EE232E Project 4

IMDb Mining

Hongyan Gu – 205025476

Xin Liu – 505037053

Aoxuan (Douglas) Li - 905027231

Jiawei Du - 404943853

In this project, we studied properties of Internet Movie DataBase (IMDB). In part 1, we describe the directed actor-actress network, and reported the actor pairing and ranking using page rank algorithm. Undirected network is presented in Part 2, in which community detection, neighborhood detection and rating predictions are studied.

Part 1

In this part of the project, we created a directed network using the data from actor/actress list files and studied the properties of the network.

Question 1

Before we construct the actor-actress network, necessary cleaning and filtering of the raw data should be made. As required in problem statement, we first eliminate the duplication of movie names in various actor/actress lists. Regular expressions are used to pair and eliminate the characters in brackets, i.e. any character lies in "()" are deleted. However, we skipped the brackets with four digits, since it indicates the year of movie published and could distinguish various movie shot in different times. What's more, "{{SUSPENDED}}" string in the movie names are also deleted.

We also filtered the actors by the total number of movies. Actors that are in less than 10 movies are not in consideration in this project.

We also note that there are three duplicated actor names "de la Torre, Antonio (I)", "McKinney, Bill (I)", "Shin, Sung-il" in raw material. This can cause many problems while generating graphs. After looking up through the Internet carefully, there are no two distinct people named the same and those 3 repeated names in the actor/actress list can be deleted.

After cleaning procedure described above, the total number of actor/actresses is 113132 and they have been in 455917 movies.

Question 2

The in-degree distribution of the network is shown as below:

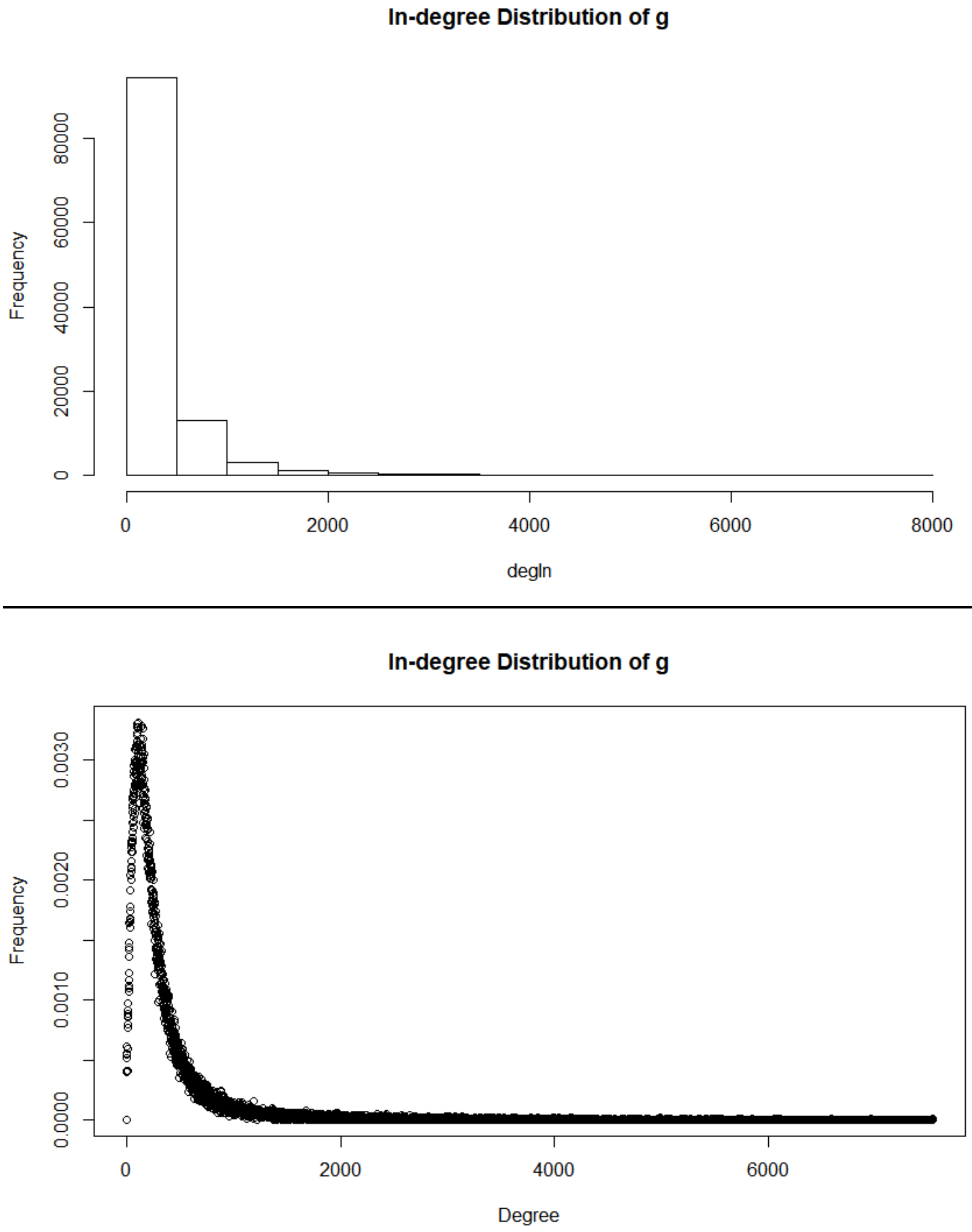

Fig. 1 In-degree distribution of the network

From the plot, we could see that most of the nodes have a degree of 200-1000, which is also a reasonable number for the average level of an actor/actress who shares the common movies with another actor/actress. The highest frequency of the distribution is around 0.0035. The maximum number of the in-degree of an actor is over 7000, which means these particular actor/actresses are very experienced and had participated in many movies. The whole distribution is like a Poisson distribution which has a long tail.

<u>Question 3</u>

In this question, we wanted to find the most wanted actor/actress to work with for a given actor/actress. We decided to take the target with the highest weights of edge for a given actor/actress. The results are shown below:

Table. 1 Actor/actress pairs

Input Actor/Actress	Output Actor/Actress	Weights of edges
Tom Cruise	Nicole Kidman	0.1746032
Emma Watson (II)	Daniel Radcliffe	0.52
George Clooney	Matt Damon	0.119403
Tom Hanks	Tim Allen (I)	0.1
Dwayne Johnson	Steve Austin (IV)	0.2051282
Johnny Depp	Helena Bonham Carter	0.08163265
Will Smith(I)	Darrell Foster	0.122449
Meryl Streep	Robert De Niro	0.06185567
Leonardo DiCaprio	Martin Scorsese	0.1020408
Brad Pitt	George Clooney	0.09859155

The results did not go beyond our expectations. As the weight of edges is positively relative to the number of movies the two actors/actresses cooperate, bigger weights indicate that the two actors/actresses are more likely to cooperate, so that they may be wanted pairs for each other. From the result, for example, the output actor for Emma Watson is Daniel Radcliffe. In the series of Harry Potter movie, Emma Watson acted the role of Hermione Jane Granger, and Daniel Radcliffe took the part of Harry Potter. Not surprisingly, they had deep cooperation, and would rationally be the best pairs. Additionally, the weight of their edges is much more than others.

Another example is the famous actor Leonardo DiCaprio and the famous director Martin Scorsese. From Wikipedia, Director-actor duo Martin Scorsese and Leonardo DiCaprio have frequently collaborated, making a total of five feature films and one short film since 2002. The pair's films explore a variety of genres, including crime, thriller, biopic and comedy. Several have been listed on many critics' year-end top ten and best-of-decade lists. The duo's films have been nominated for thirty-one Academy Awards, winning nine. In 2013, the duo was awarded National Board of Review Spotlight award for career collaboration. Scorsese's work with DiCaprio is considered to be as vital as his work with Robert De Niro. Not surprisingly, they could be best pairs.

Question 4

Table. 2 Page rank score, number of movies and in-degree for the top ten actors/actresses

Actor/Actress	Page Rank Score	Number of Movies	In-degree
Roberts, Eric (I)	0.000109269	298	3898
Tatasciore, Fred	9.803026e-05	355	3986
Jeremy, Ron	9.771786e-05	637	3019
Trejo, Danny	9.482154e-05	241	3597
Flowers, Bess	9.352609e-05	828	7537
Hitler, Adolf	8.706432e-05	379	3873
Riehle, Richard	8.650928e-05	197	3281
Harris, Sam (II)	8.561108e-05	600	6960
Jackson, Samuel L.	8.507126e-05	159	3462
De Niro, Robert	8.182877e-05	138	3434

From the result, we can see that all the names listed above with high page ranks are not listed in the actor/actress list of question 3. For me, most of the names are whom I'm not familiar with (maybe except Adolf Hitler). However, this is pretty common for them to get a high page rank. The page rank algorithm is the calculation of importance of page and is related to the number of other important pages pointing to them. By simple search on Google, we could find without difficulties that the names listed above are all big names in Filmmaking or other movie-related industries.

For example, from Wikipedia, Bess Flowers (November 23, 1898 – July 28, 1984) was an American actress best known for her work as an extra in hundreds of films. By some counts considered the most prolific actress in the history of Hollywood. She was known as "The Queen of the Hollywood Extras," appearing in more than 350 feature films and numerous comedy shorts in her 41-year career.

All the names listed above did great contributions and exerted deep influences on film industries. Most of them are Oscar Nominations, so that they were discussed widely among social medias now or in the past, and as a consequence, they gained much more page importance. By comparison, the actors/actresses listed in question 3 are almost all the famous stars in a young generation. Although they are famous indeed, it is still hard for them to compare with those super names in older generations with the page ranks.

