Large-Scale Data Mining: Models and Algorithms
ECE 232E Spring 2018

Project 4
IMDb Mining

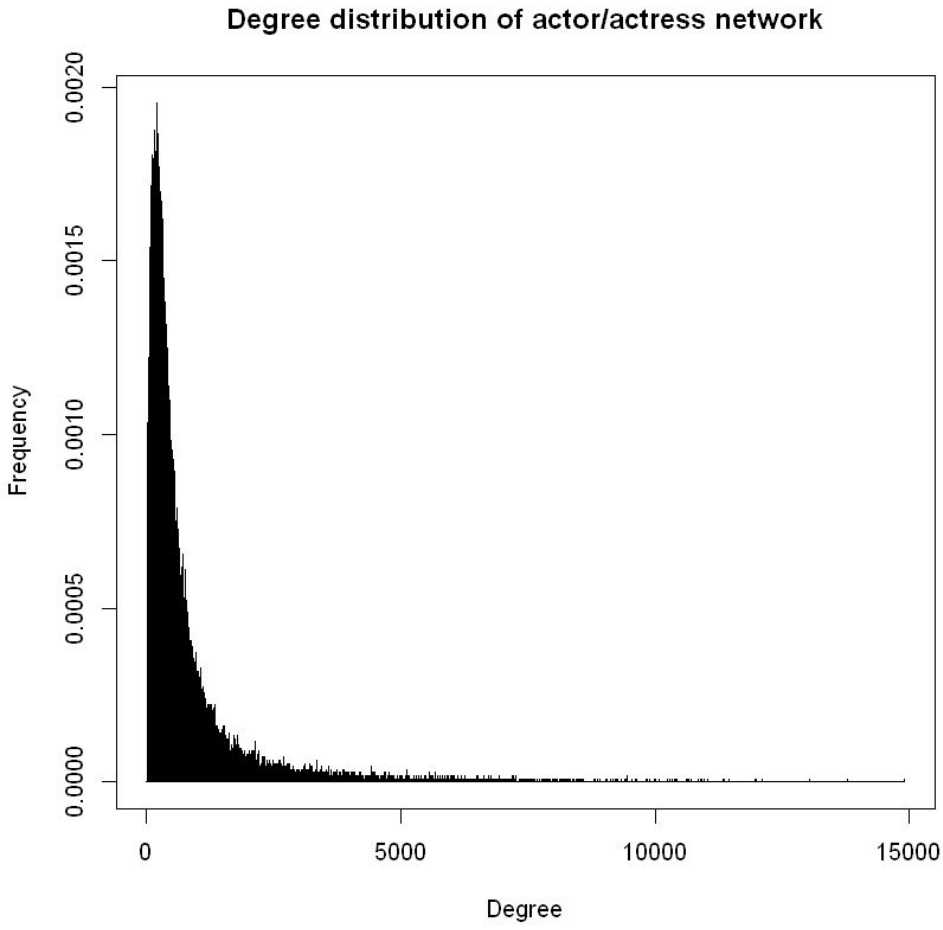
Qinyi Tang (204888348)
Shuo Bai (505032786)
Jinxi Zou (605036454)
Xuan Hu (505031796)
6/5/2018

1 Actor/Actress network

Answer for Question 1 :

We first merge “actor_movies.txt” and “actress_movies.txt” using shell script (“mergefile.sh”), and then we do the preprocess of the data using python. As the statement require, we first removed the movie with less than 10 actors/actresses. Then, we used the regular expression to match the name. The format we used to match is '\(\d\d\d\d(.*?)\)|\(\?\?\?\?(.*?)\)', trying to find the movies’ name ended with the year or (????). Also, in this question, we mapped the actors/actresses to ids, because we find that there are some actors have the same name, but are different individuals.
The total number of actors and actresses is 113132, and the total number of unique movies that these actors and actresses have acted in is 468201.

Answer for Question 2 :



From the plot above, we could get the conclusion that most actors / actresses only cooperate with limited number of actors/actresses. Few people have the opportunity to work with more than 5000 actors/actresses.

Answer for Question 3：

Actor	Pairing Actor	Weight
Cruise, Tom	Kidman, Nicole	0.174603
Watson, Emma (II)	Grint, Rupert	0.56
Clooney, George	Damon, Matt	0.119403
Hanks, Tom	Ratzenberger, John	0.1375
Johnson, Dwayne (I)	Calaway, Mark	0.205128
Depp, Johnny	Bonham Carter, Helena	0.081633
Smith, Will (I)	Foster, Darrell	0.122449
Streep, Meryl	De Niro, Robert	0.061856
DiCaprio, Leonardo	Scorsese, Martin	0.102041
Pitt, Brad	Clooney, George	0.098592

Without a doubt, Emma Watson paired with Rupert Grint and the weight from Emma to Grint is even reach up to 0.56! A lot of pairs show the actor/actress they prefer to work with. However, there are also some pairs which make no sense. For example, Streep Meryl has shown on 49 different movies and he only cooperated with De Niro, Robert three times. Although Robert is the person he worked with most times, I don't think this implies Streep prefer to work with him.

Answer for Question 4：

Top10 actor/actress	Number of movie	In degree	Pagerank score
Flowers, Bess	828	14912	0.0002329559
Tatasciore, Fred	355	7731	0.0002037972
Blum, Steve (IX)	373	6586	0.0001979227

Harris, Sam (II)	600	13774	0.0001959681
Miller, Harold (I)	561	13030	0.0001716543
Lowenthal, Yuri	318	5249	0.000158379
Phelps, Lee (I)	647	11027	0.000157087
Downes, Robin Atkin	267	5850	0.0001549555
Jeremy, Ron	636	5647	0.0001548585
O'Connor, Frank (I)	623	10882	0.0001462057

Top 10 list doesn't have any actor/actress list in the previous section. It seems that actor/actress in top 10 list are not as famous as actor/actress in the previous section. Famous stars don't mean they show on a lot of movies. This phenomenon implies a excellent actor/actress may pay much more attention on the quality of movie instead of the quantity of movies. It is easy for a actor/actress to play a uncredited role in a movie, but it is difficult for a actor/actress to interpret protagonist well in a movie. Another reason may lead to this situation is that the actor/actress in top list wasn't active in recent year. Their works were shown on screen in the past twenty or thirty years, even more. For example, Bess Flowers (1898-1984), was well-known in the last century, however, too far for us to know her name due to time.

