EE 232E Project 4
IMDb Mining

Hengjie Yang, Sheng Chang, Wandi Cui, and Tianyi Liu

June 5, 2018

1 Actor/Actress Network

After the preprocessing process:

(1) removing the actor/actress who has acted in less than 10 movies.

(2) removing the inconsistency.

the total number of actors and actresses is: 113132

the total number of unique movies that these actors and actresses have acted in is: 468150

1.1 Directed actor/actress network creation

(1) In-degree distribution of the actor/actress network after removing the actor/actress who has acted in less than 10 movies is shown in Figure 1

(2) In-degree distribution of the actor/actress network without removing the actor/actress who has acted in less than 10 movies (sampled data) are shown in Figure 2 and Figure 3.

(3) Observing the distribution, we can see that generally the actor and actress graph is a power law graph, especially without the preprocessing. Although after removing some of the actor or actress(acted in less than 10 movies), the distribution of degree in few degree values is separated, such as those data point near 0 degree in Figure 1, we can still get a more clear result that this network obeys power law in general.

1.2 Actor pairings

To find the actor pairings for the objective actors, we found the actor or actress who has largest weight edge with them, and determined actor pairs.

The weight from actor/actress1 -> actor/actress2 is defined as $(S_1 \cap S_2)/S_1$, where S_1 is the set of movies in which actor/actress1 has involved in, and S_2 is the set of movies in which actor/actress2 has involved in. Basically, For a given actor/actress A, his/her pair B

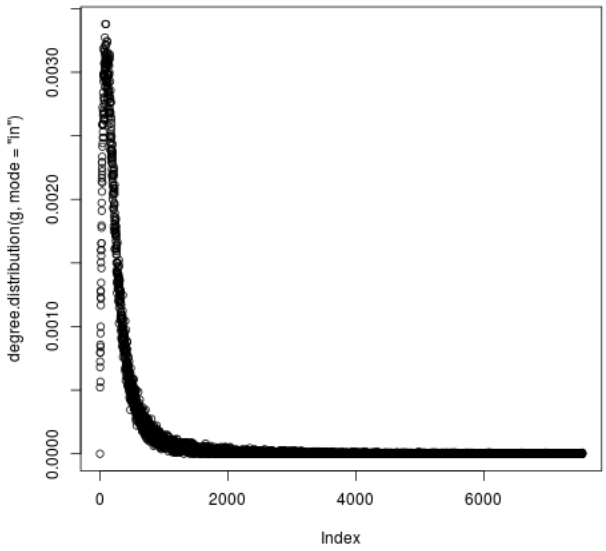

Figure 1: In-degree distribution of the actor/actress network (after preprocess)

is the one whom he/she has been collaborated the most frequently with, because B co-acted in the most of number of movies with A compared with other actors/actresses.

Then we have the following results:

(id: 14503) Tom Cruise -> (id: 92160) Nicole Kidman, score: 0.1746031746031746

(id: 111298) Emma Watson (II) -> (id: 54782) Daniel Radcliffe, score: 0.52

(id: 12812) George Clooney -> (id: 15209) Matt Damon, score: 0.11940298507462686

(id: 27258) Tom Hanks -> (id: 1193) Tim Allen, score: 0.1

(id: 32389) Dwayne Johnson (I) -> (id: 2949) Steve Austin, score: 0.20512820512820512

(id: 16878) Johnny Depp -> (id: 78310) Helena Bonham Carter, score: 0.08163265306122448

(id: 62774) Will Smith (I) -> (id: 22000) Darrell Foster, score: 0.12244897959183673

(id: 107832) Meryl Streep -> (id: 16135) Robert De Niro, score: 0.061855670103092786

(id: 17285) Leonardo DiCaprio -> (id: 60365) Martin Scorsese, score: 0.10204081632653061

(id: 53248) Brad Pitt -> (id: 12812) George Clooney, score: 0.09859154929577464

We think the actor pairings make sense. For instance, Emma Watson's pair is Daniel Radcliffe considering Emma Watson has acted in eight Harry Potter film series in which Daniel Radcliffe also acted in, which makes Daniel the actor whom Emma co-acted in movies most frequently compared with other actors/actresses, which makes Daniel the actor whom Emma prefers to work the most with among other actors/actresses.

