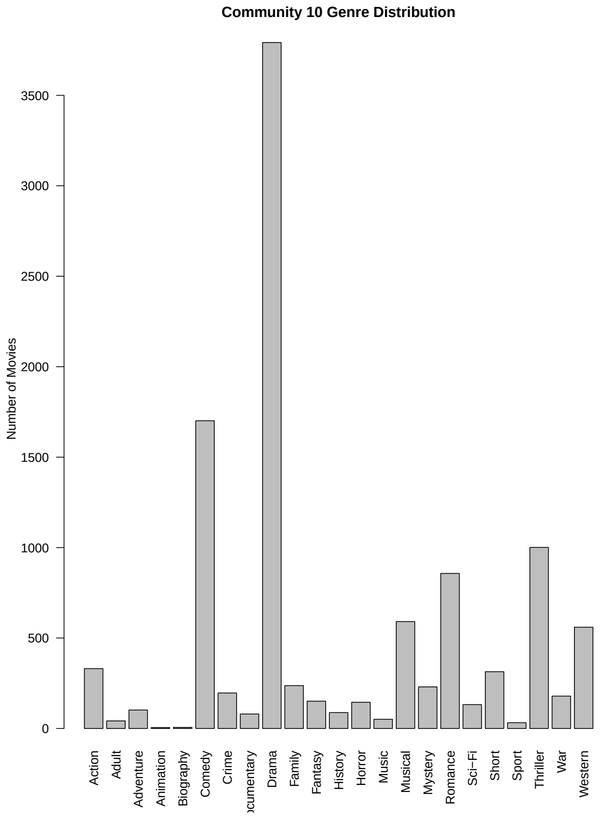

Community 5 Genre Distribution
Number of Movies
0
1000
2000
3000
4000
5000
6000
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
ocumentary
Drama
Family
Fantasy
Film−Noir
History
Horror
Music
Musical
Mystery
News
Romance
Sci−Fi
Short
Sport
Thriller
War
Western


Community 6 Genre Distribution
Number of Movies
0
2
4
6
8
Comedy
Crime
Drama
Fantasy
Romance
Sport
Thriller
War


Community 7 Genre Distribution
Number of Movies
0
500
1000
1500
2000
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
ocumentary
Drama
Family
Fantasy
Film−Noir
History
Horror
Music
Musical
Mystery
News
Romance
Sci−Fi
Short
Sport
Thriller
War
Western


Community 8 Genre Distribution
Number of Movies
0
2
4
6
8
Comedy
Drama
Romance
Short
Thriller


Community 9 Genre Distribution
Number of Movies
0
500
1000
1500
2000
2500
3000
3500
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
ocumentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci−Fi
Short
Sport
Thriller
War
Western


Community 10 Genre Distribution
Number of Movies
0
500
1000
1500
2000
2500
3000
3500
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
ocumentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci−Fi
Short
Sport
Thriller
War
Western

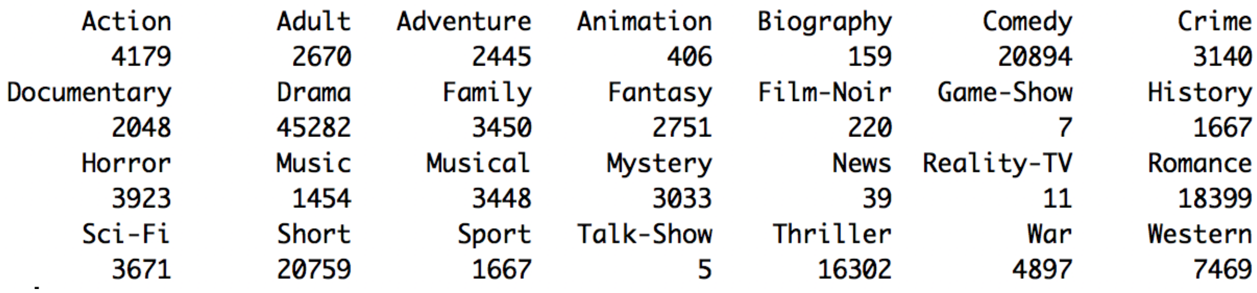
Question 8(a): In each community determine the most dominant genre based simply on frequency counts. Which generes tend to be the most frequent dominant ones across communities and why?