Answer for Question 5 :

Listed actor/actress	Number of movie	In degree	Pagerank score
Cruise, Tom	63	3230	4.01162e-05
Watson, Emma (II)	25	896	1.849868e-05
Clooney, George	67	3121	4.123254e-05
Hanks, Tom	80	4008	5.100117e-05
Johnson, Dwayne (I)	78	2700	4.347792e-05
Depp, Johnny	98	4181	5.401969e-05
Smith, Will (I)	49	2550	3.088614e-05
Streep, Meryl	97	3072	3.927147e-05

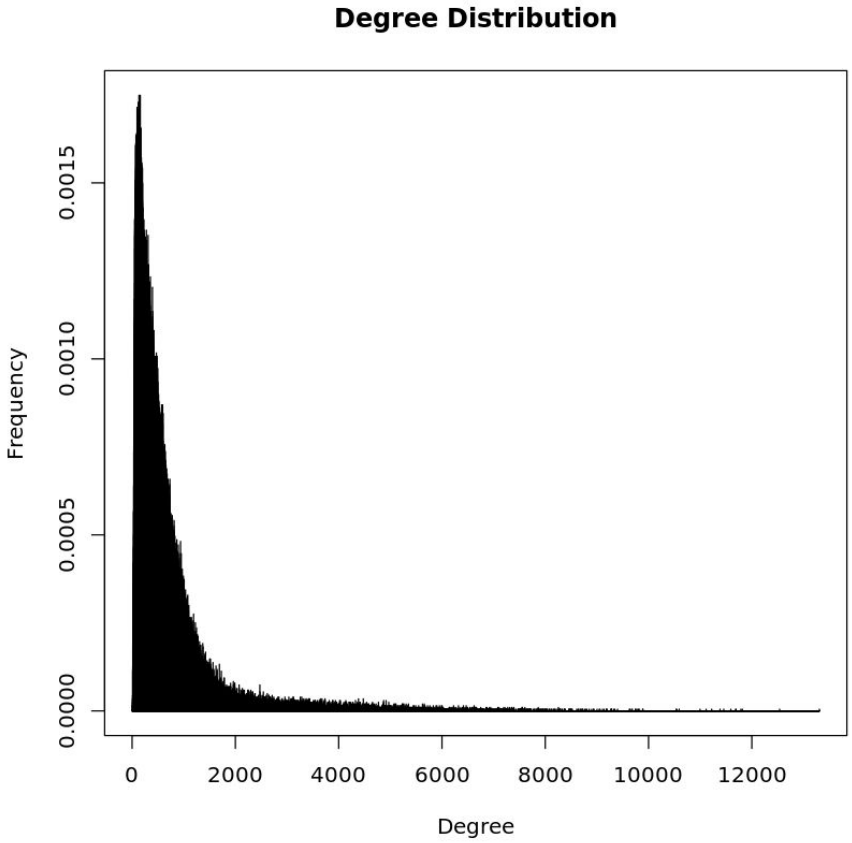
DiCaprio, Leonardo	49	2554	3.203155e-05
Pitt, Brad	71	3358	4.238431e-05

2 Movie network

Answer for Question 6 :

In this question, we used python (question6_py.ipynb) to do the txt file preprocessing and to generate the edgelist file, and used R (question6_r.ipynb) to read the graph and to plot the degree distribution. When we generate the edgelist file, we iterate the movies and use the movie dictionary to find the actor name, and then we use the actor name to find his other movies. Therefore, we can know the intersection of the actor list of any two movies in an efficient way. Comparably, if we iterate any two movies to find their actors in common, it will waste a lot of time in trying to checking whether there is any intersection in two movies that actually do not have actors in common, considering the large scale of our data.

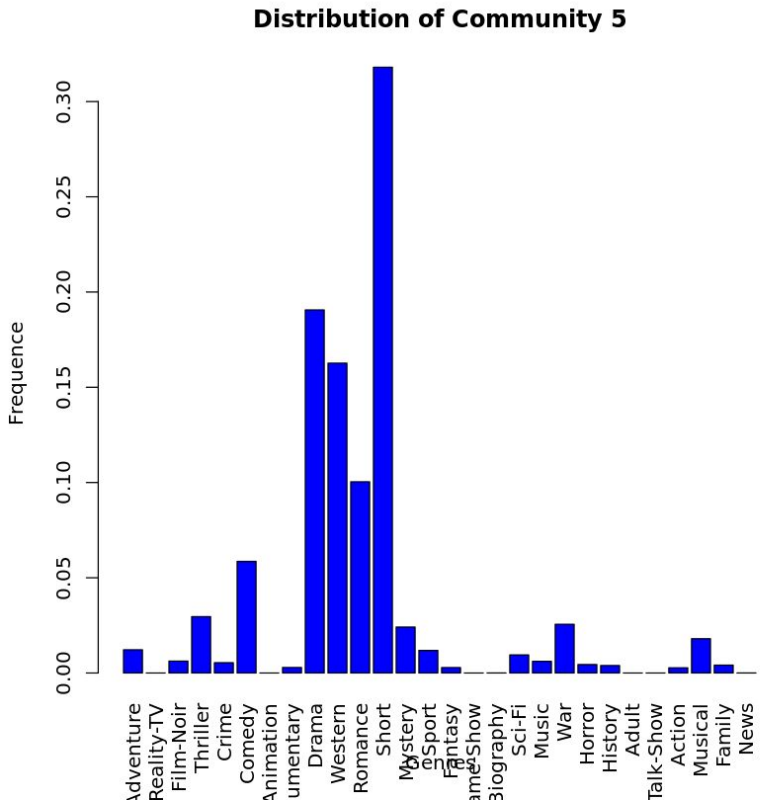
The degree distribution plot of the movie network is as following:

