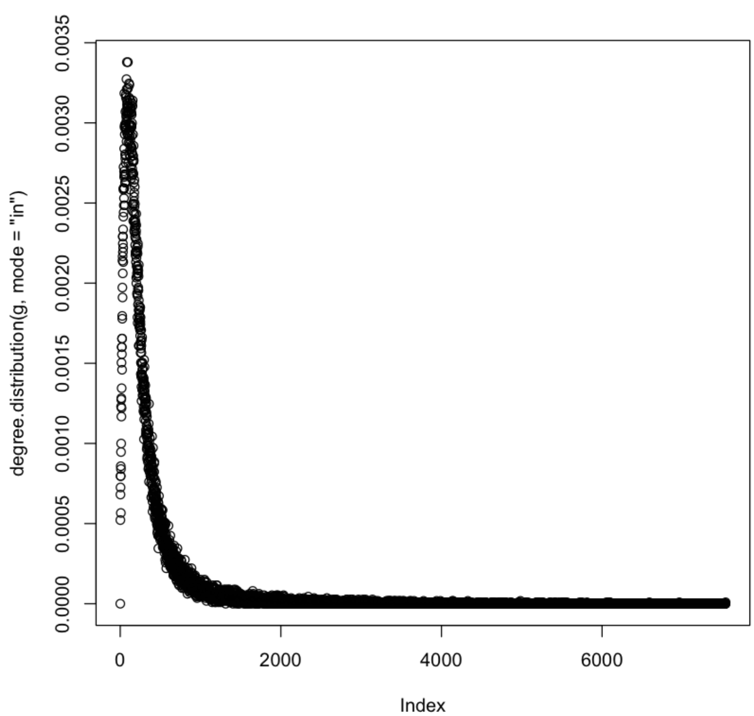
Project 4: IMDb Mining

Duan Li 005026839
Di Ma 004945175
Yuan Shao 504880181
Weijie Tang 305029285

Part 1: Actor/Actress Network

Q1: Total actor+actress: 113132
Movie amount: 468144

Q2: We can tell from the in-degree distribution that most of the actors/actresses have in degree between 0 - 1000, which is reasonable since each movie involves around 100-500 actors/actresses and the average movie amount actor/actress attends



Q3:
[1] "Given actor: Cruise, Tom , Paired actor: Kidman, Nicole , Weight: 0.174603174603175"
[1] "Given actor: Watson, Emma (II) , Paired actor: Radcliffe, Daniel , Weight: 0.52"
[1] "Given actor: Clooney, George , Paired actor: Damon, Matt , Weight: 0.119402985074627"
[1] "Given actor: Hanks, Tom , Paired actor: Allen, Tim (I) , Weight: 0.10126582278481"
[1] "Given actor: Johnson, Dwayne (I) , Paired actor: Austin, Steve (IV) , Weight:

0.205128205128205"
[1] "Given actor: Depp, Johnny , Paired actor: Bonham Carter, Helena , Weight: 0.0816326530612245"
[1] "Given actor: Smith, Will (I) , Paired actor: Foster, Darrell , Weight: 0.122448979591837"
[1] "Given actor: Streep, Meryl , Paired actor: De Niro, Robert , Weight: 0.0618556701030928"
[1] "Given actor: DiCaprio, Leonardo , Paired actor: Scorsese, Martin , Weight: 0.102040816326531"
[1] "Given actor: Pitt, Brad , Paired actor: Clooney, George , Weight: 0.0985915492957746"

We found out the actor/actress that has the largest weight edge with given actor. The result seems to make sense for most of the actors given. For instance, Watson, Emma (II) was paired with Radcliffe, Daniel and they were both in Harry Potter 1 - 7. And Cruise, Tom was paired with Kidman, Nicole who he has been working with a lot.

Q4:

Actor/Actress	Num of movies	In-degree	Pagerank
Flowers, Bess	828	34763	0.0002667548
Harris, Sam (II)	600	28779	0.0002322196
Tatasciore, Fred	353	11335	0.0002098905
Miller, Harold (I)	561	25275	0.0002041256
Jeremy, Ron	637	6815	0.000192747
Lowenthal, Yuri	317	9275	0.0001918877
Phelps, Lee (I)	647	25055	0.0001791656
O'Connor, Frank (I)	623	23395	0.0001644493
Farnum, Franklyn	565	21413	0.0001639422
Sayre, Jeffrey	430	19933	0.0001637879

We could notice that none of the top 10 pagerank actors were listed in the previous section. And most of them are not famous actors (at least now famous to my knowledge). So I think the reason they have high pagerank scores are: 1. Most of them are very old actors/actresses so they have attended a lot of movies 2. Most of them have been in the same movie with famous hollywood stars so they own connection to a lot of important actors.