Drama tends to be the most dominant genre because it takes up the most portion of all movie genres. It occupies 25.97% of all movies. It is reasonable that drama appears more frequent in each community since it has a large amount.

Community	Most Dominant Genre
1	Drama
2	Family
3	Drama
4	Drama
5	Drama
6	Drama
7	Drama
8	Short
9	Drama
10	Drama

Table 4: Most dominant genre based on frequency counts

Amount of movies of different genres



```
     Action       Adult   Adventure   Animation   Biography      Comedy       Crime
       4179        2670        2445         406         159       20894        3140
Documentary       Drama      Family     Fantasy   Film-Noir   Game-Show     History
       2048       45282        3450        2751         220           7        1667
     Horror       Music     Musical     Mystery        News  Reality-TV     Romance
       3923        1454        3448        3033          39          11       18399
     Sci-Fi       Short       Sport   Talk-Show    Thriller         War     Western
       3671       20759        1667           5       16302        4897        7469
```

Question 8(b): In each community, for the ith genre assign a score of $\ln(c(i)) \times \frac{p(i)}{q(i)}$ where: $c(i)$ is the number of movies belonging to genre i in the community; $p(i)$ is the fraction of genre i movies in the community, and $q(i)$ is the fraction of genre i movies in the entire data set. Now determine the most dominant genre in each community based on the modified scores. What are your findings and how do they differ from the results in 8(a).

Community	Most Dominant Genre
1	Family
2	Family
3	Adult
4	War
5	Comedy
6	Sport
7	Drama
8	Short
9	Action
10	Musical

Table 5: Most dominant genre based on scores

As we can see in Table 5, drama now is not the most dominant genre across all communities and each community has their own genre. In 8(a), drama is the most dominant since it has a large amount. However, in 8(b) we considered $q(i)$, the fraction of genre i movies in the community, thus the score of drama is be divided by a rather large $q(drama)$. As a result, even though drama has a large amount across many communities, it does not become the most dominant genre of most communities. In my opinion, score in 8(b) is a better and more reasonable method to determine the most dominant genre in each community.

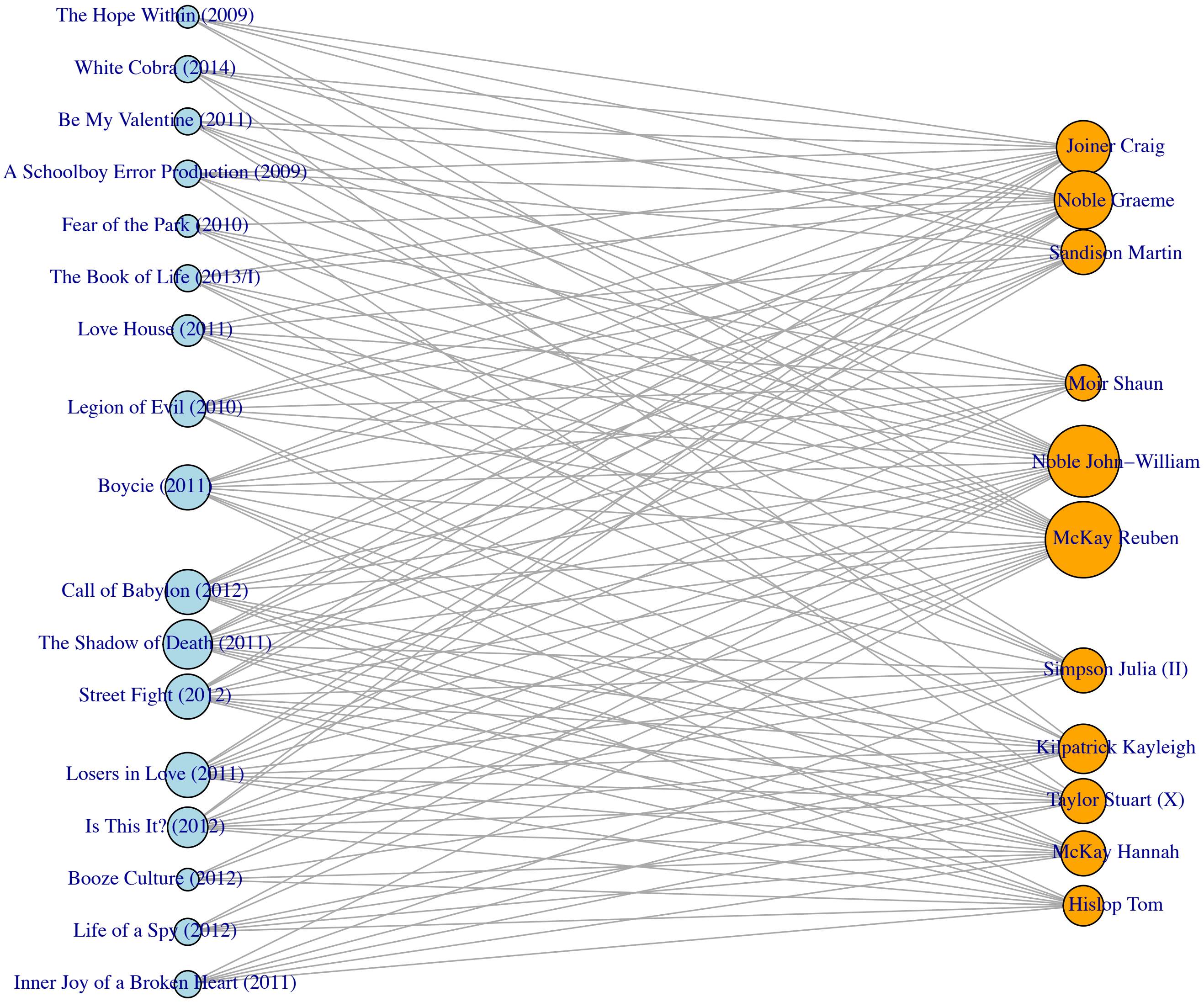
Question 8(c): Find a community of movies that has size between 10 and 20. Determine all the actors who acted in these movies and plot the corresponding bipartite graph (i.e. restricted to these particular movies and actors). Determine three most important actors and explain how they help form the community. Is there a correlation between these actors and the dominant genres you found for this community in 8(a) and 8(b).