Question 5

The results for the ten actor/actresses listed are shown below:

Table. 3 Page rank score, number of movies and in-degree for the ten actors/actresses

Actor/Actress	Page Rank Score	Number of Movies	In-degree
Tom Cruise	4.176239e-05	63	1681

Emma Watson (II)	1.278372e-05	25	453
George Clooney	3.951519e-05	67	1575
Tom Hanks	4.940027e-05	80	2064
Dwayne Johnson	3.585114e-05	78	1360
Johnny Depp	5.356372e-05	98	2168
Will Smith(I)	3.477341e-05	49	1384
Meryl Streep	3.950573e-05	97	1602
Leonardo DiCaprio	3.519180e-05	49	1379
Brad Pitt	4.545450e-05	71	1799

From the results, we could see that these actor/actresses who are famous for young generations have a reasonably high page rank score, number of movies and in-degree in the network. However, compared to the top ten actor/actresses listed in question 4, the metrics for these actors are relatively low. Notice that the page rank score for a page is largely depending on the influence of it to the whole network, that is to say, if an actor/actress have more number of movies or higher in-degree, they are possibly to have higher page rank scores. To see this, among the actors/actresses in the list above, Johnny Depp has the highest rank score, and in the meantime, the number of movies he participates in and the in-degree of his node are also the most. In spite of this, he still cannot even compare to De Niro, Robert, who is the tenth page ranking score owner. That's why the listed moviers for young generations still could not compare with those famous and influential moviers.

Part 2

In this part, we created an undirected movie network and then explore the various structural properties of that network.

Question 6

To create a movie network, we first have to convert the actor-movie list to movie-actor list, and chose the movie that has at least 5 actors in. The total amount of movies is 203480. As described in the problem statement, the weight between edges is calculated as the formula below.

$$w_{i \to j} = \frac{A_i \cap A_j}{A_i \cup A_j}$$

Fig. 2 shows the degree distribution of network degree and density curve with with mean plotted as dashed line. The mean of the degree is approximately 656. From the plot we can see that the majority movies don't have degree more than 2000, and the degree distribution is more like a Poisson distribution with a long tail on the right. For the meaning in real cases, the more degree a movie node have, the more likely the movie has a large cast and has famous actors acting in. This distribution indicates that only a very small portion of the movies are actually have a large investment, while most of the movies are acted by less, not-very-famous actors.

The movie that has the largest degree is called “Mr. Smith Goes to Washington (1939)”. It has 201 actors and has a very large impact on US film industry. The film won the 11th academy award and gained more than $9 million from box office in the 1930s, which is a lot amount of money in that age. All its actors acted in 14701 distinct movies, which is a bit more than its actual degree in the network, 13331, since we filtered the movies with less than 5 actors.

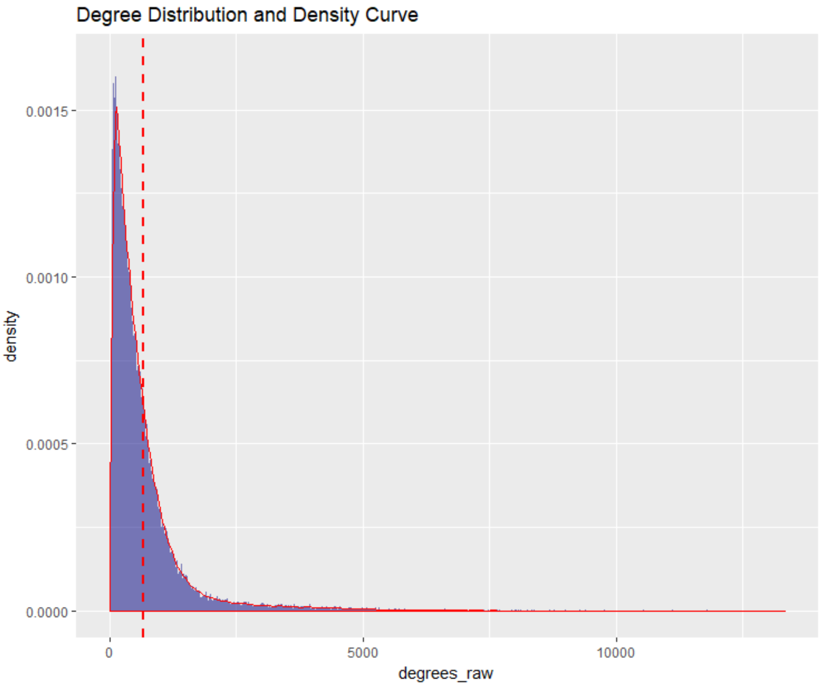

Fig. 2 Degree distribution and Density Curve (with mean plotted as dashed line)

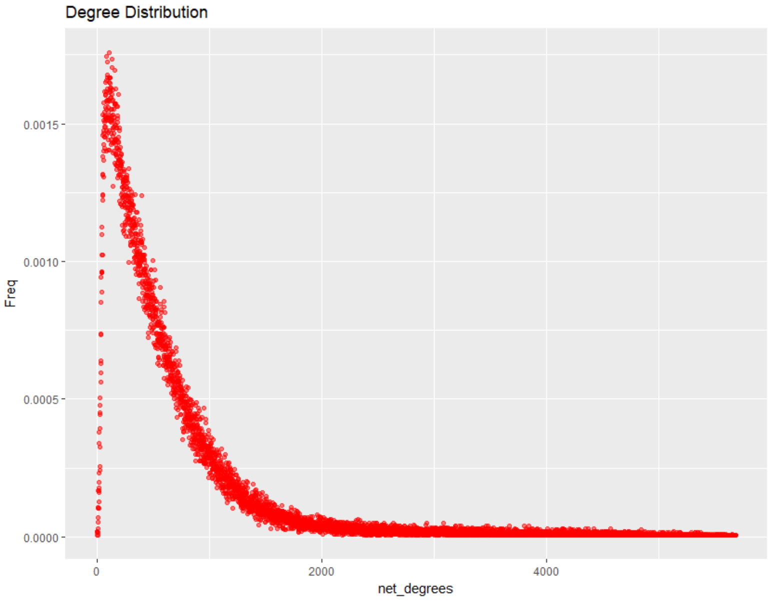

Fig. 3 Digree distribution of the movie network (scatter plor with probablity)

Qusestion 7

In this section, the communities of the movie network are detected using the Fast Greedy community detection algorithm. As a result, 29 communities are found, and their sizes are varied. Communities 1, 2, 15,17, 21, 23, 24, 25, 26, and 29 are selected as examples, and their network graphs and distribution of the genres are plotted.