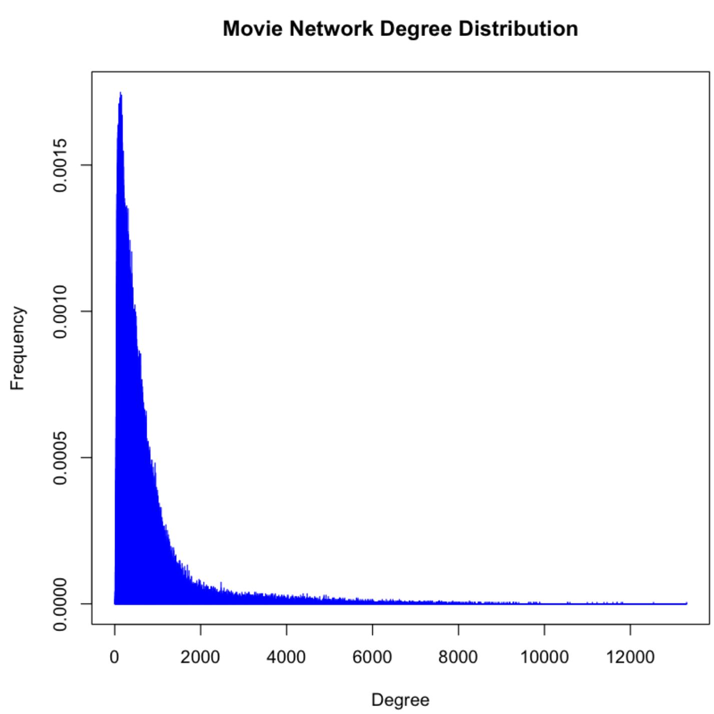
Q5:

Actor/Actress list	Num of Movie	In degree	Pagerank score
Cruise, Tom	63	1842	3.96784e-05
Watson, Emma (II)	25	752	1.745899e-05
Clooney, George	67	1816	3.994201e-05
Hanks, Tom	79	2487	5.096976e-05
Johnson, Dwayne (I)	78	1867	4.19496e-05
Depp, Johnny	98	2534	5.372149e-05
Smith, Will (I)	49	1439	3.196209e-05
Streep, Meryl	97	1863	3.953991e-05
DiCaprio, Leonardo	49	1408	3.161366e-05
Pitt, Brad	71	1972	4.290647e-05

Part 2: Movie Network

2.1. Undirected movie network creation

Q6 Graph below is the degree distribution of the movie network



We observe that

- Median degree = 418 (frequency = 0.0017), Min degree = 3, Max degree = 13307.
- Nodes are more frequently to have degree between 0 and 2000.
- Nodes are less frequently to have degree less than 418 or more than 2000.
- Nodes with degree of 0 or more than 12000 are very rare.