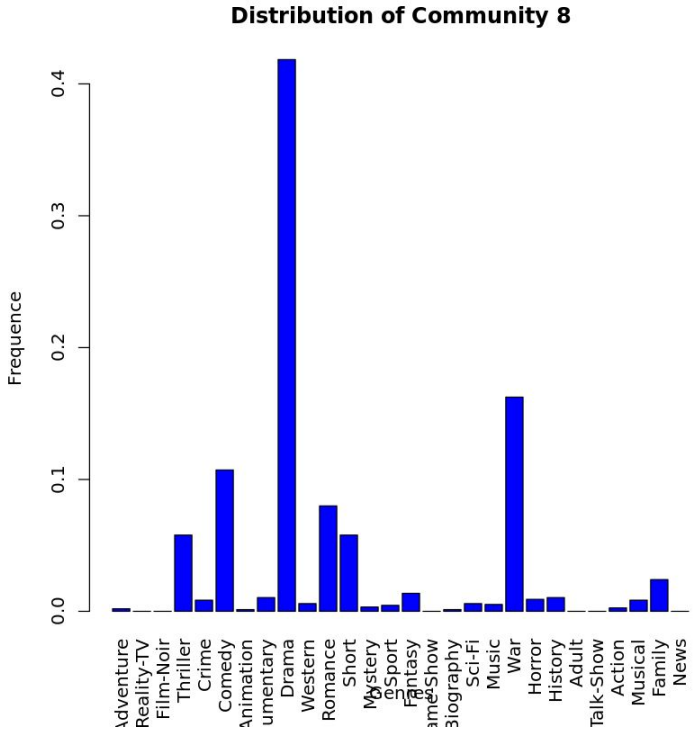


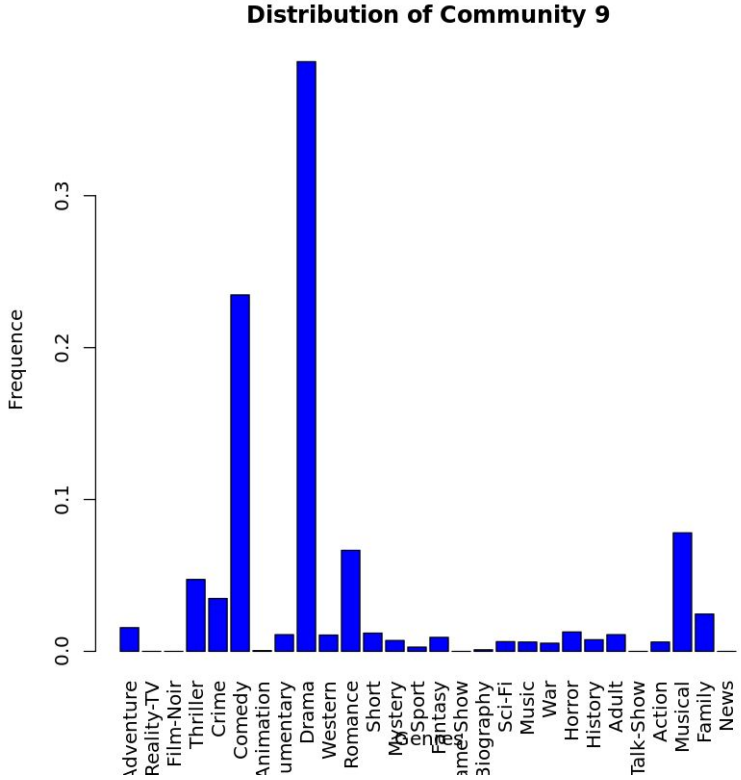
From the degree distribution plot, it matches our common sense. There are few vertices with the degree approaching to 0. In real life, there almost none actor or actress have only acted in only one movie. Also, there are few vertices having a very large degree, because the number of the actors/actresses in one movie is limited, and the actors have the limited movies. In our dataset (after pruning), in average, each actor/actress acted in 28 movies, and each movie has 14 actors/actresses. Thus, the most frequent degree number is about 150.(Some movies may have may than one actors/actresses in common).

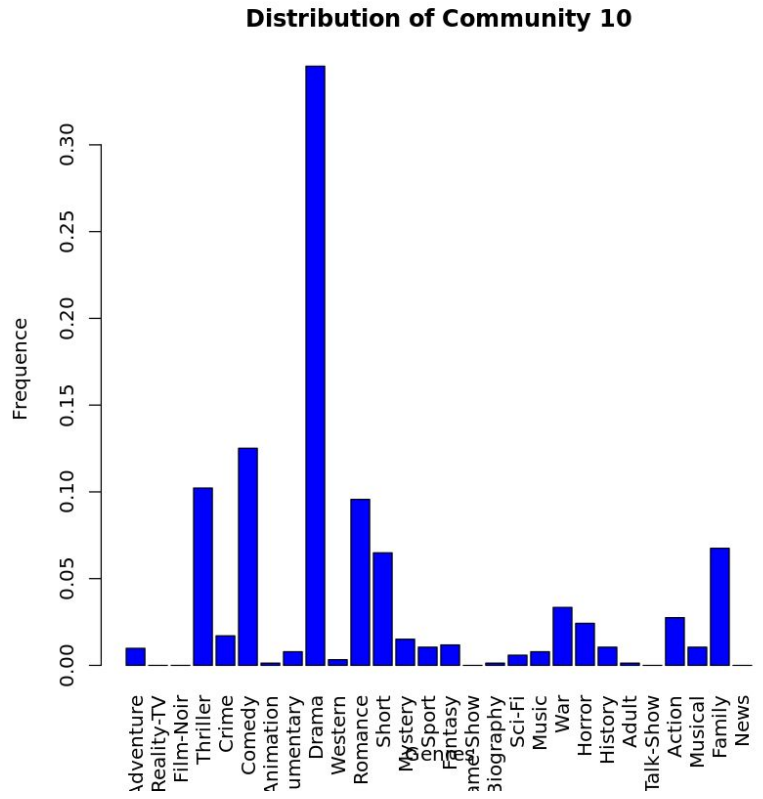
Answer for Question 7：

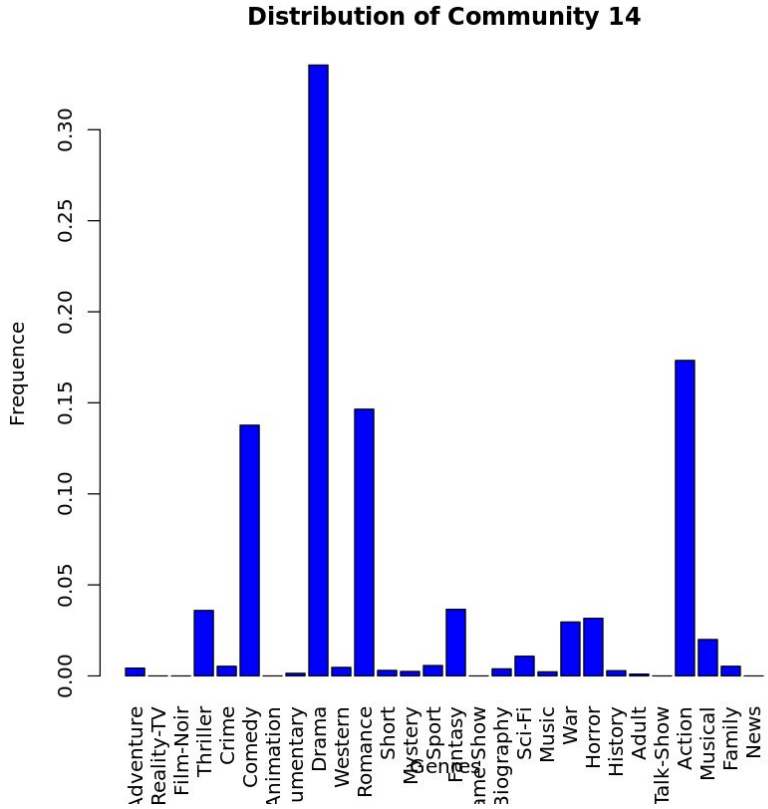
We first load the movie_genere.txt file to the graph by adding the genre properties to the vertices. Some vertices'(movies') genre information is missing, so we set their genre info as "null". We used the fast greedy algorithm to find the communities by applying "cluster_fast_greedy". After running the cluster_fast_greedy algorithm, we find there are 30 communities in this network. We picked Community 5,8,9,10,14,15,17,19,25,28. When we calculate the frequency for the distribution of different genres, we ignore the movies without genre information(not taking into account in genre_count). Their degree distribution plots are as following:

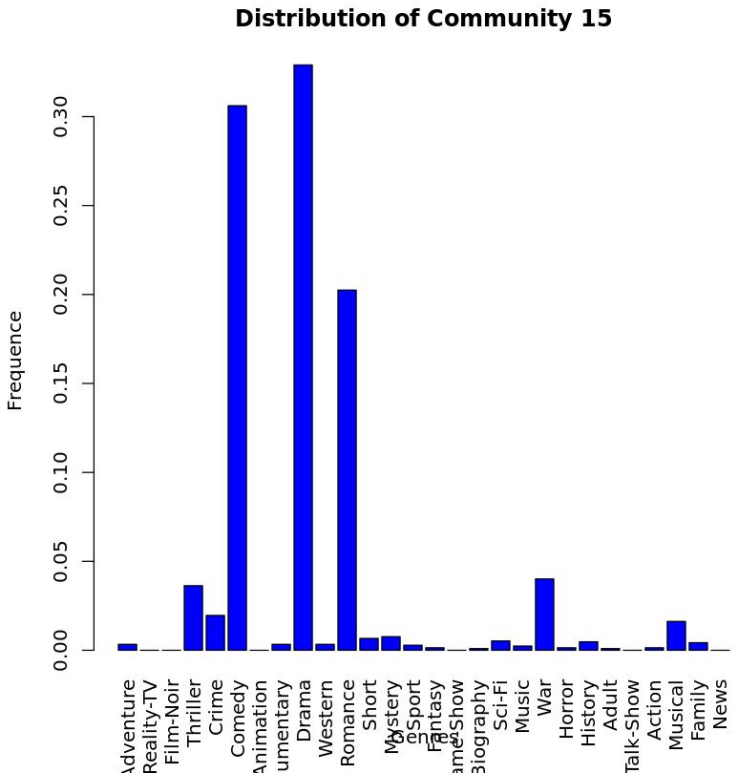


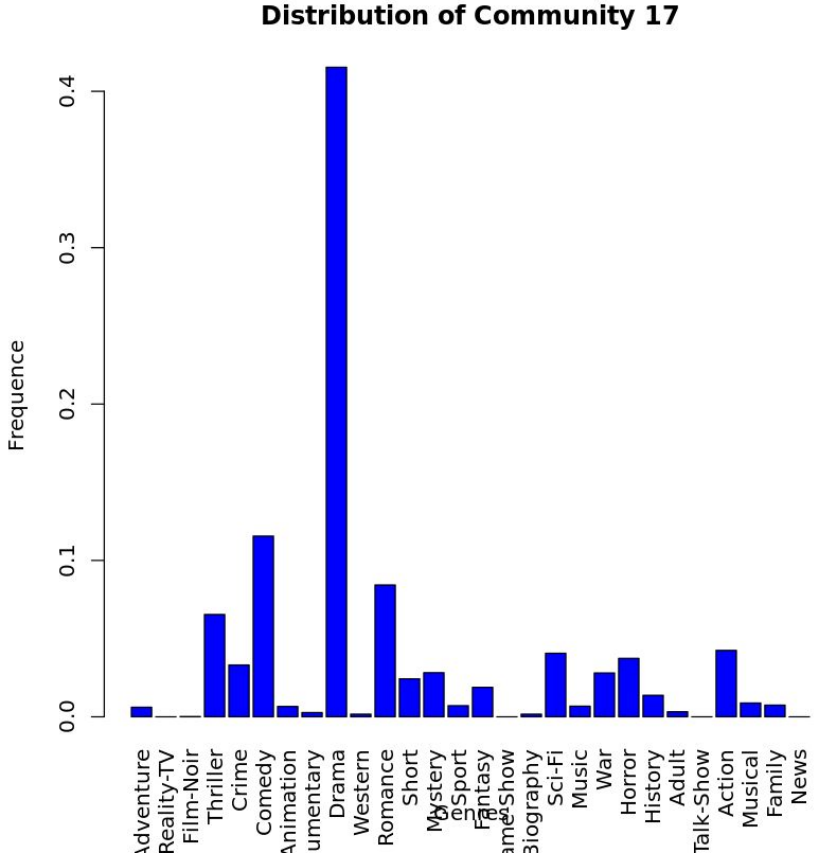

Distribution of Community 8
Frequence
0.0
0.1
0.2
0.3
0.4
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News

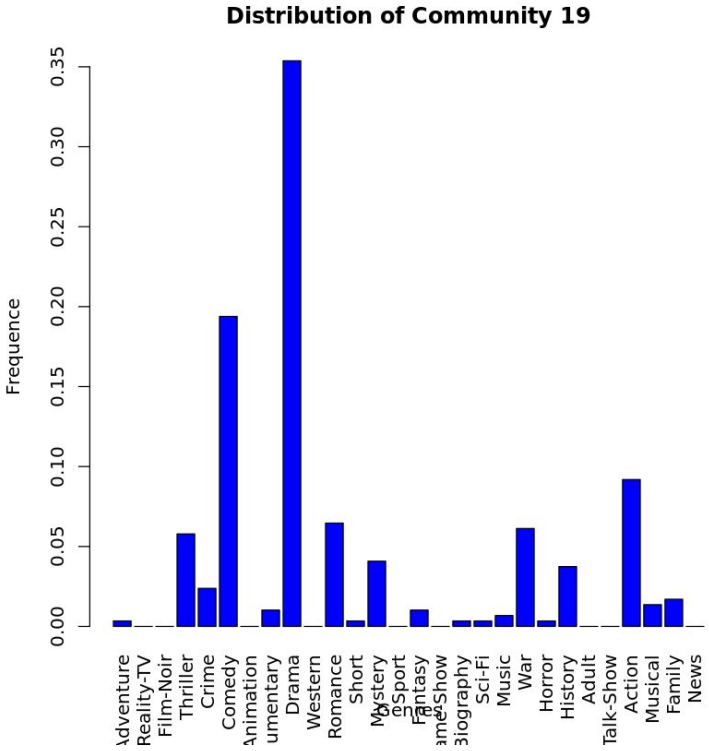

Distribution of Community 9
Frequence
0.0
0.1
0.2
0.3
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News

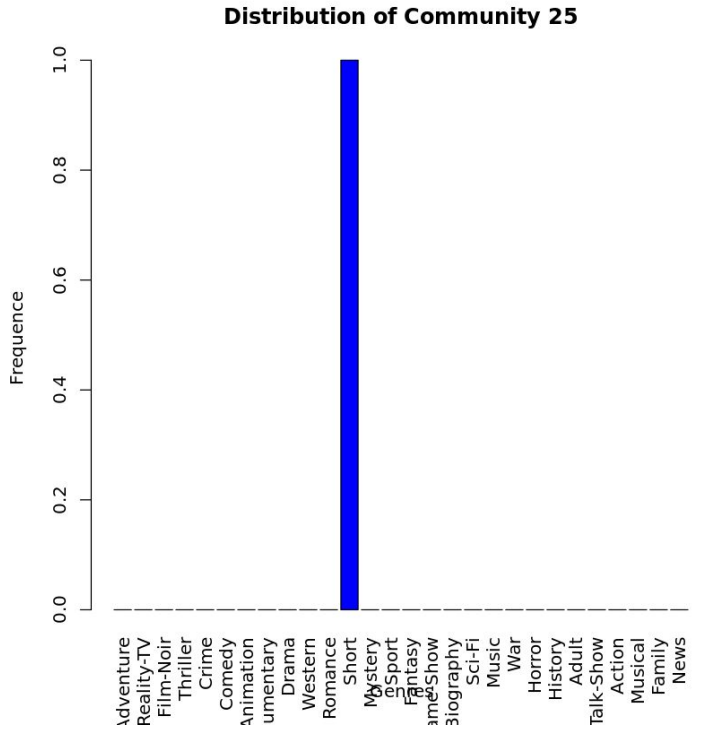

Distribution of Community 10
Frequence
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
umentary
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Show
Biography
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News