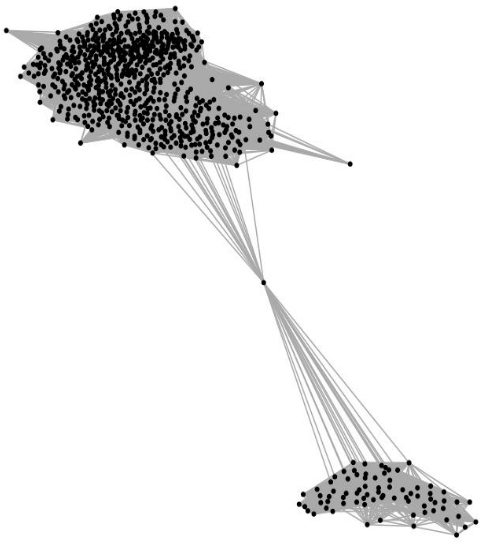

Fig 4. Graph Community 1

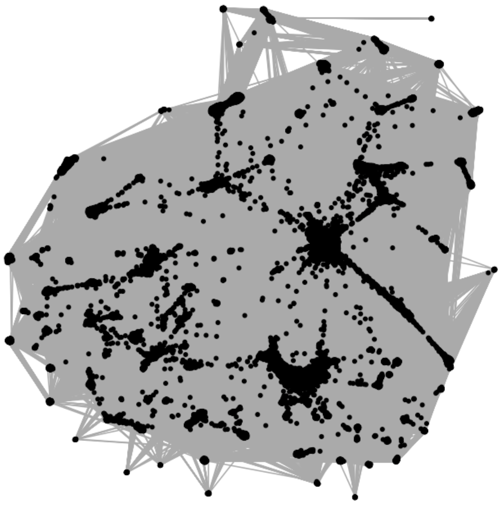

Fig 5. Graph Community 2

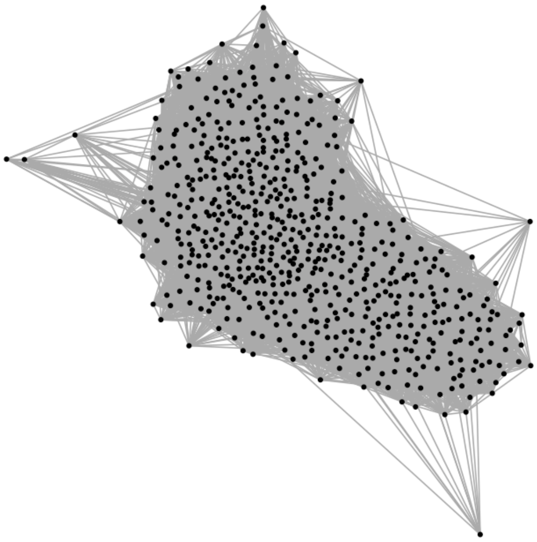

Fig 6. Graph Community 15

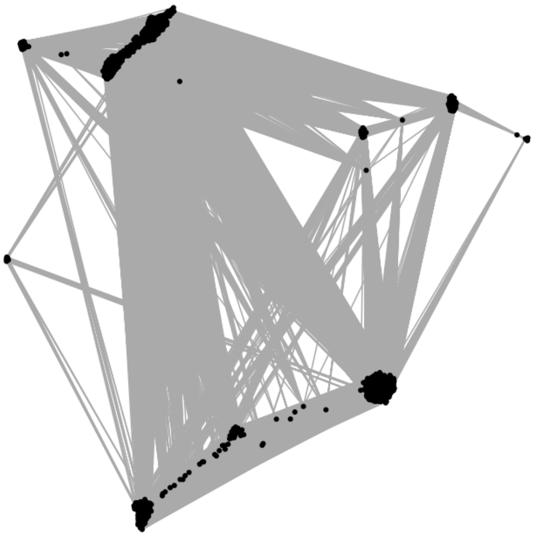

Fig 7. Graph Community 17

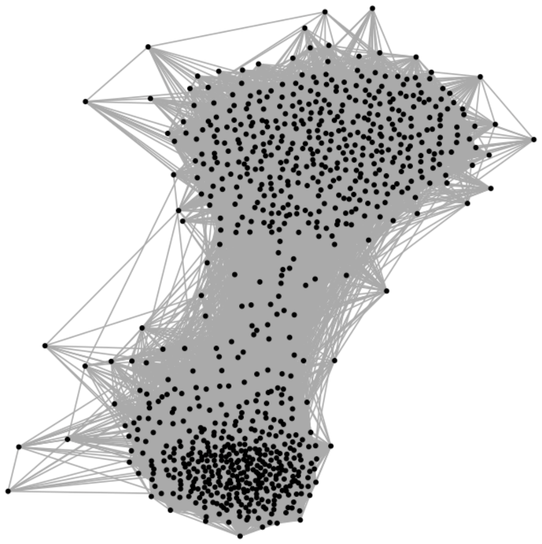

Fig 8. Graph Community 21

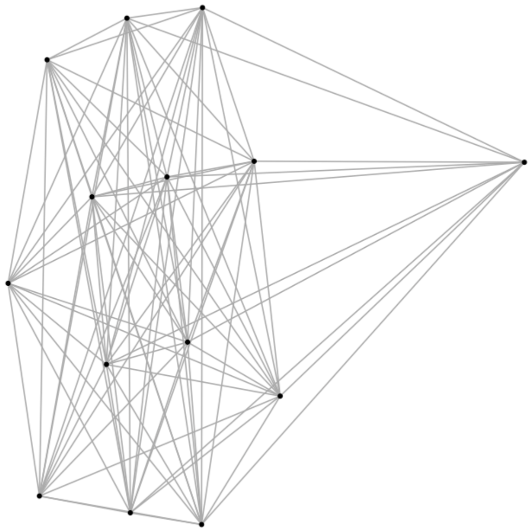

Fig 9. Graph Community 23

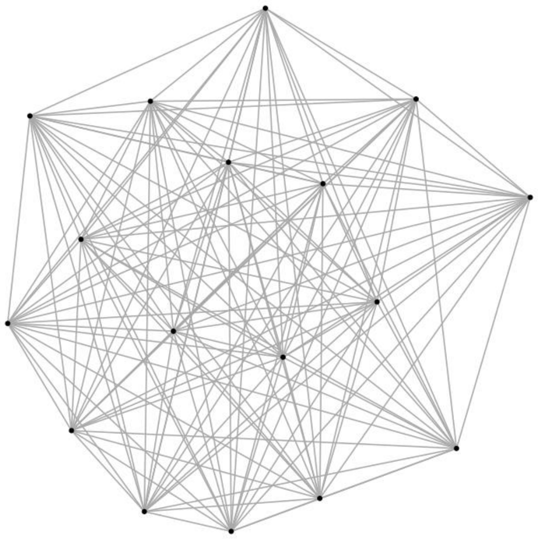

Fig 10. Graph Community 24

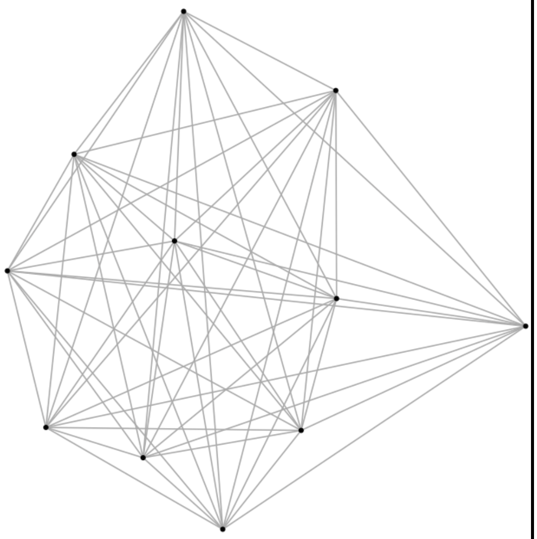

Fig 11. Graph Community 25

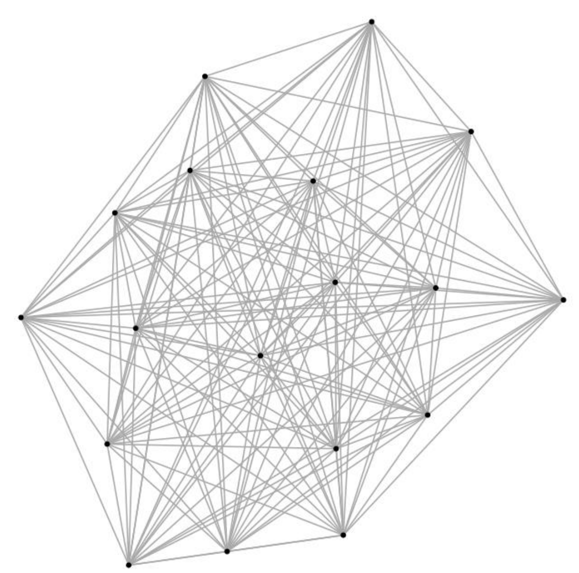

Fig 12. Graph Community 26

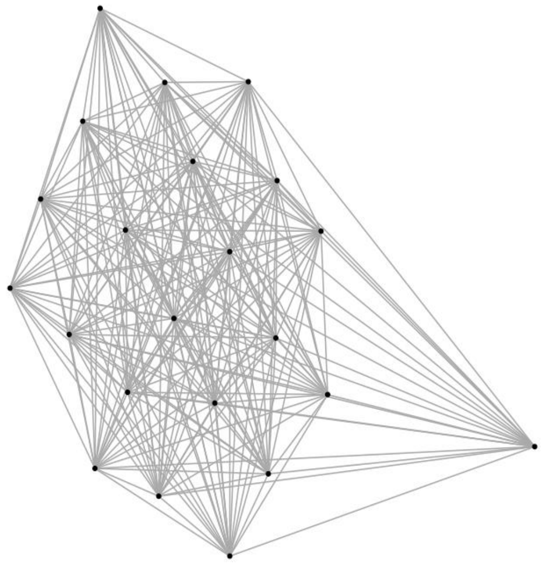

Fig 13. Graph Community 29

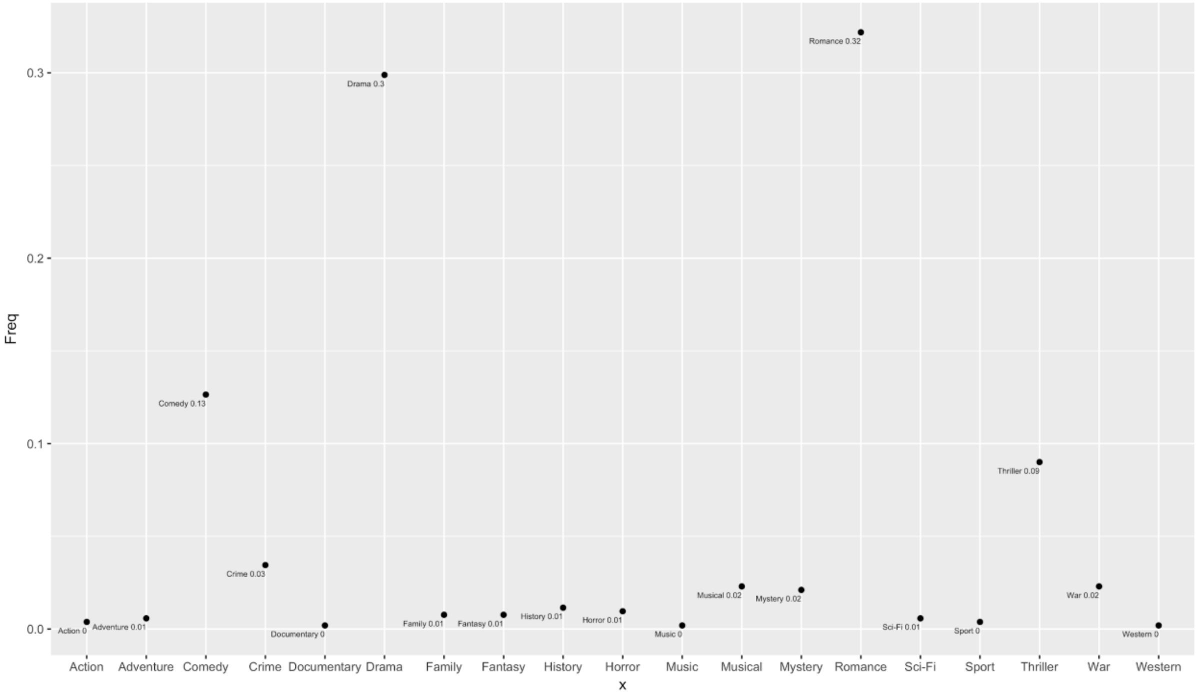

Fig 14. Distribution of the genres Community 1

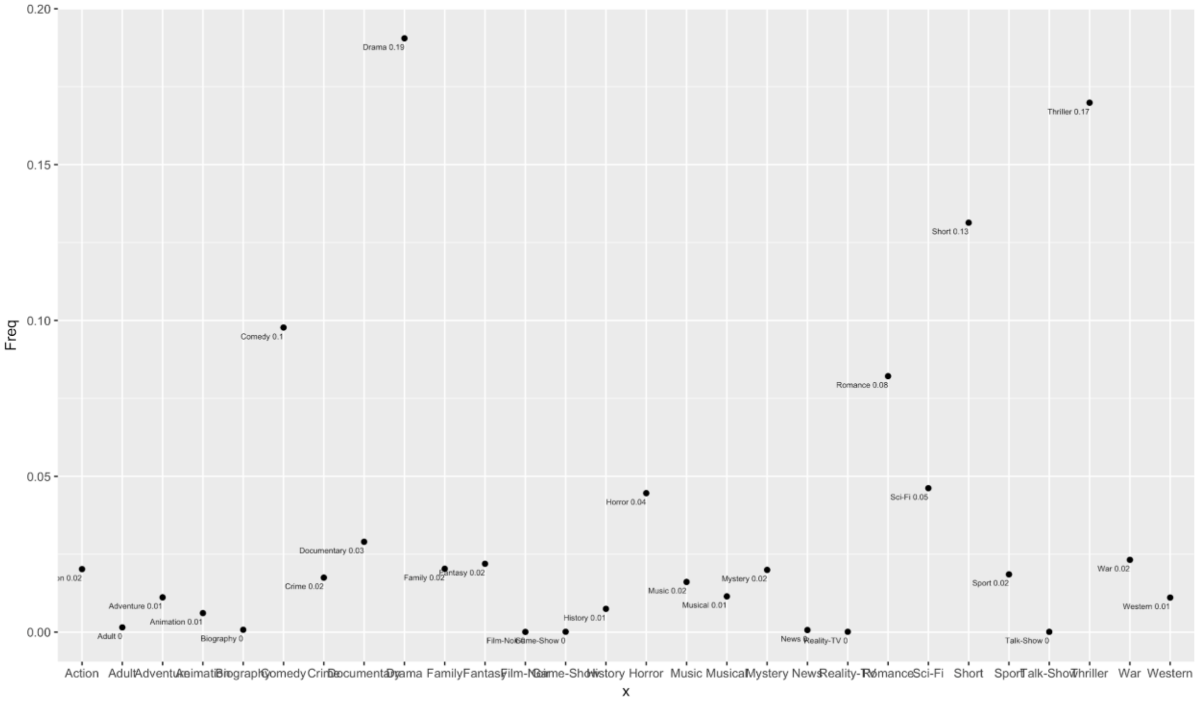

Fig 15. Distribution of the genres Community 2

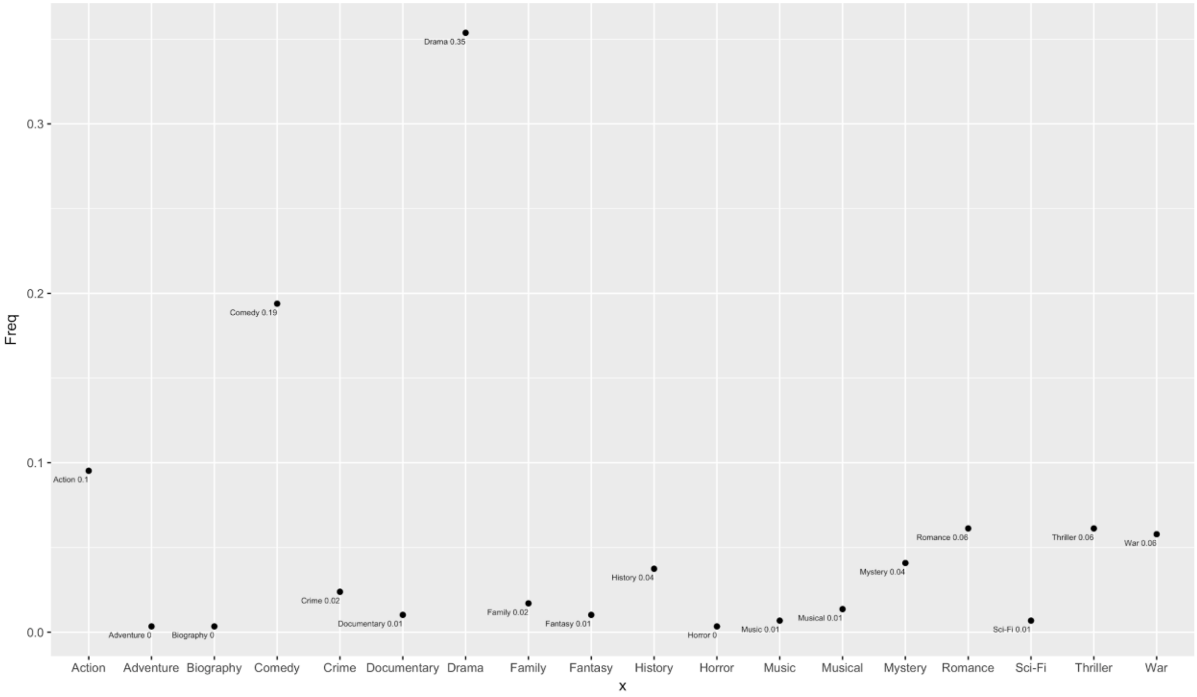

Fig 16. Distribution of the genres Community 15

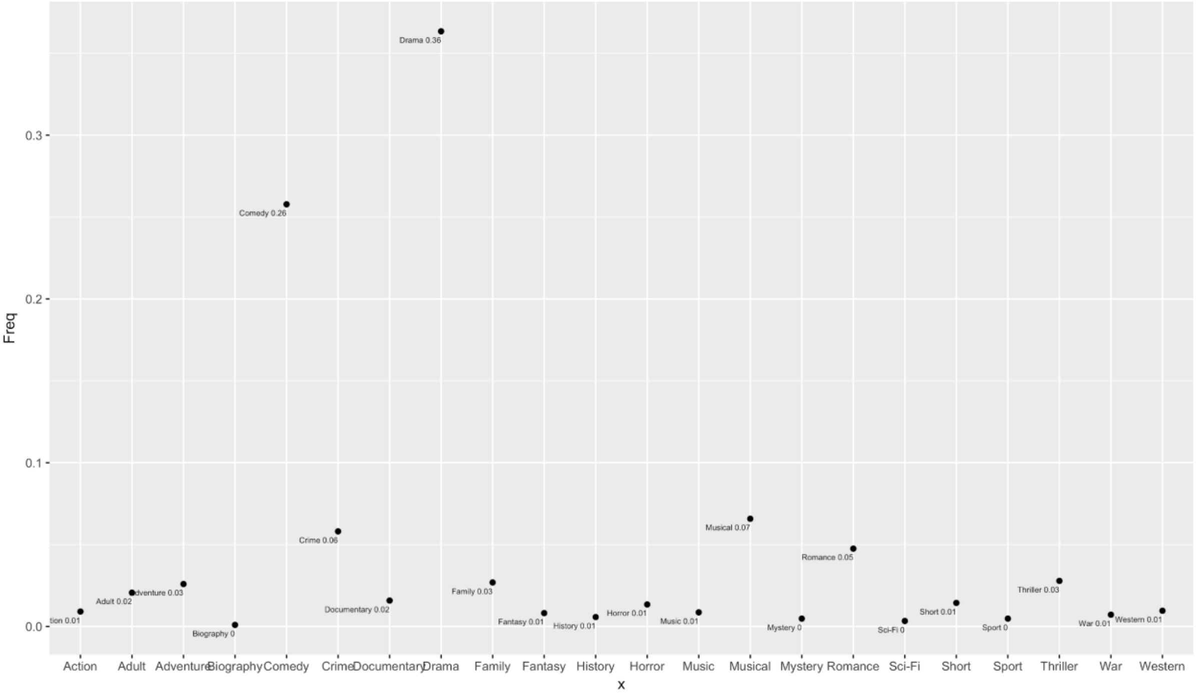

Fig 17. Distribution of the genres Community 17

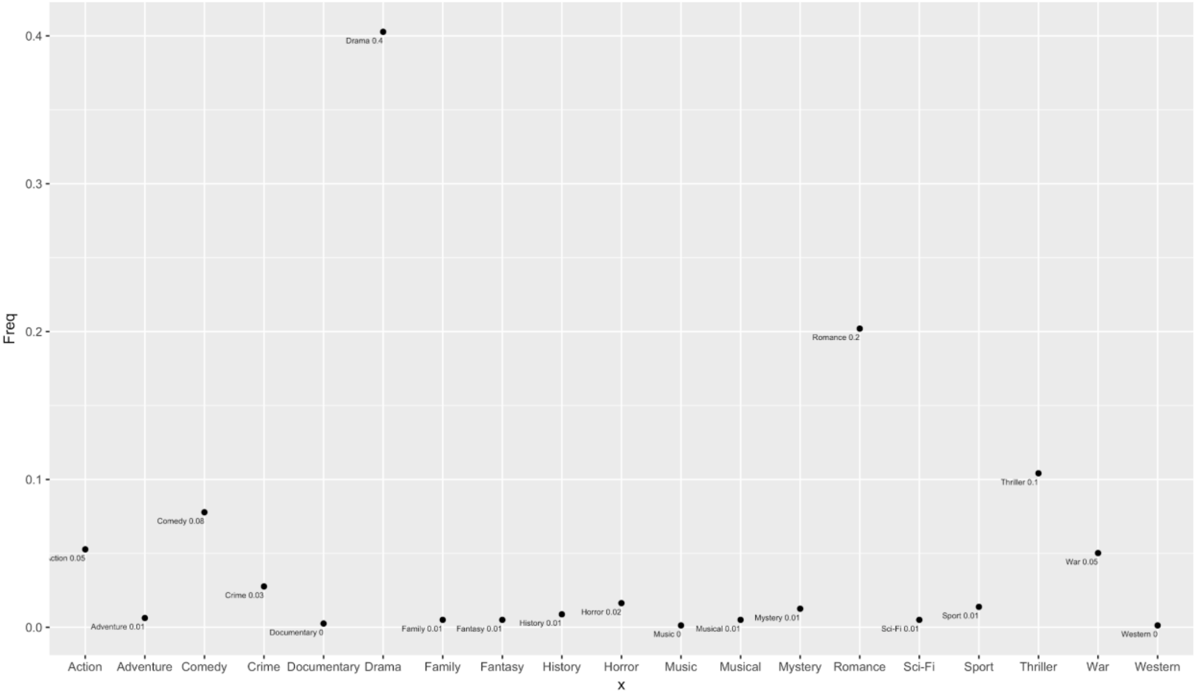

Fig 18. Distribution of the genres Community 21

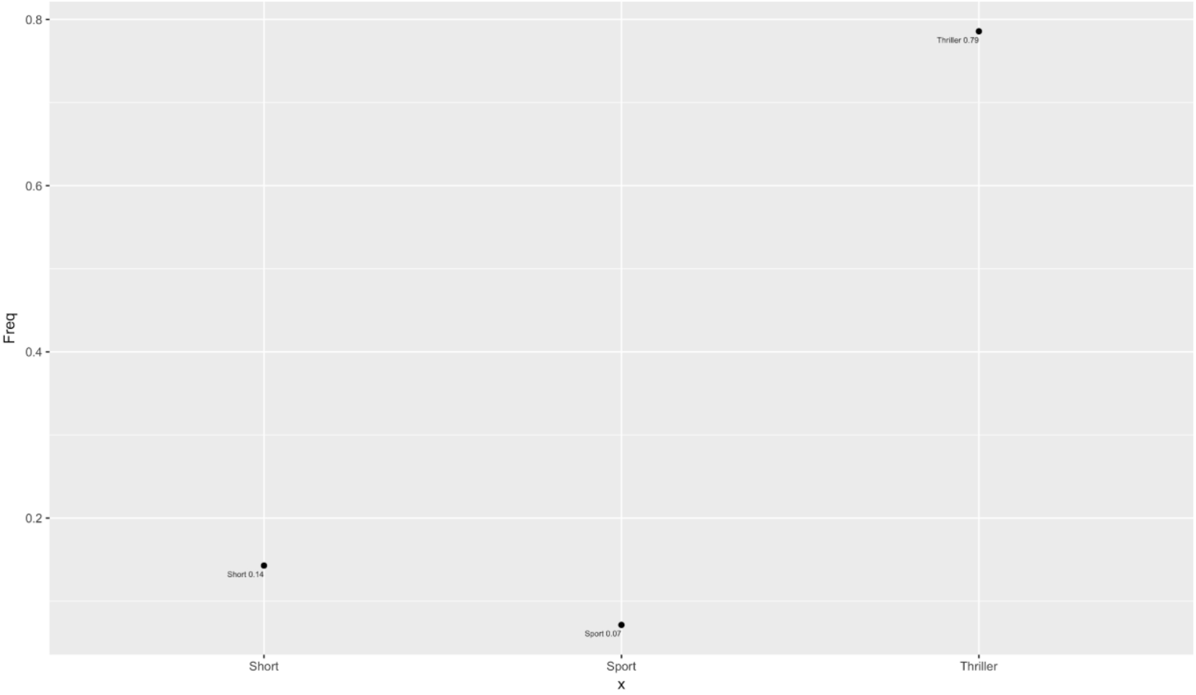

Fig 19. Distribution of the genres Community 23

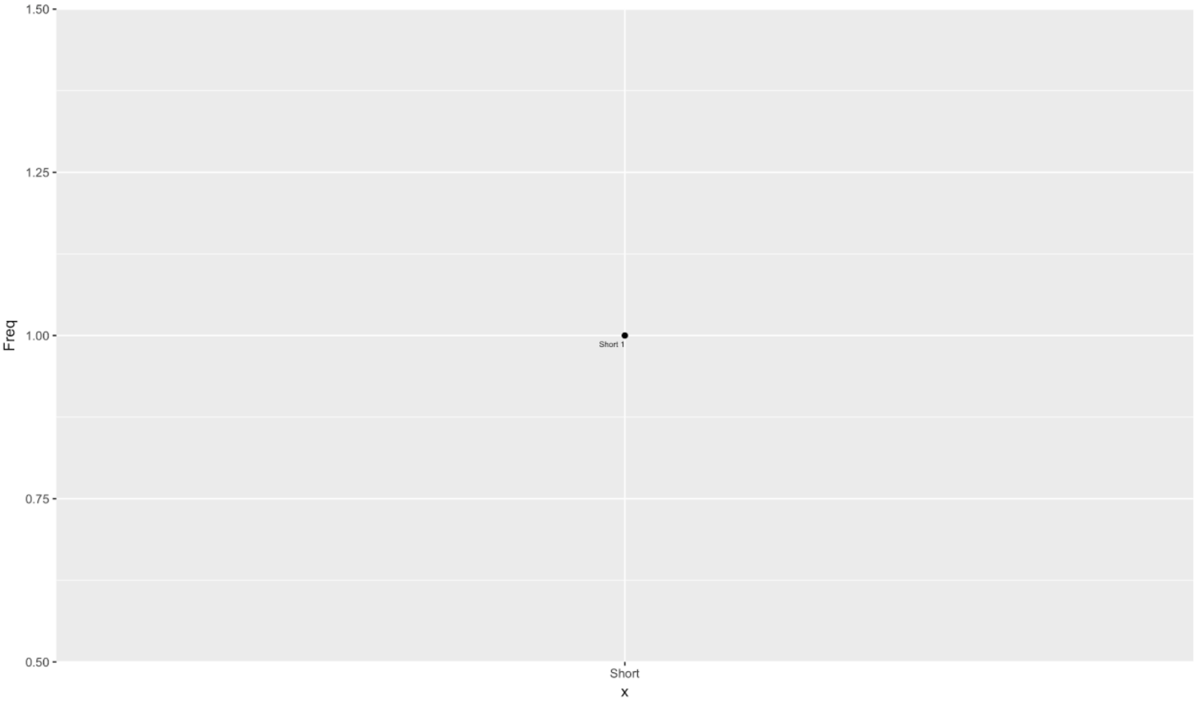

Fig 20. Distribution of the genres Community 24

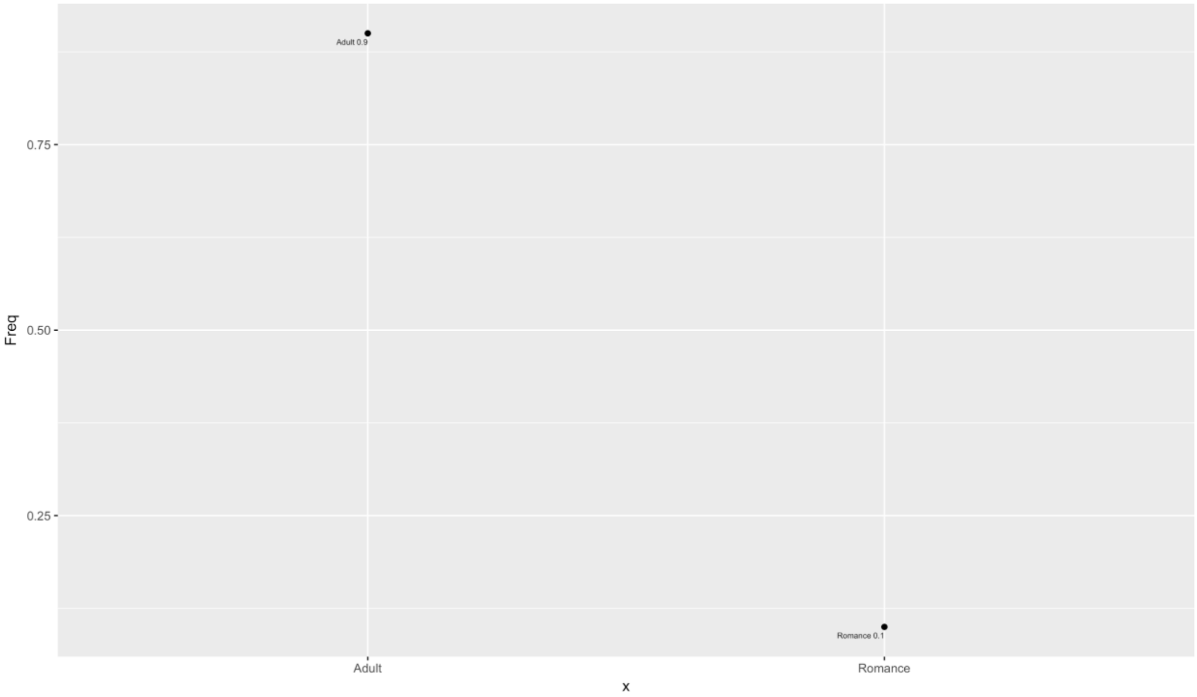

Fig 21. Distribution of the genres Community 25

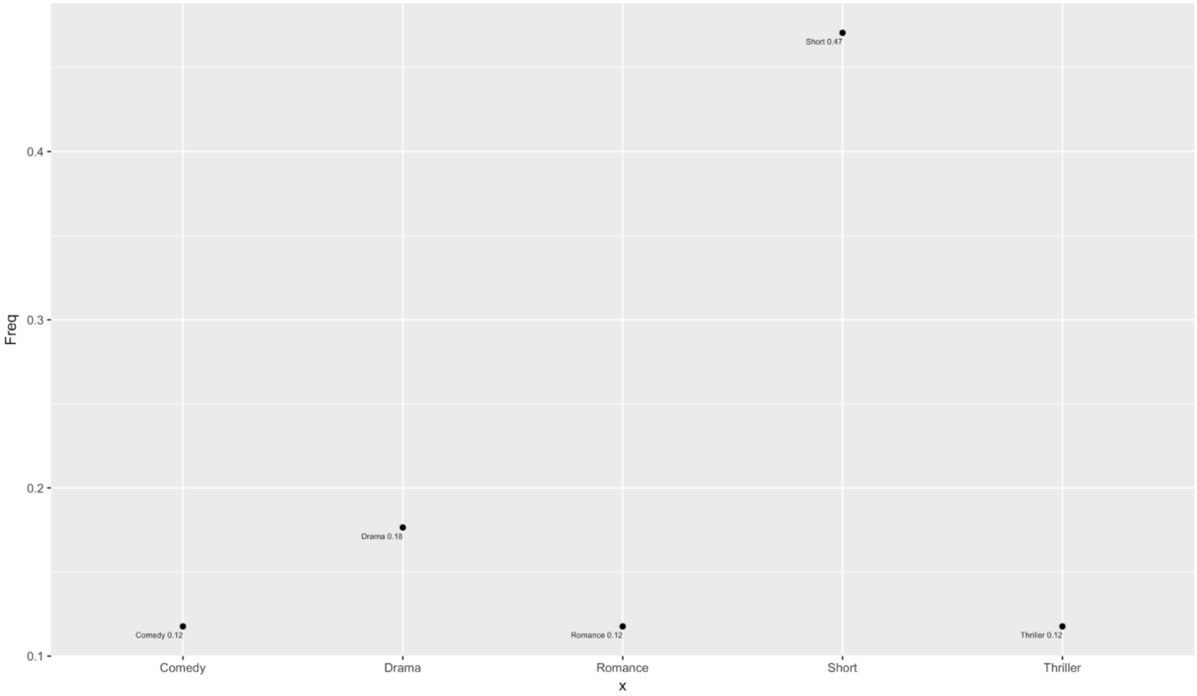

Fig 22. Distribution of the genres Community 26

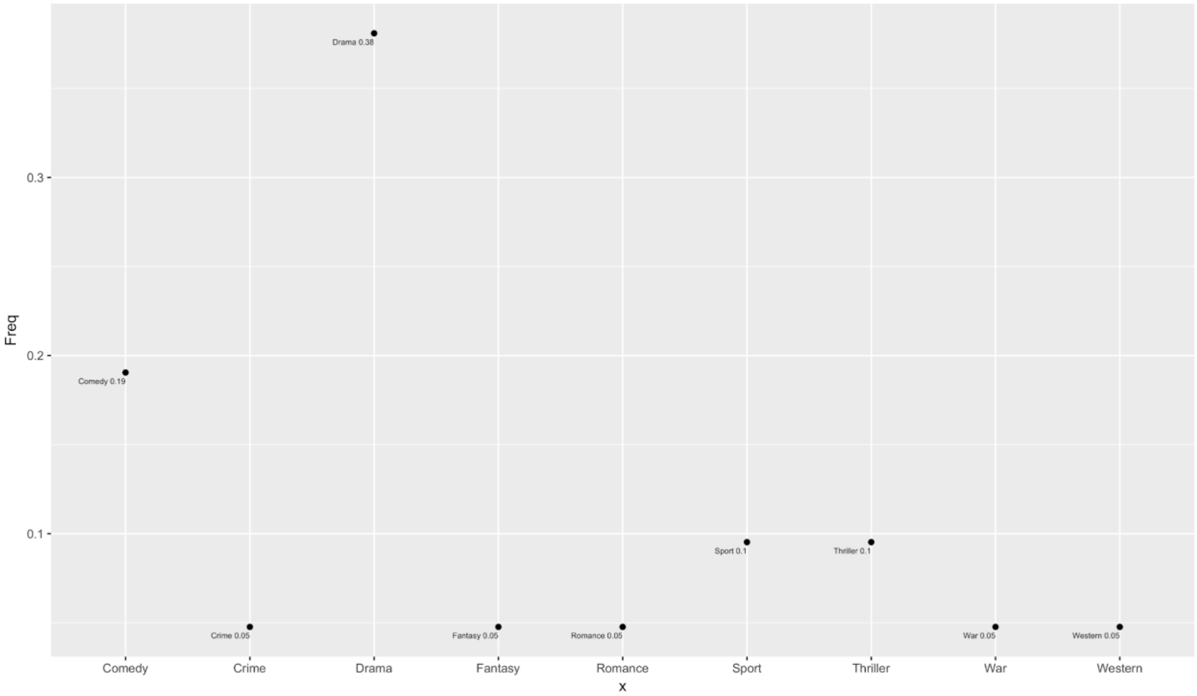

Fig 23. Distribution of the genres Community 29