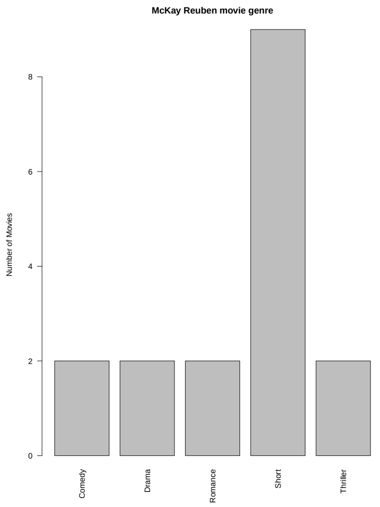
We picked community 8 with 17 movies. Table 6 are the actors in community 8. The 3 most important actors are McKay Reuben, Noble John-William and Noble Graeme. The three actors help form the community since they acted most of the movies in this community. For McKay Reuben, he acted in all 17 movies in community 8. Noble John-Williamm acted in movies 16 and Noble Graeme acted in 13 movies in community 8.

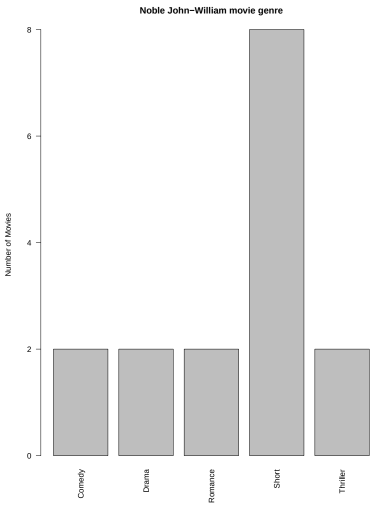
Yes, there is a correlation between these actors and the dominant genres in both 8(a) and 8(b). The dominant genre in 8(a) is short, since there are 9 short movies in total 17 movies. The dominant genre in 8(b) considered the genres in the all communities and the dominant genre is also short. Both three actors acted 9 or 8 short movies. Thus, the more a movie genre these important actors act, the more dominant a genre is in the community.