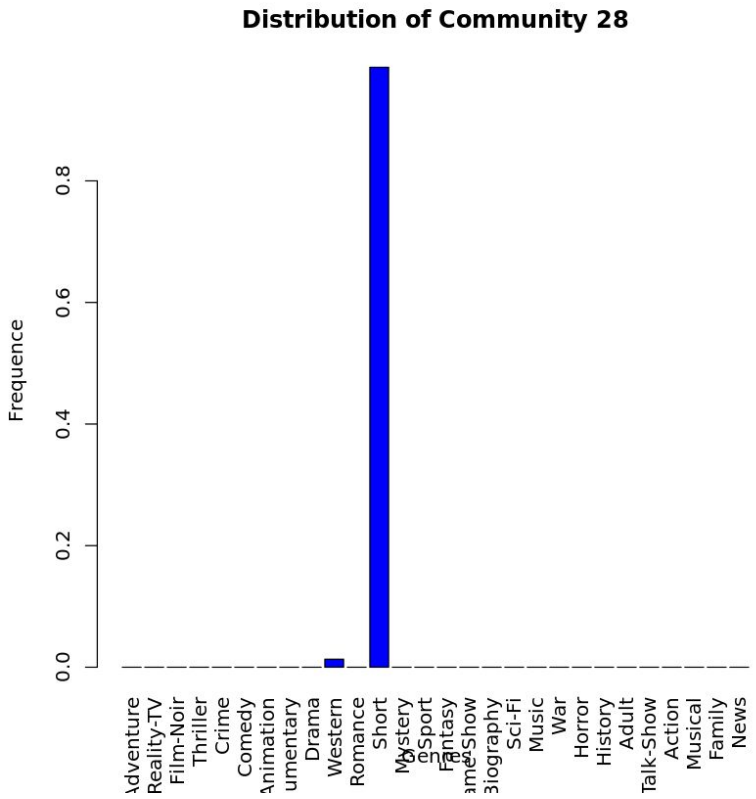

Distribution of Community 14
Frequence
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
umentary
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Show
Biography
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News


Distribution of Community 15
Frequence
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
Documentary
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Game-Show
Biography
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News


Distribution of Community 17
Frequence
0.0
0.1
0.2
0.3
0.4
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
Documentary
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Game-Show
Biography
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News


Distribution of Community 19
Frequence
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
umentary
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Show
Biography
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News


Distribution of Community 25
Frequence
0.0
0.2
0.4
0.6
0.8
1.0
Adventure
Reality-TV
Film-Noir
Thriller
Crime
Comedy
Animation
umentary
Drama
Western
Romance
Short
Mystery
Sport
Fantasy
Show
Biography
Sci-Fi
Music
War
Horror
History
Adult
Talk-Show
Action
Musical
Family
News



Answer for Question 8(a) :

We try to find the most dominant genre by simply calculating the frequency. The results are as shown in the following figure:

Community No.	Most Dominant Genre	Max Frequency
1	Thriller	0.188
2	Drama	0.255
3	Drama	0.402
4	Drama	0.302
5	Short	0.318
6	Drama	0.316
7	Adult	0.623
8	Drama	0.418

9	Drama	0.388
10	Drama	0.345
11	Drama	0.313
12	Drama	0.404
13	Drama	0.282
14	Drama	0.336
15	Drama	0.329
16	Drama	0.290
17	Drama	0.415
18	Drama	0.297
19	Drama	0.354
20	Drama	0.339
21	Adult	0.818
22	Romance	0.352
23	Thriller	0.786
24	Drama	0.623
25	Short	1
26	Short	0.5
27	Musical	0.484
28	Short	0.987
29	Drama	0.364
30	Short	1

From the most dominant movies in each community, we think that “Drama” tend to be the most frequent dominant one across communities, because there 19 communities‘ most dominant genre is Drama. Then, we confirmed our guess by calculating the frequency across communities. Across communities, the most dominant genre is “Drama”, its frequency is 0.258

Answer for Question 8(b) :

We calculate the array c, p,q as defined in the statement, and then use ln(c(i))*p(i)/q(i) as the score to determine the most dominant genre.

Community No.	Most Dominant Genre	Max Score
1	Documentary	19.921
2	Mystery	11.369
3	History	21.364
4	Comedy	15.772
5	Western	32.928
6	Musical	14.011
7	Adult	313.736
8	War	32.034
9	Musical	22.222
10	Family	15.782
11	Family	27.849
12	Romance	9.678
13	Family	26.771
14	Action	49.325
15	Comedy	16.193
16	Adventure	28.313
17	Adult	12.536
18	Fantasy	14.160
19	Action	12.778
20	Adventure	37.302

21	Adult	118.032
22	Romance	17.040
23	Thriller	20.488
24	Action	15.771
25	Short	23.662
26	Short	9.175
27	Musical	103.958
28	Short	35.584
29	Sport	6.567
30	NA	0

The Community 30 is not applicable to use this score criteria, because there are only one movies in this community with genre information. Therefore, in Community 30 all the scores are all zeros. There are no dominant movie in this community. Compared to results in Question 8(a), there are some differences. In 8(a), the most dominant genre in more than half communities is “Drama”, but in 8(b), they are some genres other than “Drama”. The reason for that is we are taking the proportion of each genre in the whole dataset into account. “Drama” is the most frequent genre across the communities, which means there are many movies belonging to “Drama”. Then each community tends to have more movies belonging to “Drama”, but this does not mean the dominant movie should be “Drama”. For example, there are 30000 “Drama” movies and 200 “Short” movies in the whole dataset, and one community has 300 “Drama” movies and 200 “Short” movies, but this community is more important for “Short” instead of “Drama”, because it has all the “Short” movies. Besides, according to the definition of score, not only local frequency(frequency inside the community) and global frequency (frequency in the entire dataset) matter, but also the absolute number of the movies matters. These three contribute to a reasonable score to find the most dominant genres.

Answer for Question 8(c) :

We picked Community 23 with size of 14 in this question.