In terms of the genres distribution, for each communities, there is/are one or two genres having relatively high frequency while frequencies of other genres are near or low than 0.1. Based on a comparison of different communities, e.g. community 1, whose community graph is extremely unbalanced, and community 26, a nearly complete graph, qualitatively, the variance of a distribution on genres is relatively high when the underlying graph is nearly fully connected. This result corresponds with the common sense in reality world, i.e. a actor is more likely to attend certain type of movie; therefore, if most movies share same actors, there is a high possibility they from the same genre. On the contrary, in community 1, corresponding with two sub-communities in the graph, its genres distribution has two peak value.

Question 8(a)

In this section, the most dominant genre based simply on frequency counts for each community is determine. From Table 4, it can be claimed that most community has dominant genre drama. The second most frequent genre is short. The result corresponds with the fact that most movies are drama in real world.

Table 4. Most Dominant Genre Based on Frequency Counts

Community	**Genre**	**Frequency**
1	Romance	168
2	Drama	8908
3	Drama	3468
4	Drama	9360
5	Drama	679

6	Drama	234
7	Drama	1868
8	Drama	1599
9	Drama	683
10	Drama	1779
11	Short	9077
12	Adult	2134
13	Drama	1614
14	Drama	2417
15	Drama	104
16	Drama	1068
17	Drama	757
18	Drama	185
19	Drama	609
20	Drama	2902
21	Drama	321
22	Short	1496
23	Thriller	11
24	Short	17
25	Adult	9
26	Short	8
27	Short	76
28	Family	18
29	Drama	8

Question 8(b)