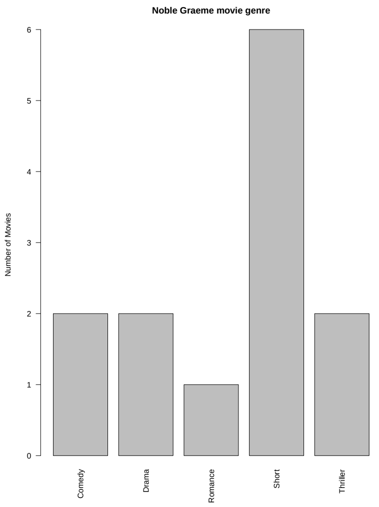
Actor Name	Degrees
Joiner Craig	12
Noble Graeme	**13**
Sandison Martin	10
Moir Shaun	8
Noble John-William	**16**
McKay Reuben	**17**
Simpson Julia (II)	10
Kilpatrick Kayleigh	11
Taylor Stuart (X)	10
McKay Hannah	10
Hislop Tom	9

Table 6: Actors' movies in community 8









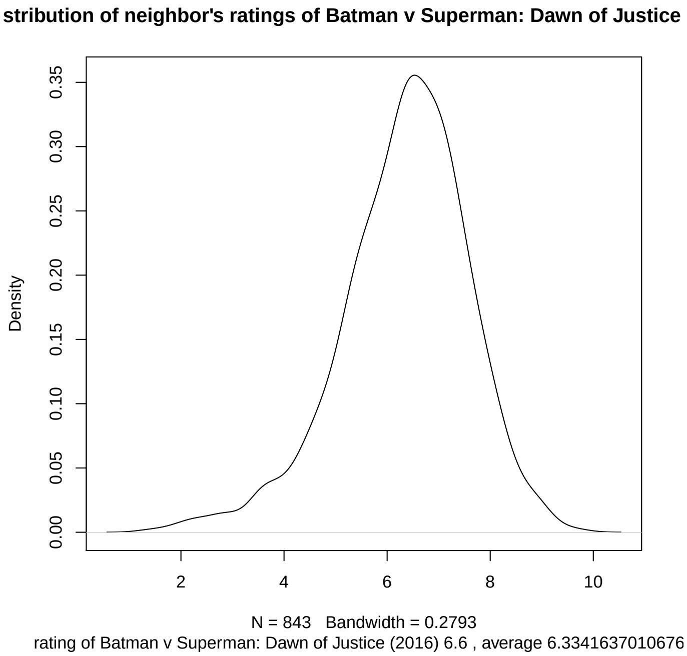
2.3 Neighborhood analysis of movies

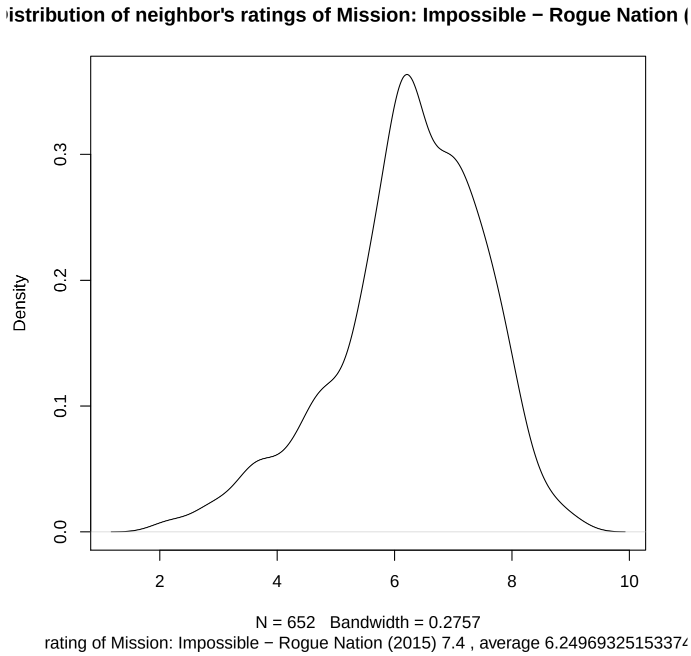
In this part of the project, you will need to load the movie_rating.txt file and we will explore the neighborhood of the following 3 movies:

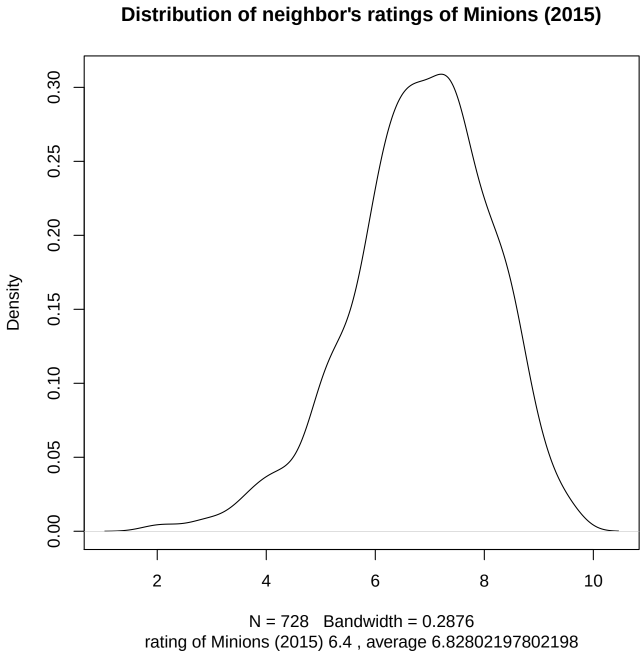
- Batman v Superman: Dawn of Justice (2016); Rating: 6.6
- Mission: Impossible - Rogue Nation (2015); Rating: 7.4
- Minions (2015); Rating: 6.4

Question 9: For each of the movies listed above, extract its neighbors and plot the distribution of the available ratings of the movies in the neighborhood. Is the average rating of the movies in the neighborhood similar to the rating of the movie whose neighbors have been extracted? In this question, you should have 3 plots.

In this part, we can use *neighbors* function in R to get the neighbors' movie name. Then we can simply extract their ratings by their names from the txt file and plot their distribution.
As you can see from the figures, the average rating of the movies in the neighborhood is similar to the rating of the movie whose neighbors have been extracted.

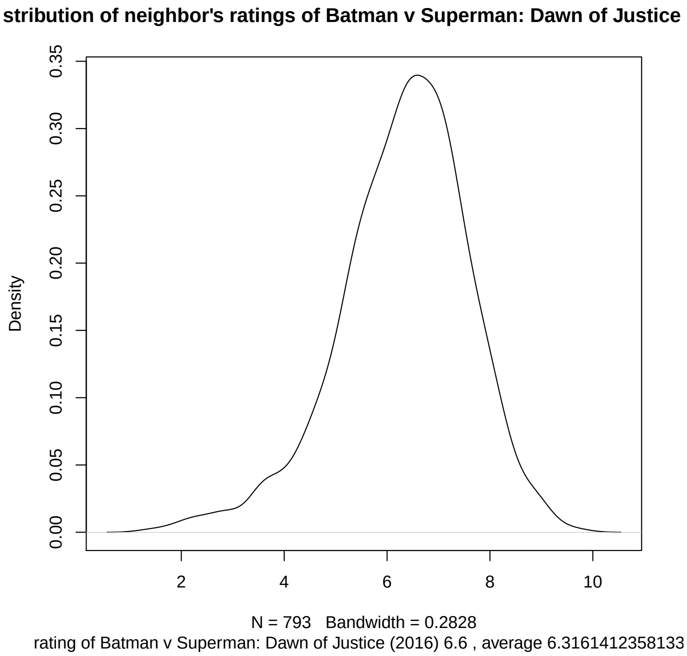

stribution of neighbor's ratings of Batman v Superman: Dawn of Justice
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
10
N = 843 Bandwidth = 0.2793
rating of Batman v Superman: Dawn of Justice (2016) 6.6 , average 6.3341637010676