However, this method may tend to pair "Super Stars" to other actors in the network. Here "Super Stars" means those who acted a large number of movies and could be seen as

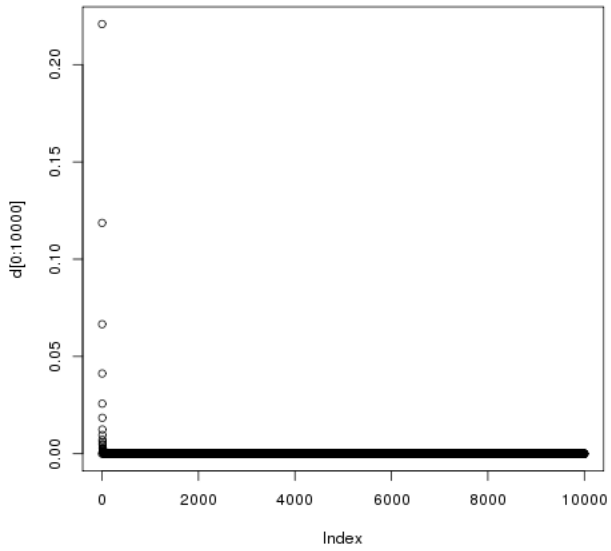

Figure 2: In-degree distribution of the actor/actress network (without preprocess)

experienced since they must be added to the movie network earlier than others. Because of these properties of "Super Stars", they are most likely to be a pair of many actors. If we think it in the way that young actors prefer to work with those "Super Stars" since the movie "Super Stars" involve in has high possibility to be a high-quality movie, which may make young actors become famous sooner. However, if we think it in another way that those young actors are lack of experience so that they have to act some unimportant characters in some famous movies with "Super Stars".

Therefore, working preference is a difficult concept to definite, so if we want to get more accurate or more reasonable results, we still need more information, though our method here is already make some sense.

1.3 Actor rankings

We aimed to find to find the top 10 actor/actress in the network using the google's pagerank algorithm. Those information of the top 10 actor/actress is shown in Table 1, including the name, the number of movies and the in-degree of each of the actor/actress in the top 10 list.

We can see from the result that it does not have any of the actor/actress listed in the previous section. In general, the more movie they took part in, the high pagerank they may had, because that means they had more changes to cooperate with other actor/actress and

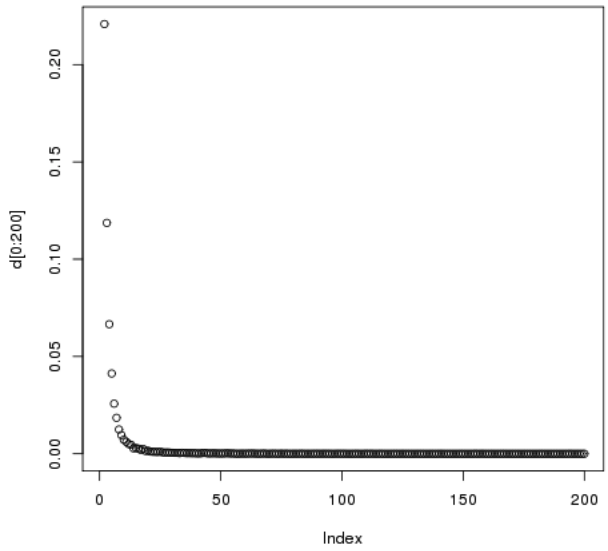

Figure 3: In-degree distribution of the actor/actress network2 (without preprocess)

it's obvious that they may have higher degree in the network. After googling it, we found that most people int the top 10 are actually voice actors. That's why they can take part in hundreds of movies and that also explains why those famous previous actor/actress are not included in the top 10. Even though those movie super stars acted so many movies, it's very common that they still act less than those voice actors.

What's more, the same information of the actor/actress listed in the previous section is shown in Table 2.

2 Movie Network

2.1 Undirected movie network creation

We create a weighted undirected movie network. And the degree distribution of the movie network is shown in Figure 4. We can see from the result that most movies have a degree between 500 and 1000, also there're only a few movies that have very large or very small degrees. The result is not super surprising since it is very common that lots of movies share same popular movie starts for the box office.