In this section, the most dominant genre based on a score given by ln(c(i))*p(i)/p\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. The result has significant difference with result in 8(a): (1) The genres are varied and nearly uniform distribution, and (2) if a community has dominant genre other than drama in 8(a), it may also has the same dominant genre. while if a community has dominant genre drama in 8(a) the result may differ.

Table 5. Most Dominant Genre Based on Score

Community	**Genre**	**Score**
1	Romance	53.32
2	Sci-Fi	32.58
3	War	37.39
4	Comedy	37.93
5	Comedy	40.48

6	History	28.24
7	Western	41.71
8	Family	41.94
9	Musical	61.27
10	Adventure	57.49
11	Film-Noir	162.63
12	Adult	73.35
13	Action	96
14	Drama	29.85
15	Action	25.96
16	Adventure	77.29
17	Crime	42.19
18	Action	26.29
19	War	68.18
20	Romance	56.58
21	Romance	33.19
22	Short	13.04
23	Thriller	34.35
24	Short	6.38
25	Adult	31.63
26	Romance	2.64
27	Short	9.51
28	Family	195.96
29	Drama	7.31

Question 8(c)

In this section, the bipartite graph of the community 26, which contains 18 movies, are plotted. All the actors are Joiner, Craig, McKay, Reuben, Noble, Graeme, Noble, John-William, Sandison, Martin, Taylor, Stuart (X), Moir, Shaun, Simpson, Julia (II), Hislop, Tom, Kilpatrick, Kayleigh, McKay, Hannah, Dietz, David (I), Dasz, Steven, Chan, Juju, and Marshall, Scarlett.

In this graph, three most important actors are McKay, Reuben with degree 18, Noble, John-William with degree 17, and Noble, Graeme with degree 14. The degree of a actor is the number of movies he/she acts in. The result states that McKay, Reuben attands all movies in this community, and Noble, John-William and Noble, Graeme acts in almost all movies. Since edges are given by common actors, this community is a complete graph; moreover, those three actors make movies tight together.