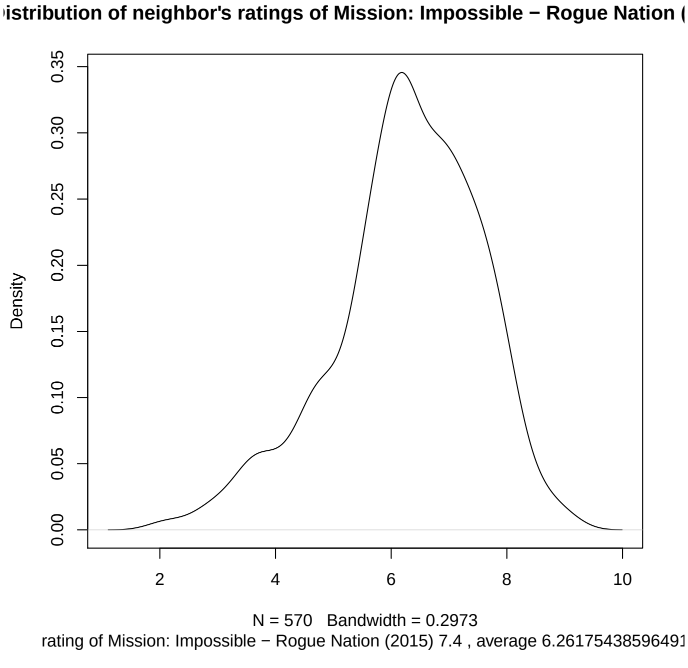

istribution of neighbor's ratings of Mission: Impossible − Rogue Nation (
Density
0.0
0.1
0.2
0.3
2
4
6
8
10
N = 652 Bandwidth = 0.2757
rating of Mission: Impossible − Rogue Nation (2015) 7.4 , average 6.2496932515337

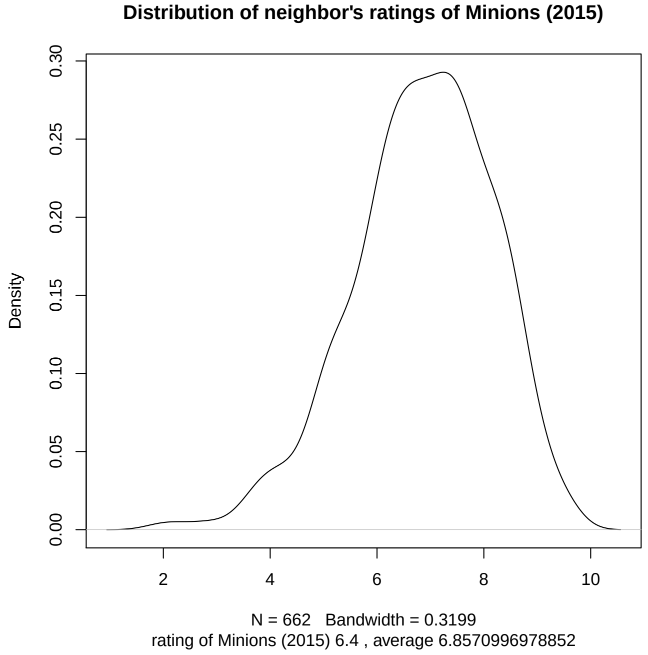


Question 10: Repeat question 10, but now restrict the neighborhood to consist of movies from the same community. Is there a better match between the average rating of the movies in the restricted neighborhood and the rating of the movie whose neighbors have been extracted. In this question, you should have 3 plots.

First we need to use *induced_subgraph* function to get a smaller network that only consist of movies from the same community. Then we can *neighbors* function as in Question 9 to get the neighbors' movie name and extract their ratings by their names from the txt file and plot their distribution. As you can see from the figures, the number of neighbors decreases (**N** in the figures) but the overall distributions do not change much. Hence the average rating of the movies in the neighborhood is also similar to the rating of the movie whose neighbors have been extracted.


stribution of neighbor's ratings of Batman v Superman: Dawn of Justice
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
10
N = 793 Bandwidth = 0.2828
rating of Batman v Superman: Dawn of Justice (2016) 6.6 , average 6.3161412358133


istribution of neighbor's ratings of Mission: Impossible − Rogue Nation (
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2
4
6
8
10
N = 570 Bandwidth = 0.2973
rating of Mission: Impossible − Rogue Nation (2015) 7.4 , average 6.26175438596491



Question 11: For each of the movies listed above, extract its top 5 neighbors and also report the community membership of the top 5 neighbors. In this question, the sorting is done based on the edge weights

After sorting the edges based on their weights, we can get the top 5 neighbors for each of the movies listed above.
The following tables show the neighbors' name, edge weights between them and their community membership.