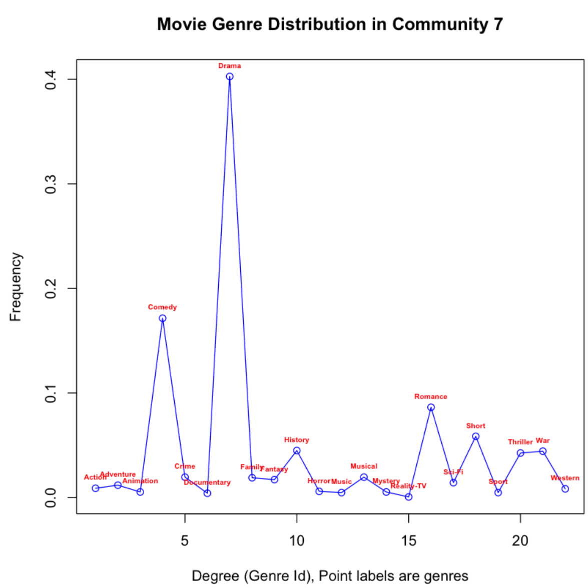
2.2 Communities in the movie network

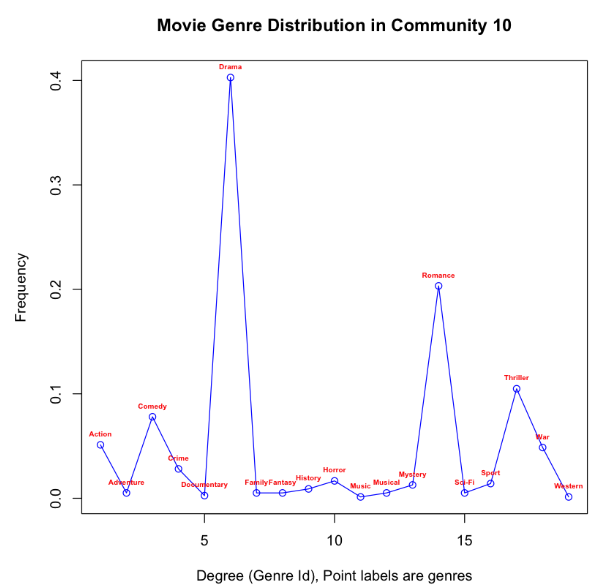
Q7

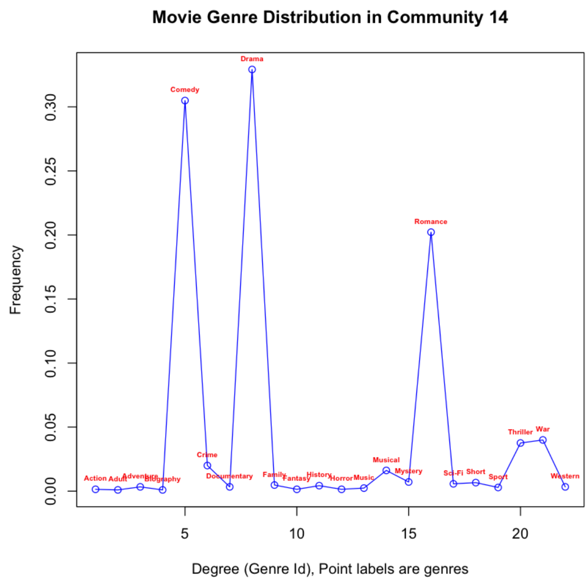
Fast greedy algorithm detects 28 communities and here are the size of each community.

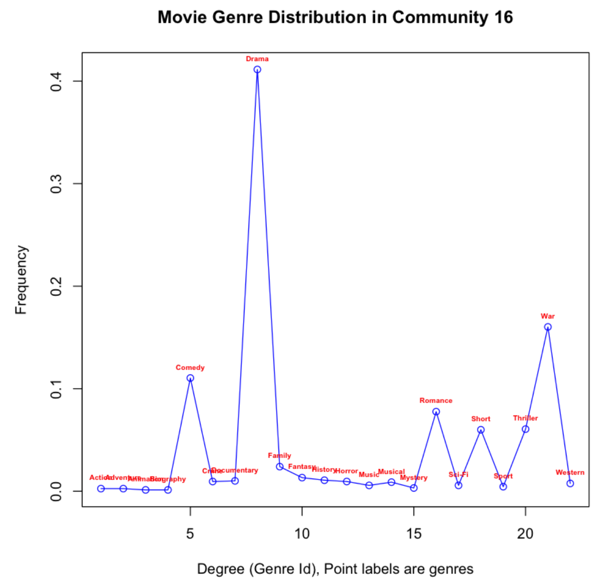
Comm 1	Comm 2	Comm 3	Comm 4	Comm 5	Comm 6	Comm 7
45029	34719	10562	35695	13574	6130	2338
Comm 8	Comm 9	Comm 10	Comm 11	Comm 12	Comm 13	Comm 14
4510	3506	834	7229	6958	4821	2113
Comm 15	Comm 16	Comm 17	Comm 18	Comm 19	Comm 20	Comm 21
3538	1705	2334	9477	1152	620	5944
Comm 22	Comm 23	Comm 24	Comm 25	Comm 26	Comm 27	Comm 28
12	687	14	17	18	22	14

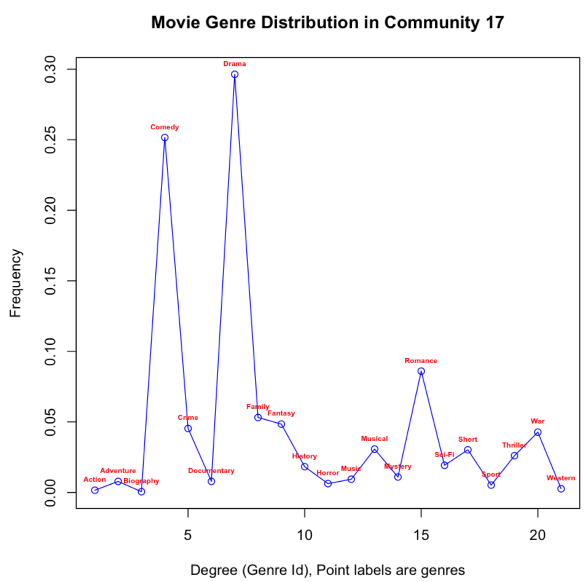
Among 28 communities, we pick community 7, 10, 14, 16, 17, 19, 20, 23, 26, 27 to plot the distribution of the genres of the movies in the community.