In 8(a), the most dominant genre is short, and in 8(b), the most dominant genre is romance. Reuben McKay is an actor who most acts in short movies, same as Graeme Noble and John-

William Noble; however, Reuben McKay also acts in some romance movie. Moreover, in this community, romance movies have higher fraction comparing with the whole dataset. Thus, the community has the most dominant genre as shown in 8(a) and 8(b).

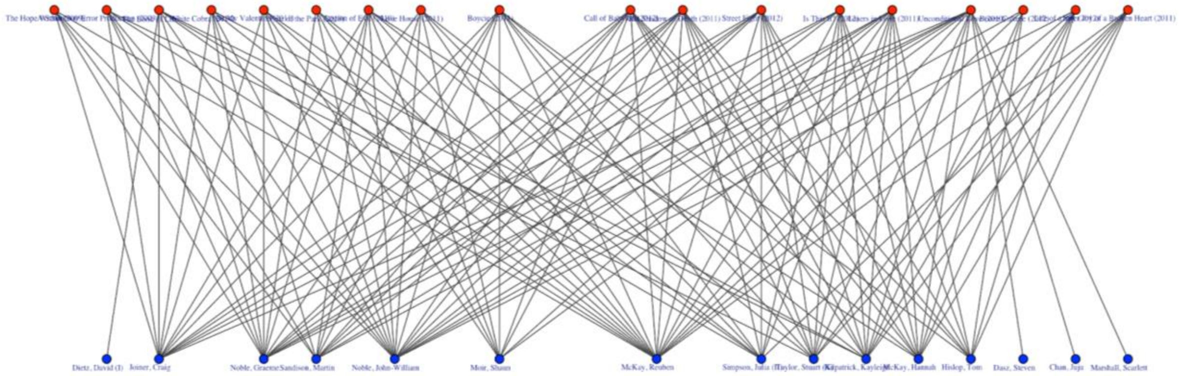

Fig. 24. Bipartite Graph (community 26)

Question 9

The distribution of the available ratings of the movies in the neighborhood of “Batman v Superman: Dawn of Justice (2016)” is depicted in Figure 25. The vertical dashed red line represents the mean value of all rates, which is **6.3454**. The actual rating of “Batman v Superman: Dawn of Justice (2016)” is 6.6, which is similar to the average rating of the movies in its neighborhood.

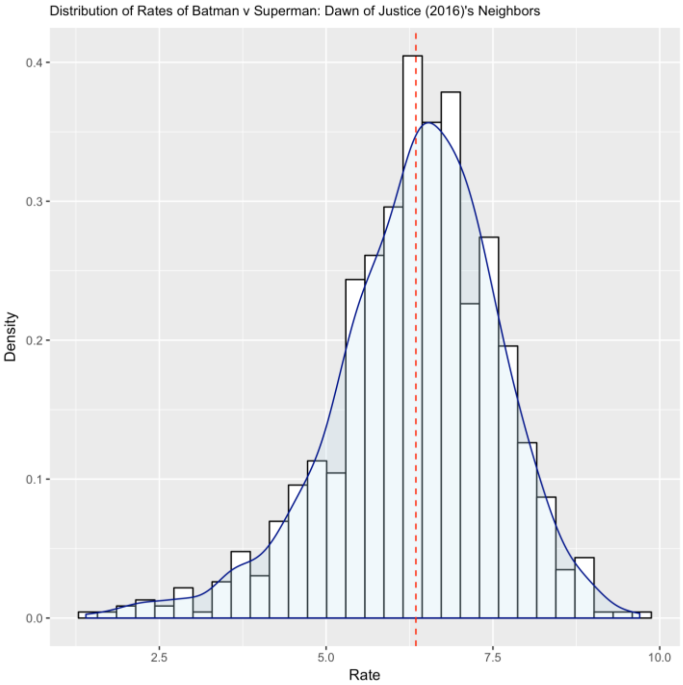

Fig 25. Distribution of Rates in the Neighborhood of “Batman v Superman: Dawn of Justice (2016)”

The distribution of the available ratings of the movies in the neighborhood of “Mission: Impossible - Rogue Nation (2015)” is depicted in Figure 26. The vertical dashed red line represents the mean value of all rates, which is **6.2383**. The actual rating of “Mission: Impossible - Rogue Nation (2015)” is 7.4, which is quite different from the average rating of the movies in its neighborhood.