name	weight	community
Eloise (2015)	0.3	3
Man of Steel (2013)	0.233333	3
Into the Storm (2014)	0.233333	3
Real Steel (2011)	0.172413	3
Love and Honor (2013)	0.161290	3

Table 7: Top 5 neighbors of Batman v Superman

name	weight	community
Fan (2015)	0.481481	3
Phantom (2015)	0.464285	3
Muppets Most Wanted (2014)	0.448275	3
Breaking the Bank (2014)	0.354838	3
Avengers: Age of Ultron (2015)	0.322580	3

Table 8: Top 5 neighbors of Mission: Impossible

name	weight	community
Cars (2006)	0.75	3
Up (2009)	0.533333	3
The Lorax (2012)	0.526315	3
Monsters University (2013)	0.473684	3
Despicable Me 2 (2013)	0.45	3

Table 9: Top 5 neighbors of Minions

Question 12: Train a regression model to predict the ratings of movies: for the training set you can pick any subset of movies with available ratings as the target variables; you have to specify the exact feature set that you use to train the regression model and report the root mean squared error (RMSE). Now use this trained model to predict the ratings of the 3 movies listed above (which obviously should not be included in your training data).

In this question, we train a linear regression model to predict the ratings of the 3 movies.
For the training set we used the movies with at least 10 actors and with director information available. The number of movies in the training set is 83533.
The features that we used are as follows:

1. Top 6 pagerank of the actors of that movie.

2. The rating of the director. In case of more than one director for a movie, we use the average ratings of the directors for that movie. The ratings of a director is defined as the average ratings of the movies filmed by that director.

This means that our feature vectors each has a length of 7.
We construct the feature vector for each movie in the training set and train a linear regression model using sklearn package. The RMSE for the training set of our linear model is 0.688.
The predicted ratings from our linear model for the 3 movies are:

movie	predicted rating	real rating (from IMDb)
Batman v Superman: Dawn of Justice (2016)	7.28	6.6
Mission: Impossible - Rogue Nation (2015)	6.84	7.4
Minions (2015)	7.15	6.4

Table 10: Predicted ratings

Our predicted rating for the 3 movies has a RMSE of 0.444.

Question 13: Create a bipartite graph following the procedure described above. Determine and justify a metric for assigning a weight to each actor. Then, predict the ratings of the 3 movies using the weights of the actors in the bipartite graph. Report the RMSE. Is this rating mechanism better than the one in question 12? Justify your answer.

In this question, we create a bipartite graph with V_1 represents actor/actress and V_2 represents movies and an edge e_{ij} between V_1 and V_2 if actor i has act in movie j.

We define the weight of the actor i (weight of the edges e_{ij}) as the average ratings of the movies that the actor has acted in.
We then predict the ratings of the movie based on the average weights of the actors for that movie. The predicted ratings using this method for the 3 movies are shown in the following table.

movie	predicted rating	real rating (from IMDb)
Batman v Superman: Dawn of Justice (2016)	6.50	6.6
Mission: Impossible - Rogue Nation (2015)	6.49	7.4
Minions (2015)	6.87	6.4

Table 11: Predicted ratings

The RMSE of this model for the 3 movies is 0.353.
The predicted ratings are fairly close to the real ratings and the RMSE in this case is lower than the linear regression model in Q12. However, if we compute the RMSE of this model for the movies in the training set in Q12, the RMSE becomes 0.974 which is higher than that in Q12. On average we would expect the prediction of rating mechanism of this model to be worse than the linear regression model in Q12 because in Q12 we can assign different weight to the top 6 actors through the fitted linear model parameter while in this case we're only taking the average of the average ratings for each actor. In addition, we did not take the average ratings of the director into account in this model.
There could be two improvements to this model. First, the leading roles, say Tom Cruise and Jeremy Renner in the movie Mission: Impossible - Rogue Nation (2015), has a long history in film history and the rating of the movies they played in their early career may not reflect the rating of their more current movies. Because we include the ratings of all movies they have been on, the weight of these leading actors are lower than one might have expected (6.8 for Tom Cruise). A more accurate metric for the weight of each actor could be the average of the ratings of the movies they acted in for the last 10 years. Second, all actors are treated equally when using their weight to compute the predicted rating of the movie. This might be the biggest assumption of this model because we would expect the star actor's ratings should weight more.