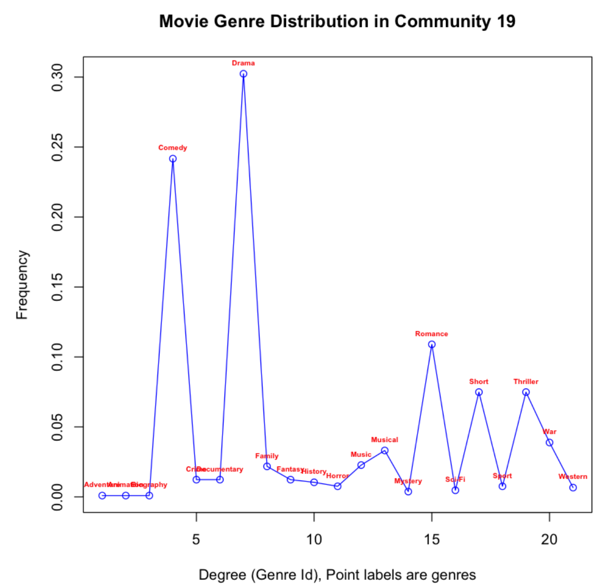


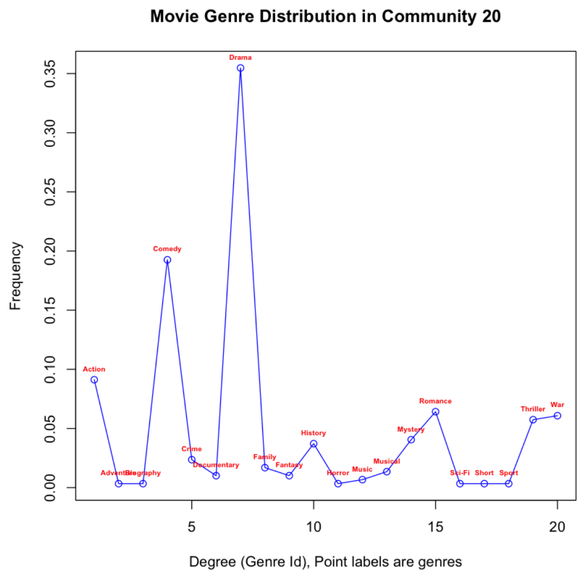


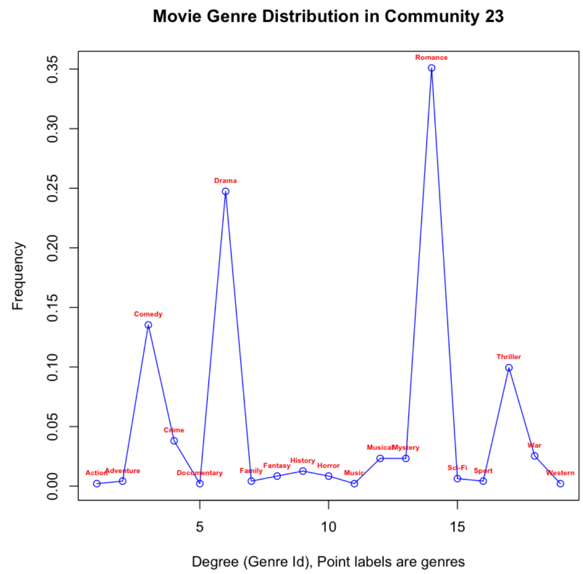

Movie Genre Distribution in Community 14
Frequency
Degree (Genre Id), Point labels are genres
Action
Adult
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Short
Sport
Thriller
War
Western

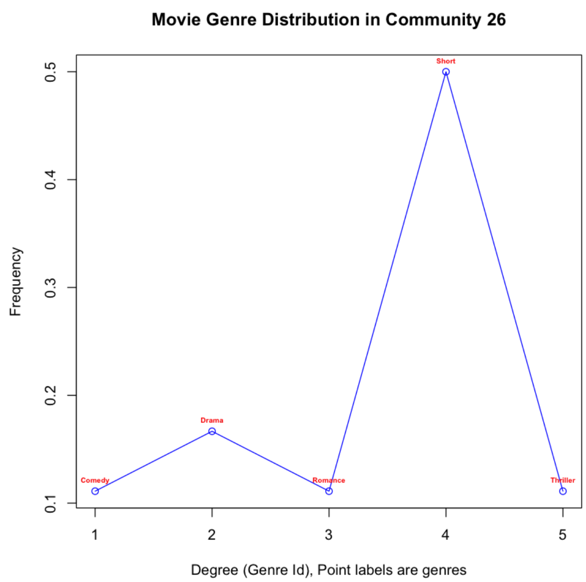

Movie Genre Distribution in Community 16
Frequency
Degree (Genre Id), Point labels are genres
Comedy
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Short
Sport
Thriller
War
Western

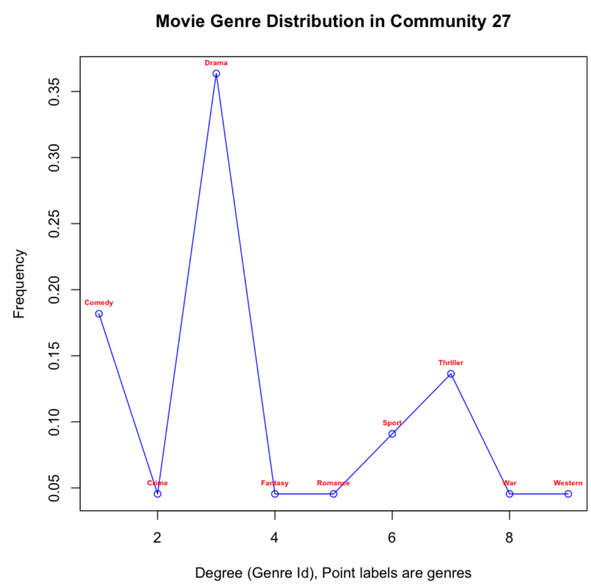

Movie Genre Distribution in Community 17
Frequency
Degree (Genre Id), Point labels are genres
Action
Adventure
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Short
Sport
Thriller
War
Western