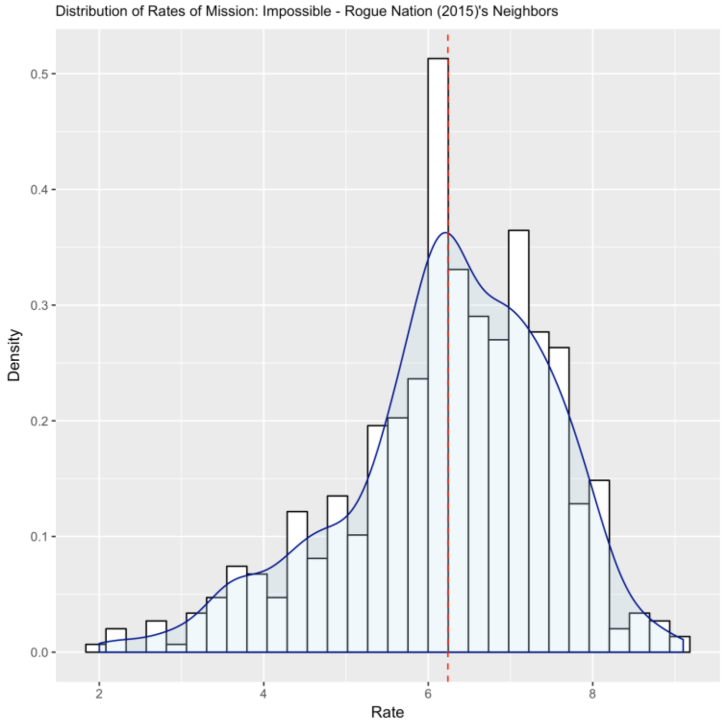

Fig 26. Distribution of Rates in the Neighborhood of “Mission: Impossible - Rogue Nation (2015)”

The distribution of the available ratings of the movies in the neighborhood of “Minions (2015) ” is depicted in Figure 27. The vertical dashed red line represents the mean value of all rates, which is **6.7962**. The actual rating of “Minions (2015) ” is 6.4, which is a little similar to the average rating of the movies in its neighborhood.

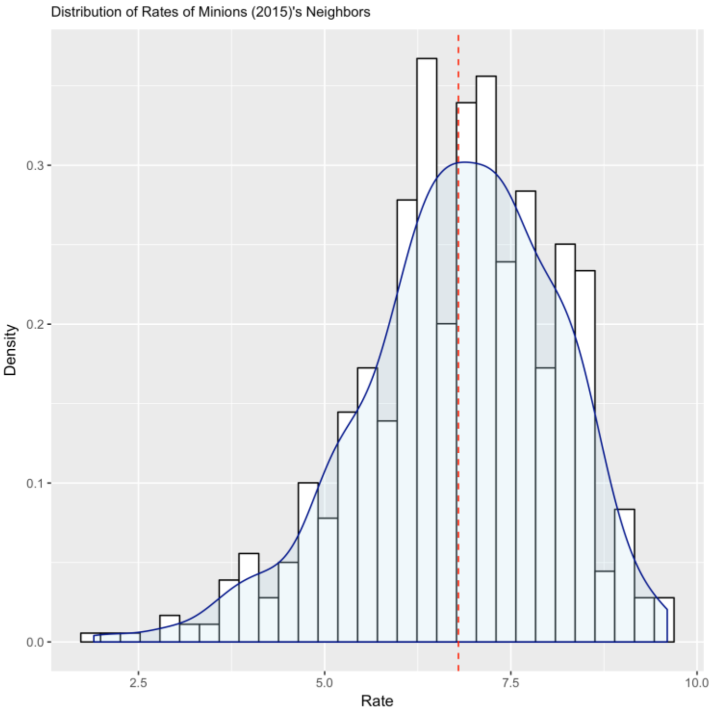

Fig 27. Distribution of Rates in the Neighborhood of "Minions (2015)"

To provide additional illustration of three distributions, we also plot box-plots for each of them, which is shown as Figure 28. Specifically, the square inside each box represents the mean value of corresponding rate distribution.

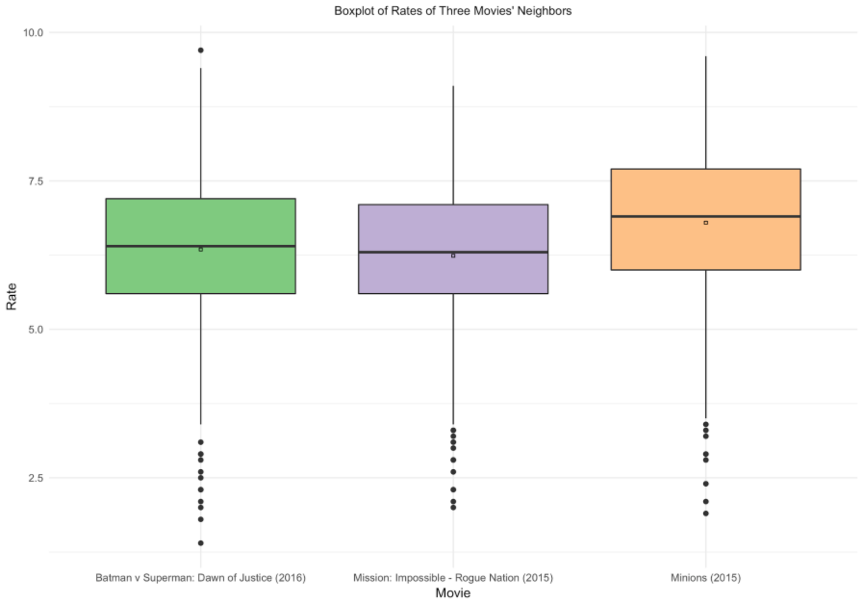

Fig 28. Box-plot of Rates of three Movies' Neighorhood

Question 10

Figure 29 depicts the distribution of the available ratings of the movies in the neighborhood of "Batman v Superman: Dawn of Justice (2016)" as well as **in the same community**(a restriction imposed by Question 10 requirement). The vertical red line represents the mean value of all rates, which is **6.3431**. The actual rating of "Batman v Superman: Dawn of Justice (2016)" is 6.6, which is similar to the average rating of the movies in its neighborhood. Additionally, it is nearly as same as the average rate obtained in Question 9.

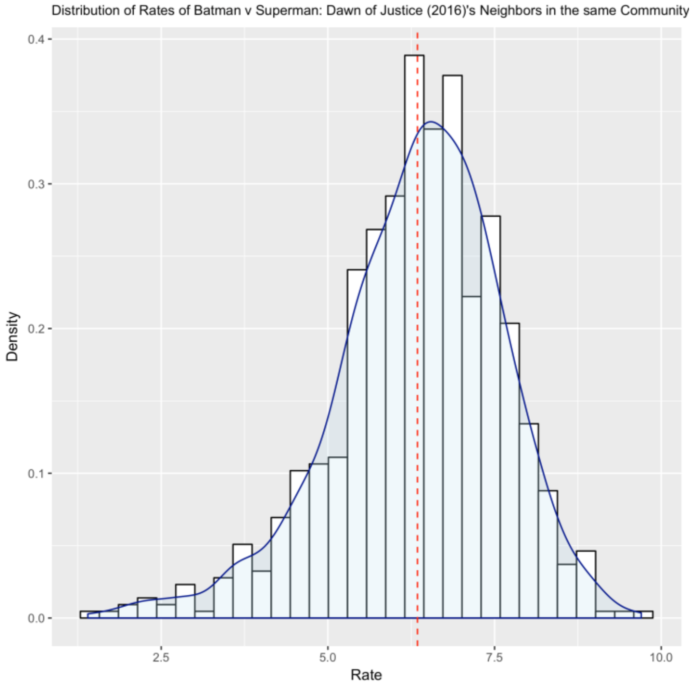

Fig 29. Distribution of Rates in the Restricted Neighborhood of "Batman v Superman: Dawn of Justice (2016)"

Figure 30 depicts the distribution of the available ratings of the movies in the neighborhood of "Mission: Impossible - Rogue Nation (2015)" as well as **in the same community**(a restriction imposed by Question 10 requirement). The vertical red line represents the mean value of all rates, which is **6.2533**. The actual rating of "Mission: Impossible **-** Rogue Nation (2015)" is 7.4, which has great difference with the average rating of the movies in its neighborhood. In addition, it is very similar to the average rate obtained in Question 9.

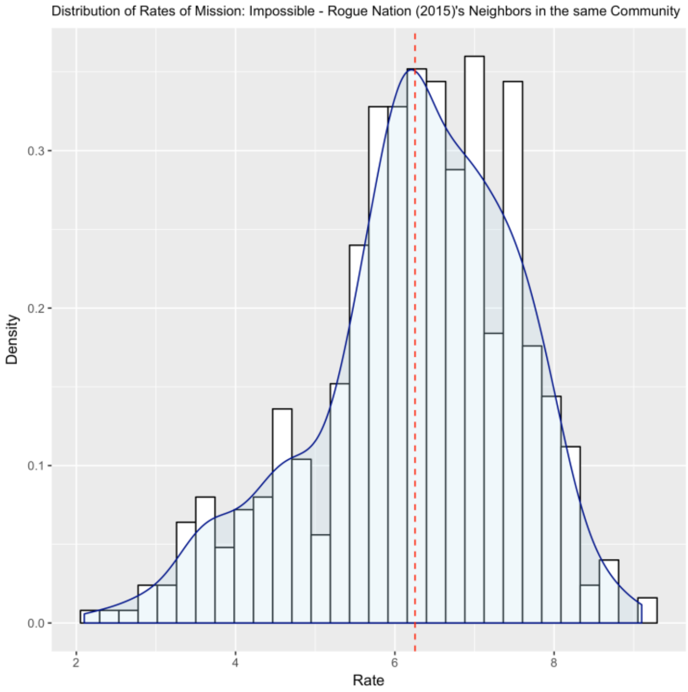

Fig 30. Distribution of Rates in the Restricted Neighborhood of "Mission: Impossible - Rogue Nation (2015)"

Figure 31 shows the distribution of the available ratings of the movies in the neighborhood of "Minions (2015) " as well as **in the same community**(a restriction imposed by Question 10 requirement). The vertical red line represents the mean value of all rates, which is **6.8110**. The actual rating of "Minions (2015) " is 6.4, which is a little similar to the average rating of the movies in its neighborhood. Additionally, it is nearly as same as the average rate obtained in Question 9.