Table 1: Top 10 highest pagerank score actor/actress

Name	the Number of Movies	In-degree
Flowers, Bess	828	7537
Tatasciore, Fred	355	3954
Harris, Sam (II)	600	6960
Blum, Steve (IX)	373	3316
Miller, Harold (I)	561	6587
Jeremy, Ron	637	3177
Phelps, Lee (I)	647	5563
Lowenthal, Yuri	318	2662
Downes, Robin Atkin	267	2953
O'Connor, Frank (I)	623	5502

Table 2: The same information table for previous actor/actress

Name	the Number of Movies	In-degree
Tom Cruise	63	1651
Emma Watson (II)	25	453
George Clooney	67	1573
Tom Hanks	80	2064
Dwayne Johnson (I)	78	1357
Johnny Depp	98	2144
Will Smith (I)	49	1319
Meryl Streep	97	1594
Leonardo DiCaprio	49	1301
Brad Pitt	71	1739

2.2 Communities in the movie network

By detecting communities with Fast Greedy algorithm for our movie network, we got 28 communities. And for Quesition 7, we just picked the first 10 communities to plot the distribution of the genres of the movies in each community.

NOTICE: Some of the movies' genre information is missing in the given dataset, so we marked them as "NAN"; however, only Question 7, we considered them in our plots.

The 10 plots are shown in Figure 5 to Figure 14.