Movie Genre Distribution in Community 19
Frequency
Degree (Genre Id), Point labels are genres
Biography
Documentary
Comedy
Drama
Family
Fantasy
History
Horror
Music
Musical
Mystery
Romance
Sci-Fi
Short
Sport
Thriller
War
Western









Q8

(a) The table below shows the most frequent dominant genres across communities based on frequency counts. The number in the parenthesis is the frequency count of the genre in the community. For example, in community 1, Thriller is the most dominant genre because it occurs 8278 times and is more than any other genres.

Comm 1	Comm 2	Comm 3	Comm 4	Comm 5	Comm 6	Comm 7

Thriller (8278)	Short (10976)	Drama (3569)	Drama (8322)	Drama (2989)	Drama (1849)	Drama (681)
Comm 8	Comm 9	Comm 10	Comm 11	Comm 12	Comm 13	Comm 14
Drama (1537)	Drama (1079)	Drama (315)	Drama (1780)	Drama (1761)	Drama (1303)	Drama (692)
Comm 15	Comm 16	Comm 17	Comm 18	Comm 19	Comm 20	Comm 21
Adult (2149)	Drama (652)	Drama (569)	Drama (2429)	Drama (319)	Drama (105)	Drama (1653)
Comm 22	Comm 23	Comm 24	Comm 25	Comm 26	Comm 27	Comm 28
Adult (9)	Romance (166)	Thriller (11)	Short (17)	Short (9)	Drama (8)	Short (10)

(b) The table below shows the most frequent dominant genres across communities based on the modified scores. The number in the parenthesis is the score of the genre in the community.

$$score = ln(c(i)) * \frac{p(i)}{q(i)} = ln(c(i)) * \frac{c(i) \,/\, size\ of\ community}{\sum_{i=1}^{28} c(i) \,/\, size\ of\ enitre\ data\ set}$$

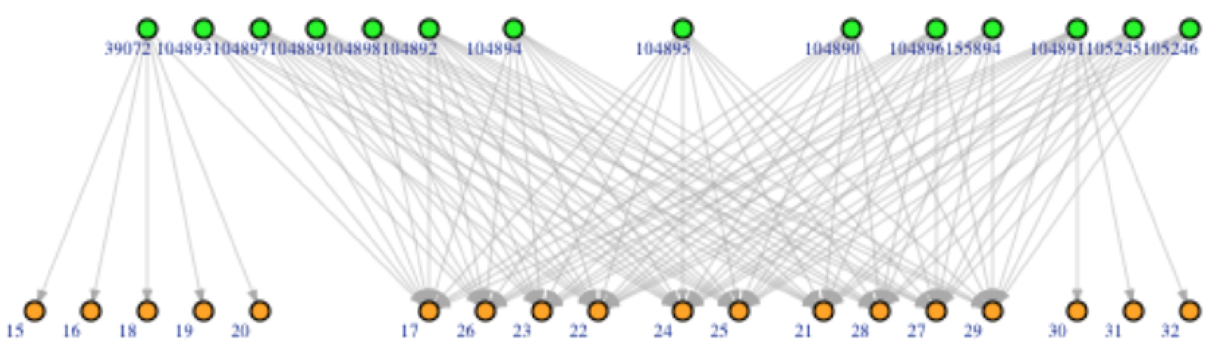
For example, in community 1, Documentary is the most dominant genre because it has the score of 21.200 and is more than any other genres. Compared to 8(a), we found that

- Among 28 communities, only 8 communities have the same most frequent dominant genres based on frequency counts and scores, including community 15, 18, 22, 23, 24, 25, 26, 27.
- In 8(a), 19 communities have Drama as their most dominant genre but in 8(b), 18 of them switch the most dominant genre from Drama to some other genres. It is because 8(a) only consider the frequent counts c(i), and 8(b) consider the fraction of genre in the community p(i) and the fraction in the entire data set q(i) as well.
  Drama is the most dominant genre in the entire data set so Drama has large c(i) and dominate most communities based on frequency counts c(i) in 8(a). However, the large number of Drama in the data set means that q(i) is large and thus the ratio of p(i) to q(i) is small. Hence, Drama has low score due to large c(i) but small ratio of p(i) to q(i) in 8(b)
- In 8(a) and (b), 8 communities have the same most dominant genre and their sizes are 3538, 9477, 12, 687, 14, 17, 18, 22, and 14. 6 of them has size between 10 and 700, given that only 8 communities has size between 10 and 700. So community with smaller size has the tendency to have the same dominant genre based on both frequency counts and scores.