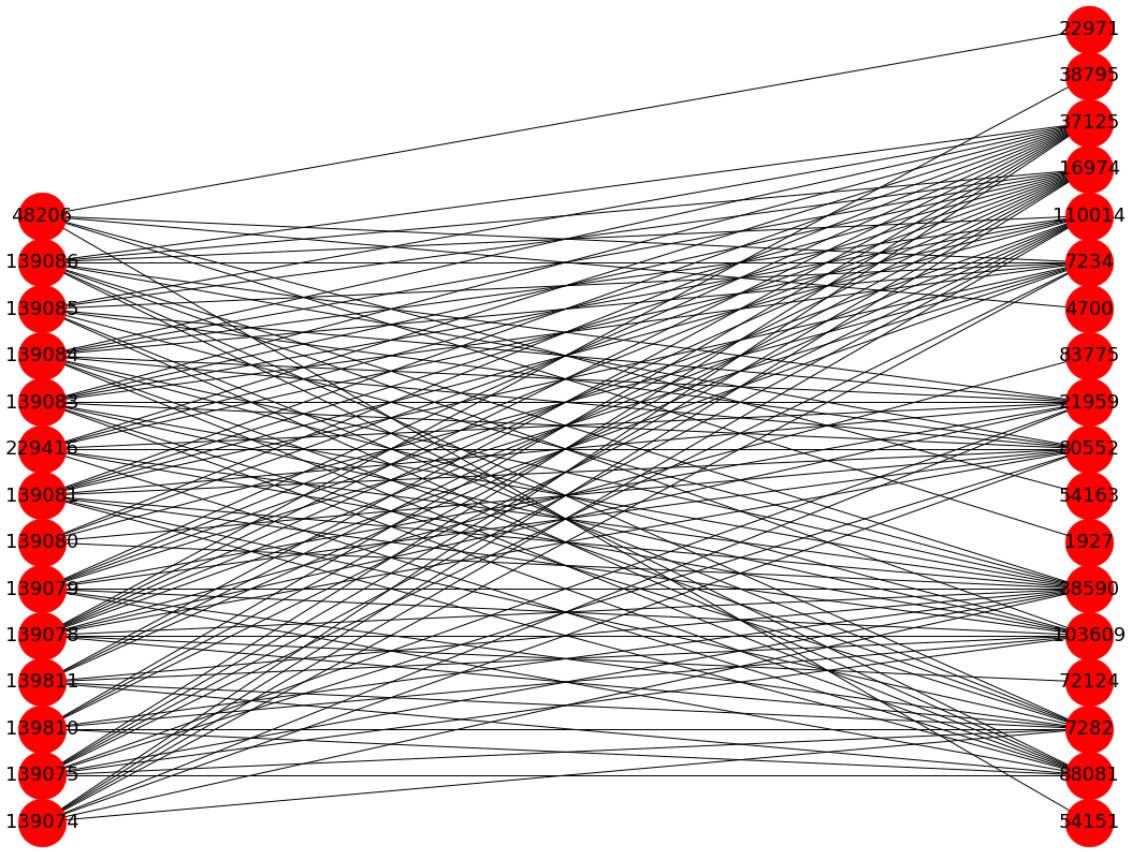
The movies in this community is as following: (their ids and names)

48206	Liverpool (2012)
139074	669: Escape the Reality (2011)
139075	An Olimatsim adventure (2011)
139078	Cent jours avant le lendemain (2015)
139079	L'affaire Hawkins (2014)
139080	La peur anonyme (2014)
139081	La Peur aux trousse (2015)
139083	Les oiseaux se cachaient pour mourir (2015)
139084	Midnight Stranger (2011)
139085	New York Vengeance (2013)
139086	October Sunset (2017)
139810	Des humains bien tranquilles (2016)
139811	Les années folles (2016)
229416	Mocakoma (2013)

Then we find all the actors who acted in these movies. These 18 actors as following: (their ids and names)

103609	Riel-Dery, Jessica
110014	Valin, Andréanne
16974	Desjardins, Nick
1927	Antaki, Joseph
21959	Fortin, Samuel (I)
22971	Gagné, David
37125	Lafond-Martel, Olivier
38590	Legros, Simon (I)
38795	Leonard, Joshua
4700	Beaulac, Sebastien
54151	Priest, Benoit
54163	Primeau, Marc
72124	Williams, Michael C.
7234	Boucher-L'Écuyer, Émile Pascal
7282	Bourassa-Simpson, Mathieu
80552	Charlebois, Jessica
83775	Donahue, Heather (I)
88081	Guimont, Mélanie

The corresponding bipartite graph is as following, we use the ids to represent to vertices, instead of the names for a clear view. The left side represents the movies, and the right side represents the actors.



We find the three most important actors based on the number of movies in this community they have acted. The three most important actors are:

16974	Desjardins, Nick
37125	Lafond-Martel, Olivier
38590	Legros, Simon (I)

They all acted in most movies compared to the other actors (13 movies) in this community. This is how they help to form this community. One actor acted 13 movies, which means that these 13 movies are all connected. If an actor acted in more movies, he will contribute more to the connections of the vertices and the density of the subgraph. Besides, we find that all these three actors' movies are in this community, which means they do not contribute to connection between any movie in the community and any movie outside the community. This is also help to explain how they help to form the community.

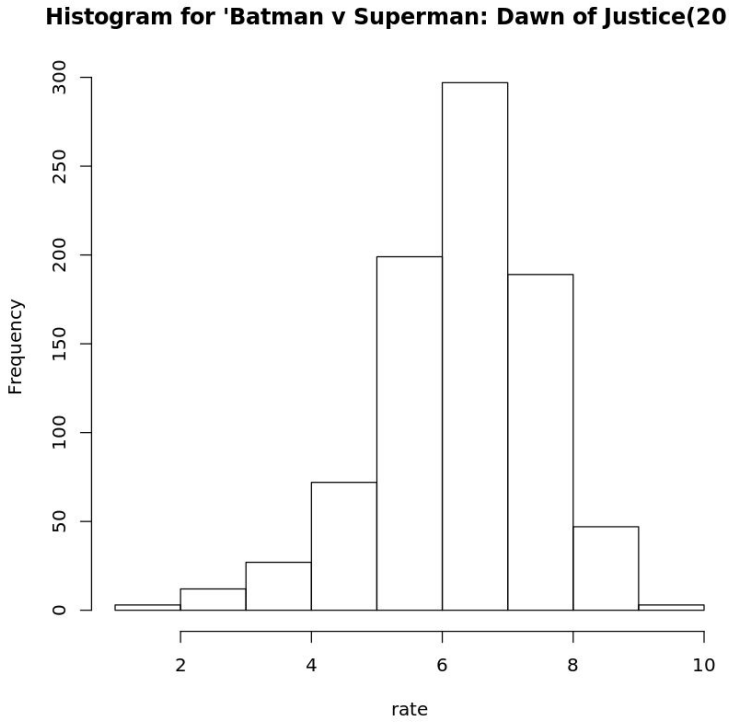
There is the correlation between these actors and the dominant genres you found for this community. The most dominant genres is "Thriller" both in 8(a) and 8(b). To find this correlation, we calculate the proportion of their movies with genre "Thriller" over all their movies in the entire data set. Then we get the proportion 0.71429,

0.55556, 0.58824, all larger than 0.5. Then, we can conclude the these actors acted a considerable proportion “Thriller” movie in all the movies they acted. The dominant genre in the community which they contribute most is very likely to be the same with the genre that most of their movies belong to.