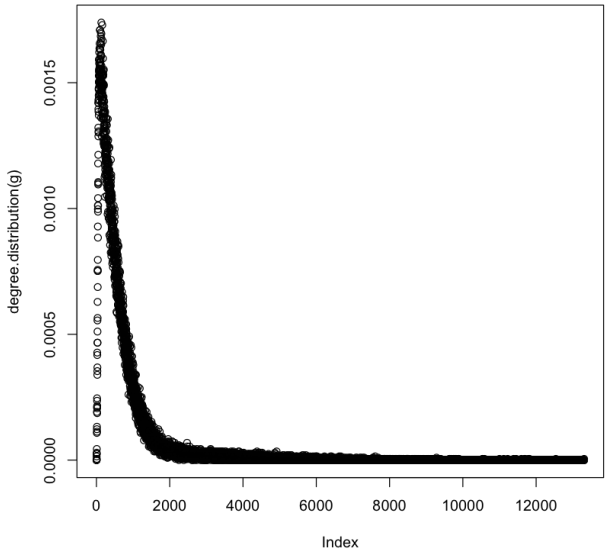

Figure 4: Degree distribution of the movie network

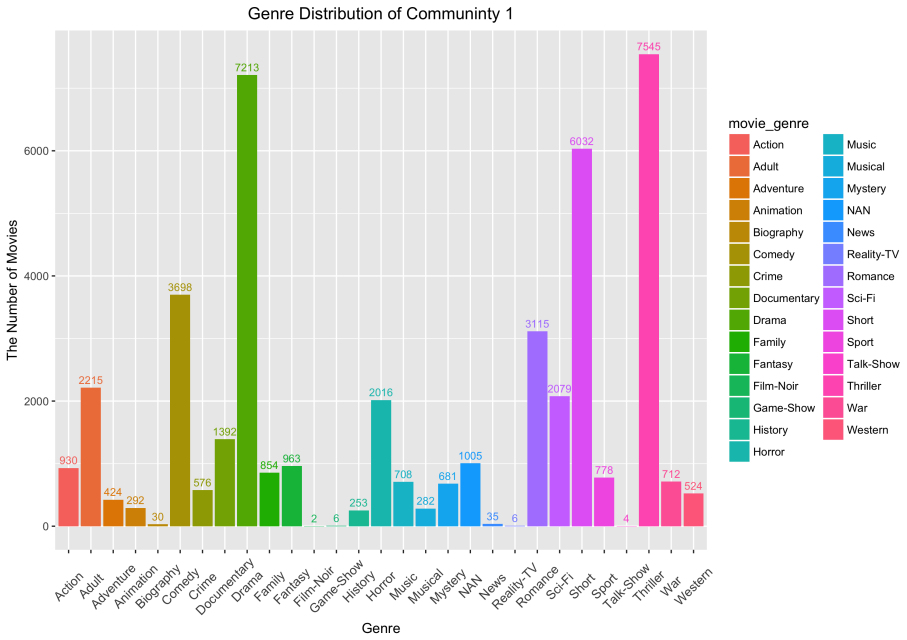

Figure 5: Genre Distribution of Community 1

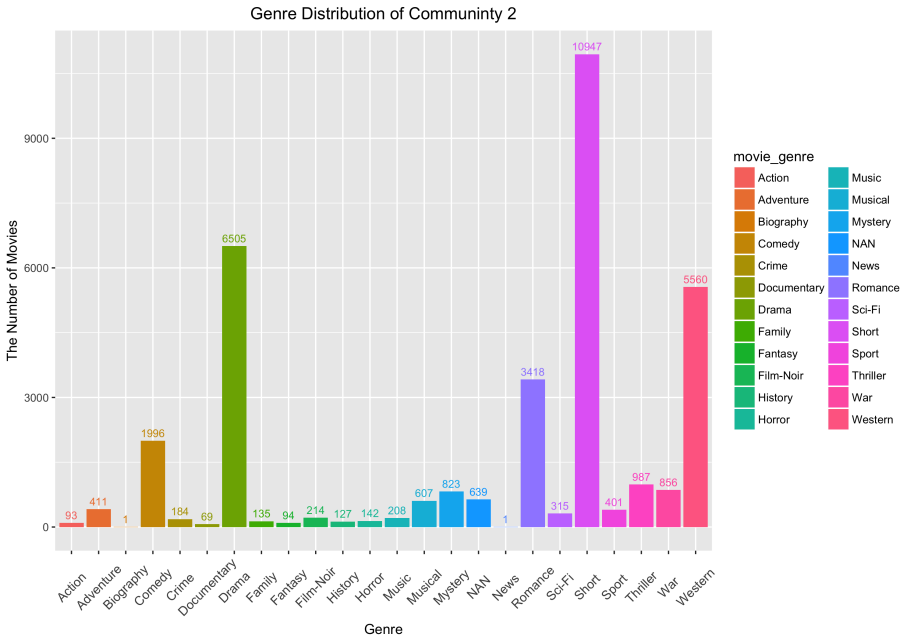

Figure 6: Genre Distribution of Community 2