Comm 1	Comm 2	Comm 3	Comm 4	Comm 5	Comm 6	Comm 7
Documentary (21.200)	Western (36.715)	War (17.763)	Comedy (14.080)	Family (22.074)	Musical (15.318)	History (16.527)

Comm 8	Comm 9	Comm 10	Comm 11	Comm 12	Comm 13	Comm 14
Musical (22.416)	Adventure (38.561)	Romance (10.398)	Adventure (27.166)	Mystery (15.203)	Family (31.895)	Comedy (18.184)
Comm 15	Comm 16	Comm 17	Comm 18	Comm 19	Comm 20	Comm 21
Adult (347.397)	War (33.418)	Fantasy (12.982)	Drama (8.765)	Comedy (11.376)	Action (6.876)	Action (46.275)
Comm 22	Comm 23	Comm 24	Comm 25	Comm 26	Comm 27	Comm 28
Adult (122.837)	Romance (13.291)	Thriller (23.261)	Short (26.816)	Short (10.398)	Short (7.457)	Short (6.383)

(c) Here is the bipartite between movies and actors in community 24 of size 14. Green nodes are movies and orange nodes are actors. Labels on green nodes are movie ids in the entire network and labels on orange nodes are ids assigned to actor for the sake of presentation. Movie and actor names are too long to show in the graph. The map between ids and the real movie/actor names is presented after the bipartite.
Three most important actors are: (1) 22 Desjardins, Nick, (2) 24 Lafond-Martel, Olivier , and (3) 25 Legros, Simon (I). They help form the community because they have acted in all the movies in community 24 expect Liverpool (2012).
In both 8(a) and 8(b), the dominant genre of community 24 is Thriller with frequency counts of 11 and score of 23.261. These actors have all acted in 10 Thrillers and 3 Shorts movies. Hence, there is a correlation between these actors and the dominant genres for this community in 8(a) and 8(b).



104891	Cent jours avant le lendemain (2015) Thriller	39072	Liverpool (2012) Thriller
104889	669: Escape the Reality (2011)	104890	An Olimatsim adventure (2011) Short

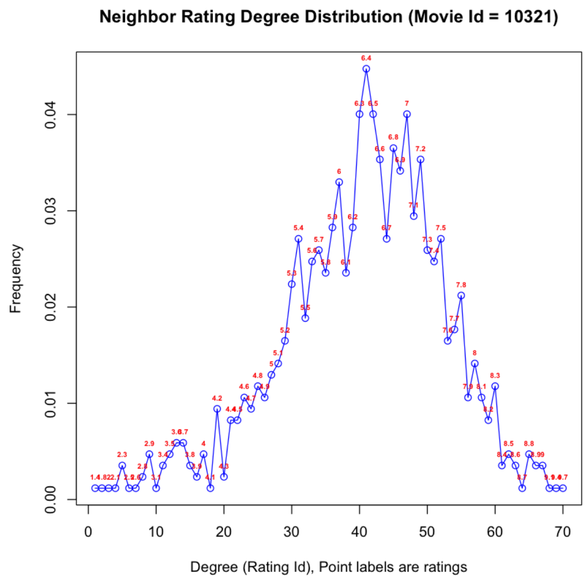
	Thriller		
104892	L'affaire Hawkins (2014) Thriller	104893	La peur anonyme (2014) Short
104894	La Peur aux trousse (2015) Thriller	104895	Les oiseaux se cachaient pour mourir (2015) Thriller
104896	Midnight Stranger (2011) Thriller	104897	New York Vengeance (2013) Thriller
104898	October Sunset (2017) Thriller	105245	Des humains bien tranquilles (2016) Thriller
105246	Les années folles (2016) Thriller	155894	Mocakoma (2013) Sport
15	Antaki, Joseph	16	Beaulac, Sebastien
17	Boucher-L'Écuyer, Émile Pascal	18	Gagné, David
19	Priest, Benoit	20	Primeau, Marc
21	Bourassa-Simpson, Mathieu	22	Desjardins, Nick
23	Fortin, Samuel (I)	24	Lafond-Martel, Olivier
25	Legros, Simon (I)	26	Charlebois, Jessica
27	Valin, AndrÈanne	28	Guimont, MÈlanie
29	Riel-Dery, Jessica	30	Leonard, Joshua
31	Williams, Michael C.	32	Donahue, Heather (I)