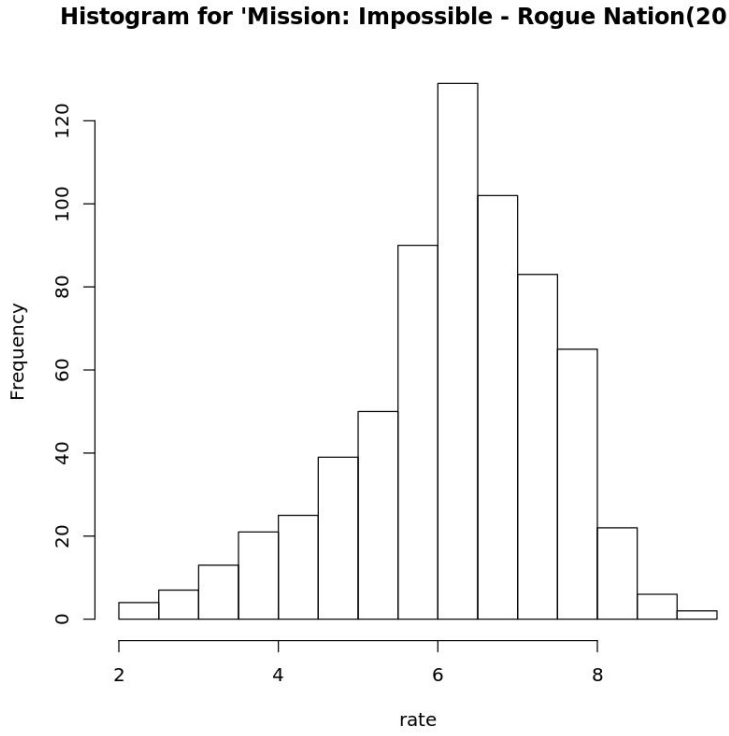
Answer for Question 9:

We firstly added ratings to the movie genre graph tree by loading “movie_rating.txt” file. Then, we extracted the neighbors of listed three movies from the movie genre rate graph and plot their distributions. The average rating of neighborhood of three movies are also calculated.

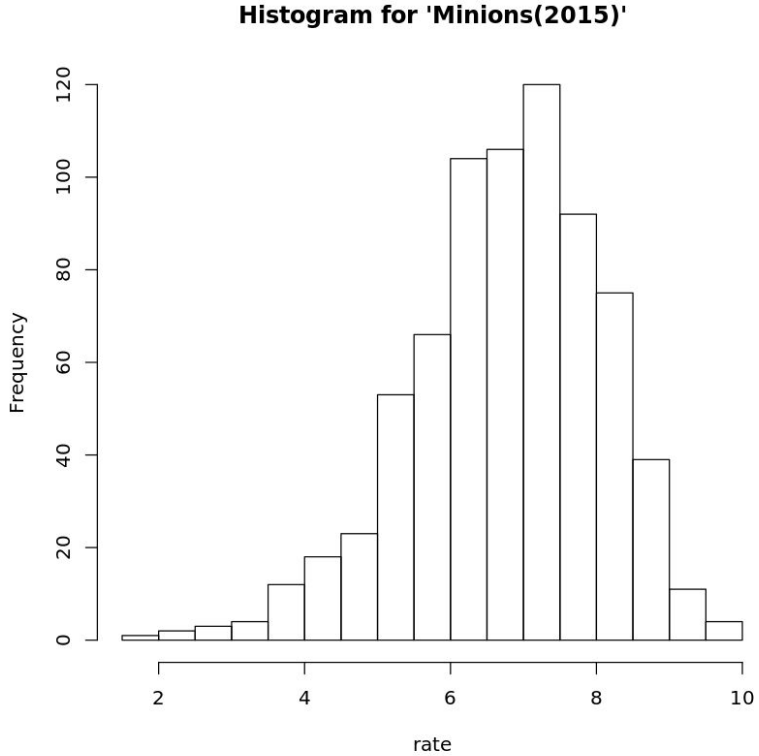
For movie “Batman v Superman: Dawn of Justice(2016)”, the average rating of its neighborhood is 6.33. Distribution of neighborhood rating is shown in following figure.



For movie “Mission: Impossible - Rogue Nation(2015)”, the average rating of its neighborhood is 6.23. Distribution of neighborhood rating is shown in following figure.



For movie “Minions(2015)”, the average rating of its neighborhood is 6.83. Distribution of neighborhood rating is shown in following figure.



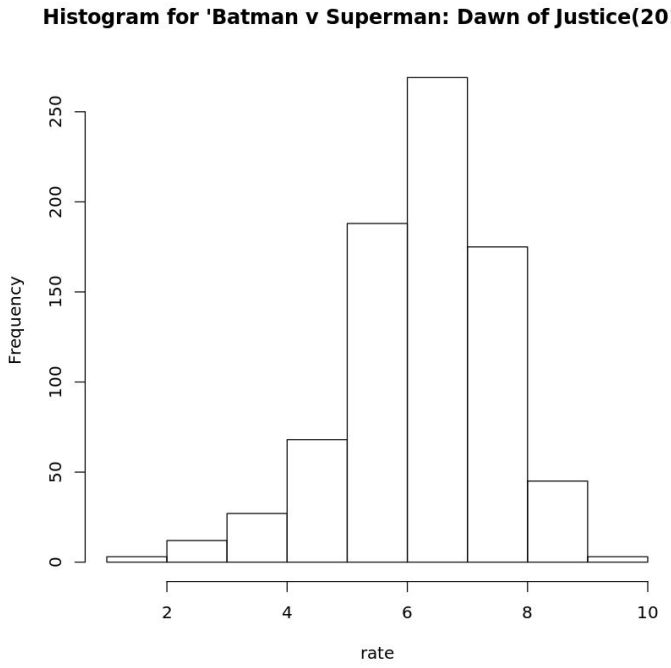
	Actual rating	Neighborhood average rating
“Batman v Superman: Dawn of Justice(2016)”	6.6	6.33
“Mission: Impossible - Rogue Nation(2015)”	7.4	6.23
“Minions(2015)”	6.4	6.83

From the distributions and average rating of neighborhood ratings, we didn’t find any obvious relation between neighborhood rating and actual movie rating.This may because we didn’t restrict community of neighbors and didn’t count the influence of weight between center and neighbor movies. In practice, the larger weight between two movie is, the more common features they should have.