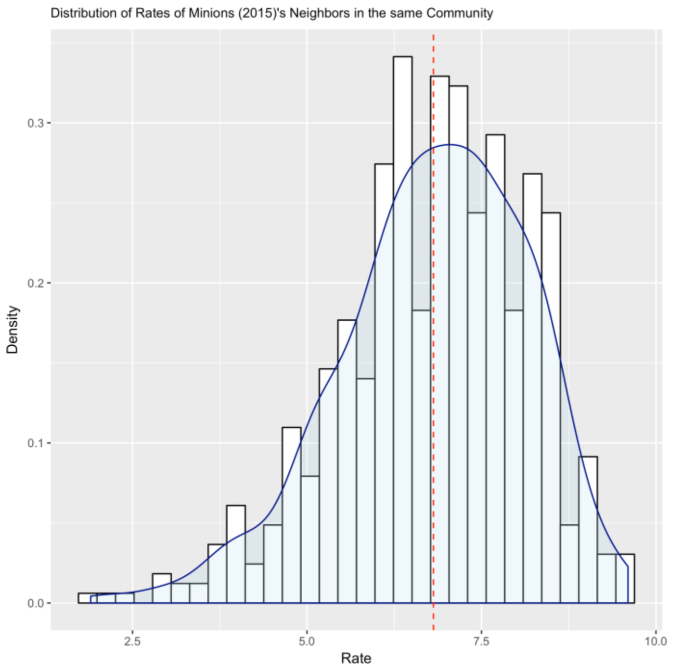

Fig 31. Distribution of Rates in the Restricted Neighborhood of “Minions (2015)”

Again, the box-plot apropos three distributions is depicted in Fig 32.

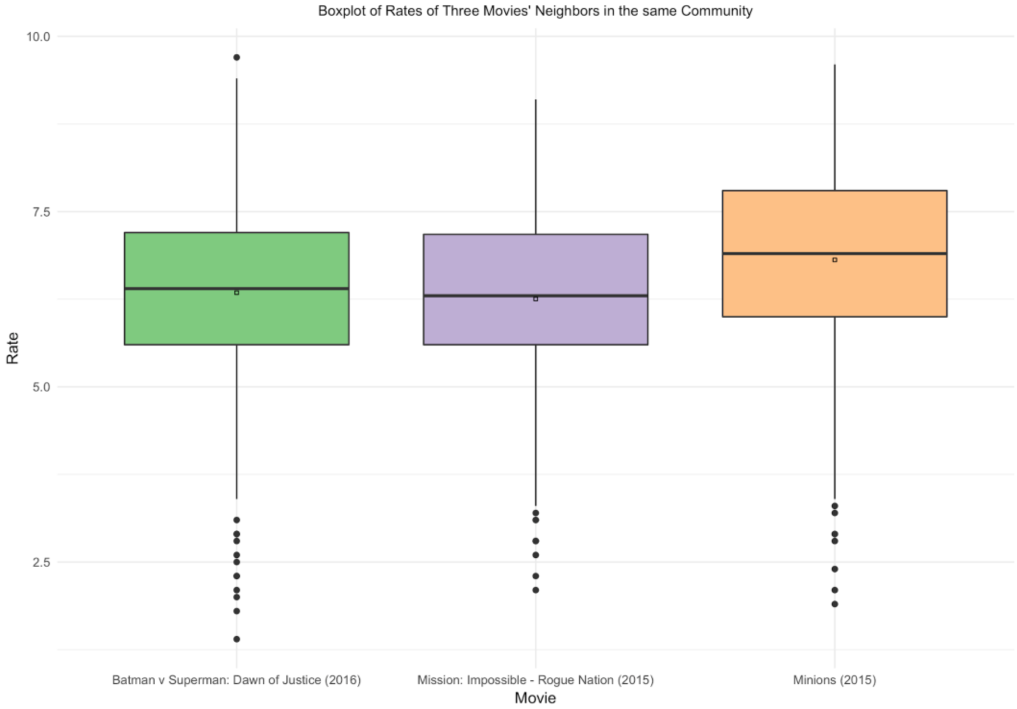

Fig 32. Box-plot of Rates of three Movies’ Neighorhood

To summary, Table 6 presents the comparison of estimates of three movies’ ratings in a more straightforward way.

Table 6. The comparison of estimates of three movies' ratings

Movie	Actual Rate	Average Rate of Neighborhood in Q9	Average Rate of Neighborhood in Q10
Batman v Superman: Dawn of Justice (2016)	6.6	6.3454	6.3431
Mission: Impossible - Rogue Nation (2015)	7.4	6.2383	6.2533
Minions (2015)	6.4	6.7962	6.8110

As is conspicuously demonstrated in table 6, the "same community" restriction imposed on the neighborhood selection does not bring noticeable improvement to the prediction of three movies' rates.

Question 11

In this question, we extract top 5 neighbors for each of three movies according to the edge weight and report their corresponding community memberships. The results are displayed in Table 7, Table 8 and Table 9 respectively.

Table 7 Top 5 Neighbors of "Batman v Superman: Dawn of Justice (2016)"

No.	Movie Name	Community Membership
1	Eloise (2015)	2
2	The Justice League Part One (2017)	2
3	Into the Storm (2014)	2
4	Love and Honor (2013)	2
5	Man of Steel (2013)	2

Table 8 Top 5 Neighbors of "Mission: Impossible - Rogue Nation (2015)"

No.	Movie Name	Community Membership
1	Fan (2015)	20
2	Phantom (2015)	20
3	Breaking the Bank (2014)	2
4	Suffragette (2015)	2

5	Now You See Me: The Second Act (2016)	2

Table 9 Top 5 Neighbors of “Minions (2015)”

No.	**Movie Name**	**Community Membership**
1	The Lorax (2012)	2
2	Inside Out (2015)	2
3	Up (2009)	2
4	Despicable Me 2 (2013)	2
5	Surf's Up (2007)	2

Question 12

In this question, we train a linear regression model to predict the ratings of movies. We select two feature as the independent variables in regression model.

Feature 1 -- Average Actor/Actress Score

We use the PageRank score computed for each actor/actress in Question 4 as a quantitative metric for each actor/actress. Then, the average value of all actors/actresses appearing in the movie is computed as the Actor/Actress Score of this movie. Note that some actors/actresses’ scores may not be available because they are filtered out in the data preprocessing period due to small number of movies they acted in. Thus, we only consider those whose pagerank scores are already computed.

Feature 2 -- Maximum Director Score

Since there is no sufficient information regarding directors available for us, we choose to compute the director score using the movie rates available and director-movie relations in file director_movies.txt.

The director score for each director is the average value of the ratings of all the movies that directors has directed before (the movie’s rate should be available from the moving_rating.txt file). The director score for each movie is the maximum director score of all the directors who direct that movie. We use the director score for each movie as the metric of influence that directors has brought to the movie.

The target value(dependent variable) of regression is the numerical value of movie rate.

Table 10 Details of Linear Regression Model

Method	**Least Squares**
R-squared	0.983
F-statistic	5.965e+06
Df Residuals	201071
Log-Likelihood	-2.4348e+05

The RMSE for training data is **0.8122**.

Now we use the trained model to predict the rate of three target movies. Table 11 shows the comparison between actual rates and predicted rates of three movies.

Table 11 Comparison of Actual Rate and Predicted Rate using Linear Regression

Movie	Actual Rate	Predicted Rate
Batman v Superman: Dawn of Justice (2016)	6.6	7.18
Mission: Impossible - Rogue Nation (2015)	7.4	6.83
Minions (2015)	6.4	7.16

Although the deviation from the actual rate is obvious, the prediction result is still acceptable.

The RMSE for three test movies is **0.6413**.

We find that the average actor/actress score contributes very little to the accuracy of the prediction. That is, if we retain only feature 1, the RMSE would increase significantly. So, it is inferred that the average popularity of actors/actresses in a particular movie does not necessarily determine the quality or the reputation of that movie. By contrast, the impact of director score to the rate is huge. This is actually not surprising because we derive the scores of directors from the rates of movies directly. From the perspective of statistics, this is not very tenable as the multicollinearity problem arises. However, given the limited features we can resort to, we have to adopt this kind of method.

Question 13

In this question, we create a bipartite graph where one vertex party represents the actor/actress and the other vertex party represents the movie.

The bipartite graph constrcuted has 569036 vertices and 3176602 edges in total.

We need to assign a weight for each actor/actress. In our approach, the weight for actor/actress A is computed by the **average of the available rates of movies that A has participated in.**

After assigning the weight to each actor/actress vertex, we need to predict the rates of new movies in a similarly comprehensible way. **The average value of available weights of all actors/actresses in the movie M** is computed and regarded as the prediction value of movie M's rate.

The RMSE for training data is **1.097**, which is larger than that of model using linear regression.

Afterwards we use the bipartite graph model to predict the rate of three target movies. Table 12 shows the comparison between actual rates and predicted rates of three movies.

Table 12 Comparison of Actual Rate and Predicted Rate using Bipartite Graph

Movie	Actual Rate	Predicted Rate
Batman v Superman: Dawn of Justice (2016)	6.6	6.51
Mission: Impossible - Rogue Nation (2015)	7.4	6.47
Minions (2015)	6.4	6.87

Although the training RMSE of bipartite graph approach is worse than that of linear regression model, however, the predictions for these three movies through bipartite graph seem more accurate. To be precise, the RMSE for three test movies is **0.6016**, which is smaller than that of linear regression mode.

Considering the size of test data (3) is quite small, we can not make the conclusion that bipartite graph model has a better predictive performance on the unseen data (lower test error) than the linear regression model that we apply in Question 12.