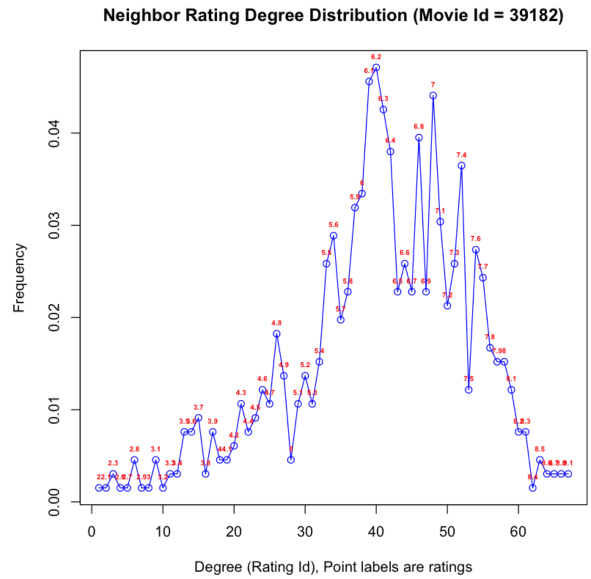
2.3 Neighborhood analysis of movies

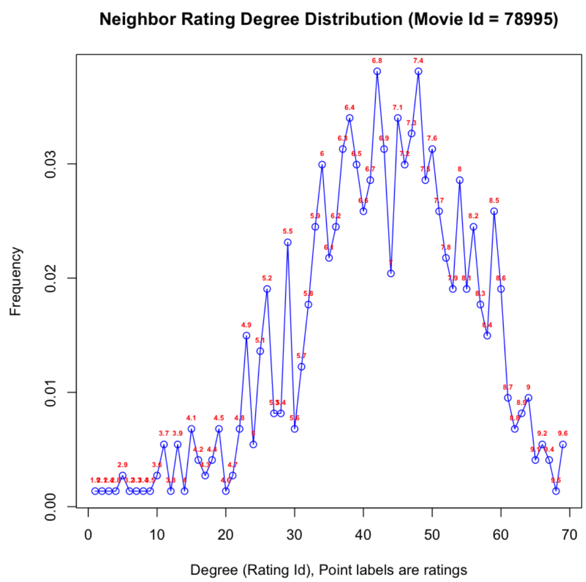
Q9 The average rating of the movies in the neighborhood is similar to the rating of the movie whose neighbors have been extracted.

Movie Id	Movie Name	Real Rating	Average rating of neighbors	Most freq rating of neighbors
10321	Batman v Superman: Dawn of Justice (2016)	6.6	6.287	6.4
39182	Mission: Impossible - Rogue Nation (2015)	7.4	6.091	6.2

78995	Minions (2015)	6.4	6.570	6.8





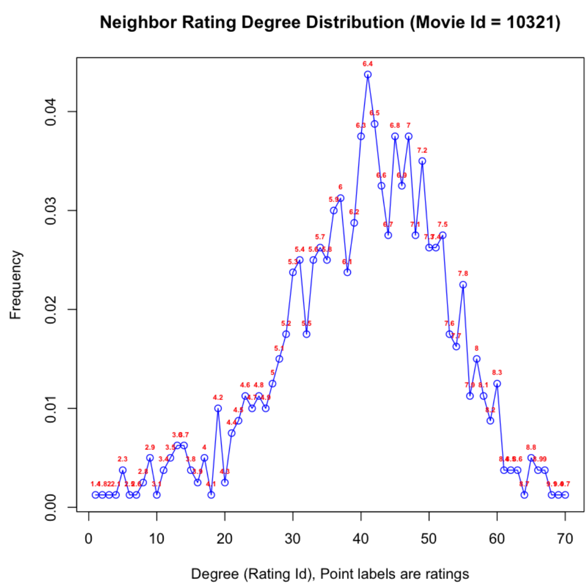


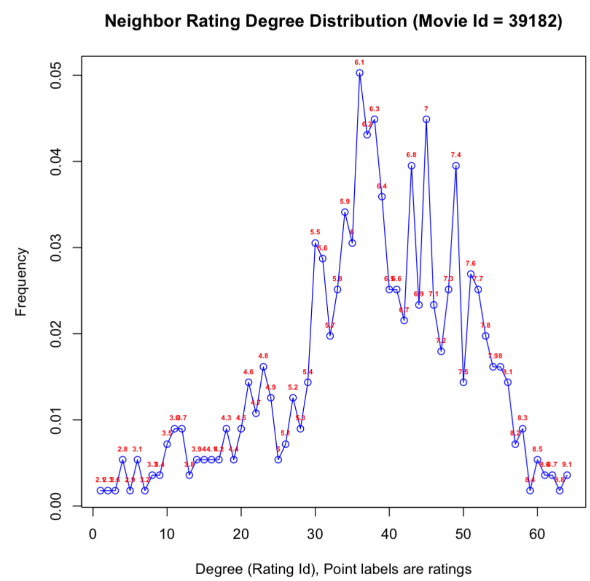
Q10 By observing the two tables and graphs in Q9 and Q10, we found that