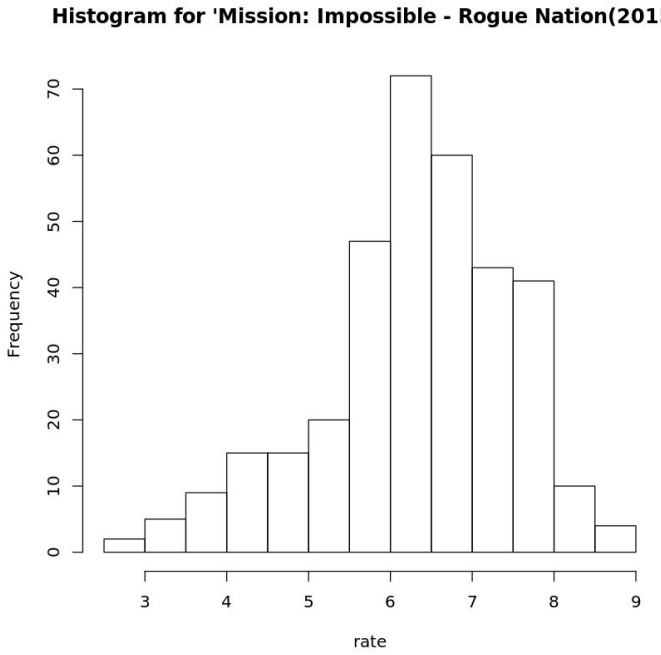
Answer for Question 10:

To avoid influence caused by different community and improve the accuracy of rating prediction by neighbor movies, we restrict the community of neighbors to be the same with center movie. This time, we got following average ratings and distribution.

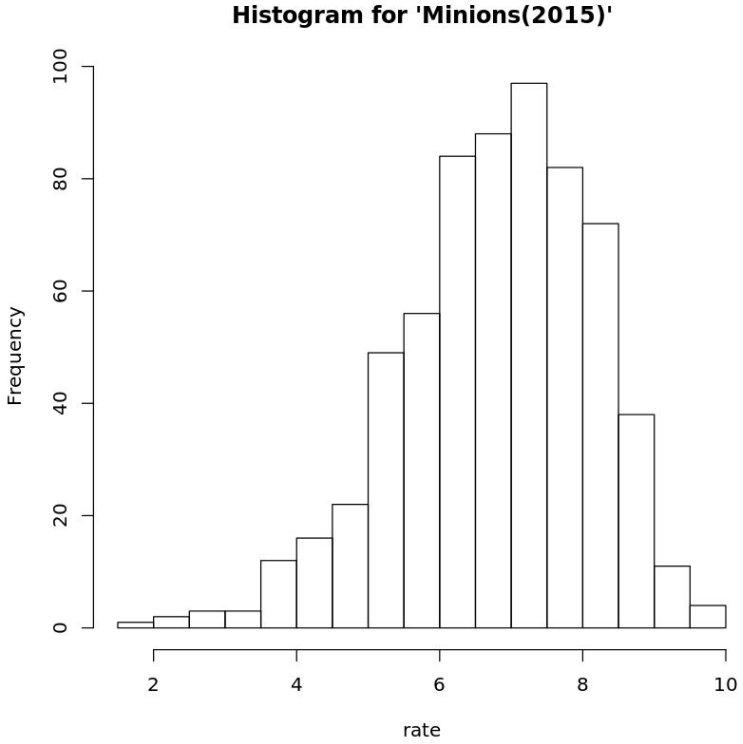
For movie “Batman v Superman: Dawn of Justice(2016)”, the average rating of its neighborhood is 6.31. Distribution of neighborhood rating is shown in following figure.



For movie “Mission: Impossible - Rogue Nation(2015)”, the average rating of its neighborhood is 6.37. Distribution of neighborhood rating is shown in following figure.



For movie “Minions(2015)”, the average rating of its neighborhood is 6.85. Distribution of neighborhood rating is shown in following figure.



	Actual rating	Neighborhood average rating
“Batman v Superman: Dawn of Justice(2016)”	6.6	6.31
“Mission: Impossible - Rogue Nation(2015)”	7.4	6.37
“Minions(2015)”	6.4	6.85

From the distribution and average ratings, we didn’t find any obvious relation between neighbors and center movie. From the prediction of three movies, we can’t see restricting neighborhood to one community changes prediction accuracy a lot, but it do increase rating for the second movie.

Answer for Question 11:

To extracted 5 top neighbors from movie neighborhood rate graph, we firstly output all weights between center movie and its neighbors and sorted them in descending order to find the top 5 movie weights. Then we look for movies whose weight is equal to those weights. By this method, we extracted the top 5 neighbors of listed movies.

For movie “Batman v Superman: Dawn of Justice(2016)”,the top 5 neighbors ids and their community are

Movie id	26903	12643	40746	11623	4341
community	1	1	1	1	1

All community of top 5 neighbors are the same with the neighbor of movie “Batman v Superman: Dawn of Justice(2016)”.

For movie “Mission: Impossible - Rogue Nation(2015)”,the top 5 neighbors ids and their community are

Movie id	40100	40106	72241	87094	48391
community	13	13	1	1	1

All community of top 3-5 neighbors are the same with the neighbor of movie “Mission: Impossible - Rogue Nation(2015)”, while the first two neighbors are different.

For movie “Minions(2015)”,the top 5 neighbors ids and their community are

Movie id	46240	20421	46204	65511	77059
community	1	1	1	1	1

All community of top 5 neighbors are the same with the neighbor of movie “Minions(2015)”.

From the results, most community of top 5 neighbors are the same with center neighbors, which indicates the importance of weights between two movies. High weights usually mean high revalences.

Answer for Question 12 :

Based on previous part, we want to do some prediction job. We try to predict the ratings for three movies which has index of ('12596', '48390', '100855'). Firstly, we use a normal linear regression model to do prediction. We define our features in this way. Normally, people will see a movie because of some famous actors or actresses. Most movie will have two main male roles and two main female roles. So we pick the highest four pagerank points as features. We expect that if the scores are higher, we will earn a higher ratings.

Another thing is that in order to increase the computing speed and avoid overfitting. We randomly pick the data points as our subset. We respectively get 1/1000, 1/500, 1/100, 1/50 parts of the total set and get the result as following.

	12596	48390	100855	RMSE
	rating	rating	rating	
1/1000	6.14015	6.139553	6.157364	1.104741
1/500	6.013111	6.047674	6.118591	1.264422
1/100	6.13774	6.099596	6.186541	1.238669
1/50	6.107414	6.125062	6.152376	1.226467

From the result we can observe that the result differs not to much as the part goes larger, the RMSE increases at first and then goes a little down. It distributes near 1.2.

Answer for Question 13 :

In this question, we take another strategy to make prediction. We build a bipartite graph to observe the relationship between the actors/actresses and movies. The procedure includes the following steps:

1. Get the involved actors/actresses according to the three movies
2. Get the involved movies from the list of actors/actresses we had already
3. Build bipartite graph between them.

We avoid building a huge bipartite graph in this way. Because only those involved actors and actresses are valuable.

We define the weight as the average rating over all movies of a certain actor/actress. Therefore, for every actor/actress, we have a rating value. When making prediction, we average rating over all actors/actresses of each movie to get result.

```
predict 6.362715
predict 6.424666
predict 6.856822
```

We use the same subset as the Q12 of 1/500 division. For each data points, we have to build a specific bipartite graph to make prediction and make comparison with its original rating to compute the final RMSE.

We take the smallest bipartite graph as an example. If there are too many actors, it is hard to observe the whole graph.