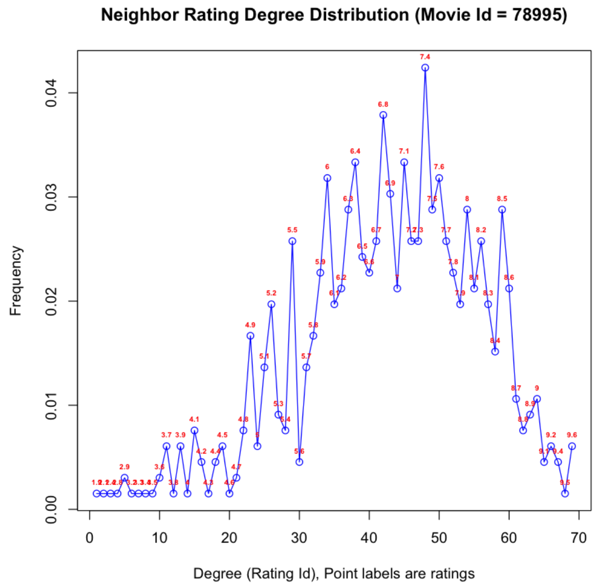
- For Batman v Superman, the average rating of the restricted neighbors is 0.004 better than the average rating of the unrestricted neighbors.
- For Mission Impossible, the average rating of the restricted neighbors is 0.146 better than the average rating of the unrestricted neighbors.
- For Minions, the average rating of the restricted neighbors is 0.003 worse than the average rating of the unrestricted neighbors.

Overall, there is a little bit better match between the average rating of the movies in the restricted neighborhood and the rating of the movie whose neighbors have been extracted.

Movie Id	Movie Name	Real Rating	Average rating of restricted neighbors	Most freq rating of restricted neighbors
10321	Batman v Superman: Dawn of Justice (2016)	6.6	6.291	6.4
39182	Mission: Impossible - Rogue Nation (2015)	7.4	6.237	6.1
78995	Minions (2015)	6.4	6.573	7.4







Q11

Batman v Superman: Dawn of Justice (2016) is in Community 1

Here are the top 5 neighbors and their community memberships:

Movie id	Movie name	Community id
22165	Eloise (2015)	1
10363	The Justice League Part One (2017)	1
33301	Into the Storm (2014)	1
9502	Love and Honor (2013)	1
3384	Man of Steel (2013)	1

Mission: Impossible - Rogue Nation (2015) is in Community 1

Here are the top 5 neighbors and their community memberships:

Movie id	Movie name	Community id
32741	Fan (2015)	5
32744	Phantom (2015)	5
57762	Breaking the Bank (2014)	1

68813	Suffragette (2015)	1
39183	Now You See Me: The Second Act (2016)	1

Minions (2015) is in Community 1
Here are the top 5 neighbors and their community memberships:

Movie id	Movie name	Community id
37617	The Lorax (2012)	1
16741	Inside Out (2015)	1
37589	Despicable Me 2 (2013)	1
52491	Up (2009)	1
61332	Surf's Up (2007)	1

We observed that almost all the top 5 neighbors are in the same community of with the movie extracted.

2.4 Predicting ratings of movies

Q12

In this question, we use the movie network and the actor pageranks from the previous sections to predict the ratings of movies. To be specific, the features are top 5 pageranks of the actors in each movie. We use all the movies with non-NA rating as the training set, which doesn't include the three movies that we are going to predict. We use linear regression model and get **RMSE = 1.243792**. The predicted ratings are shown below.

```
Batman v Superman: Dawn of Justice (2016)
Ground truth rating: NA
Predicted rating: 6.303812

Mission: Impossible - Rogue Nation (2015)
Ground truth rating: NA
Predicted rating: 6.124734

Minions (2015)
Ground truth rating: NA
Predicted rating: 6.081247
```

Q13

In this question, we are asked to predict the ratings of movies using the actor weights of each movie. We use two ways to calculate each actor's weight: 1) the average of top 5 ratings of the actor's movies and 2) the average of all the ratings of the actor's movies. Then the features are the top 5 actor weights for each movie. We also use two ways for the model: 1) average the features and 2) linear regression. So there are in total four methods.

- Method 1: actor weight - the average of top 5 ratings of the actor's movies
  model - average the features
- Method 2: actor weight - the average of top 5 ratings of the actor's movies
  model - linear regression
- Method 3: actor weight - the average of all the ratings of the actor's movies
  model - average the features
- Method 4: actor weight - the average of all the ratings of the actor's movies
  model - linear regression

	RMSE	Dawn of Justice	Rogue Nation	Minions
Method 1	2.10236	8.511	8.26	9.212
Method 2	1.186899	6.544315	6.458406	6.917641
Method 3	1.099779	7.9	7.733143	7.568849
Method 4	1.003335	8.646708	8.272409	7.856296

We can see that all methods except method 1 get better RMSE, and the method 4 (using the average of all the ratings of the actor's movies as actor weight and linear regression for the model) has the best result. We also observe that the movie Minions tend to get lower ratings in both Q12 and Q13, because in both questions, the features are highly dependent on the actors of each movie, and since the Minions is a cartoon movie which doesn't have many famous actors as the other two. As a result, all the models need improvements for cartoon movies like the Minions. One possible way is to include the director information in the features.