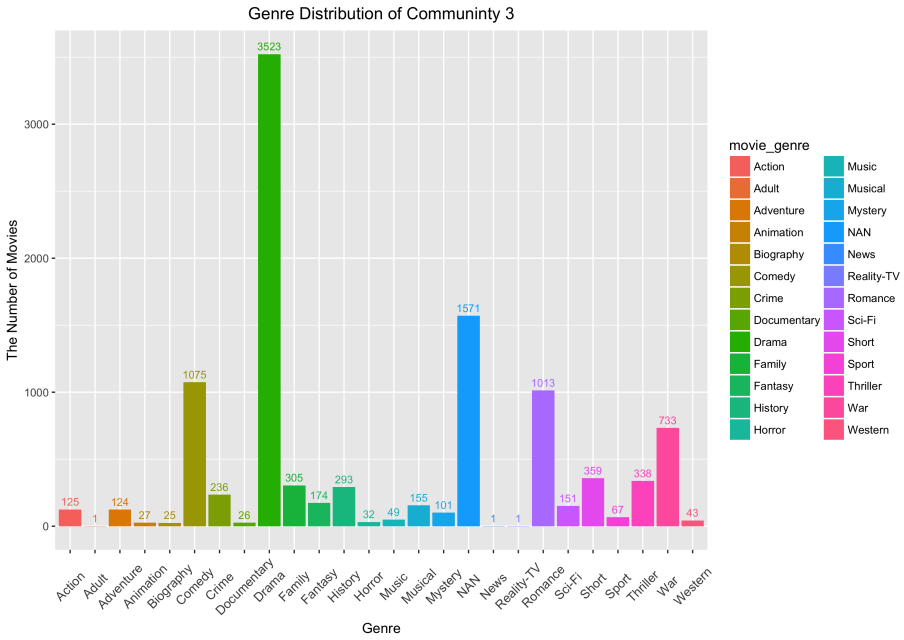

Figure 7: Genre Distribution of Community 3

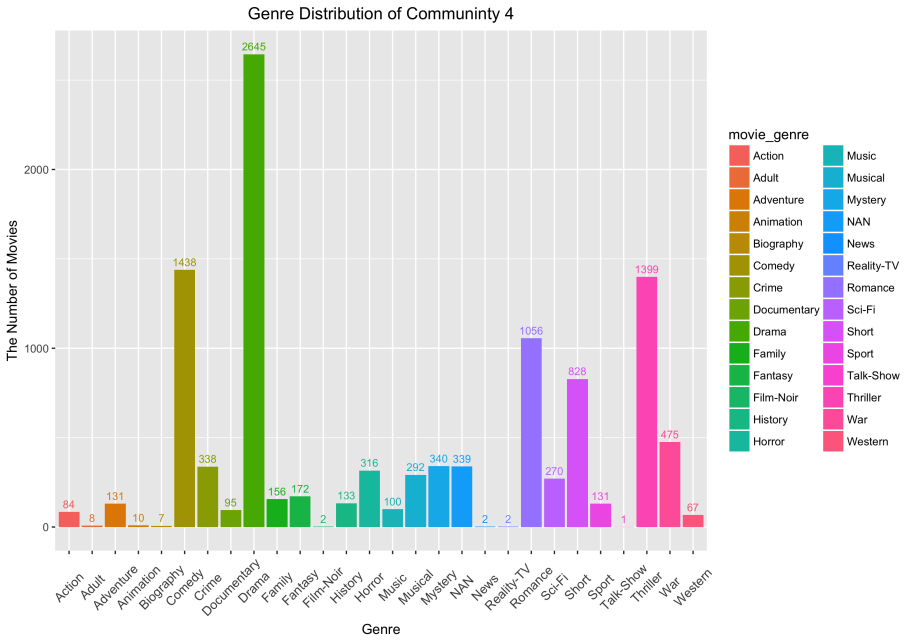

Figure 8: Genre Distribution of Community 4

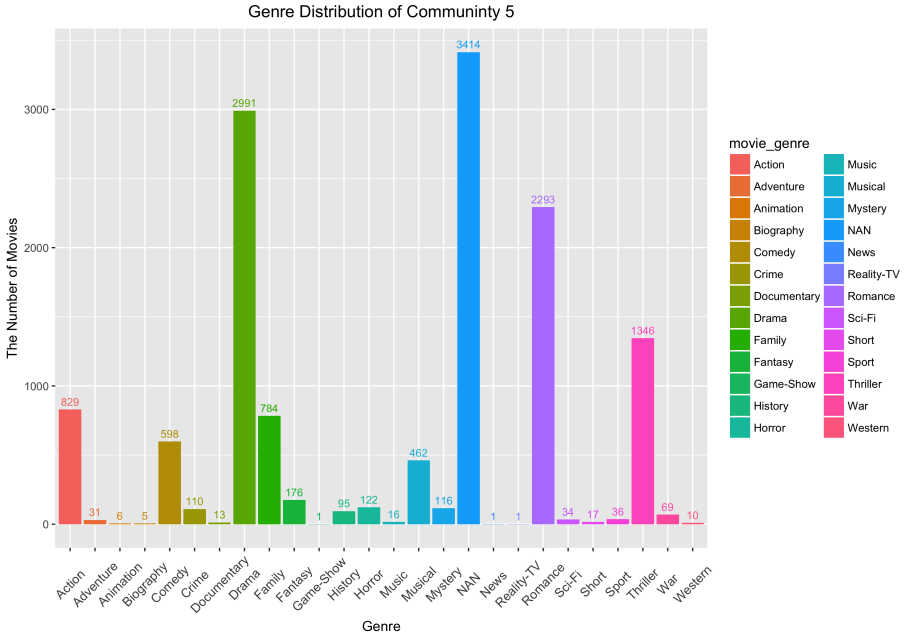

Figure 9: Genre Distribution of Community 5

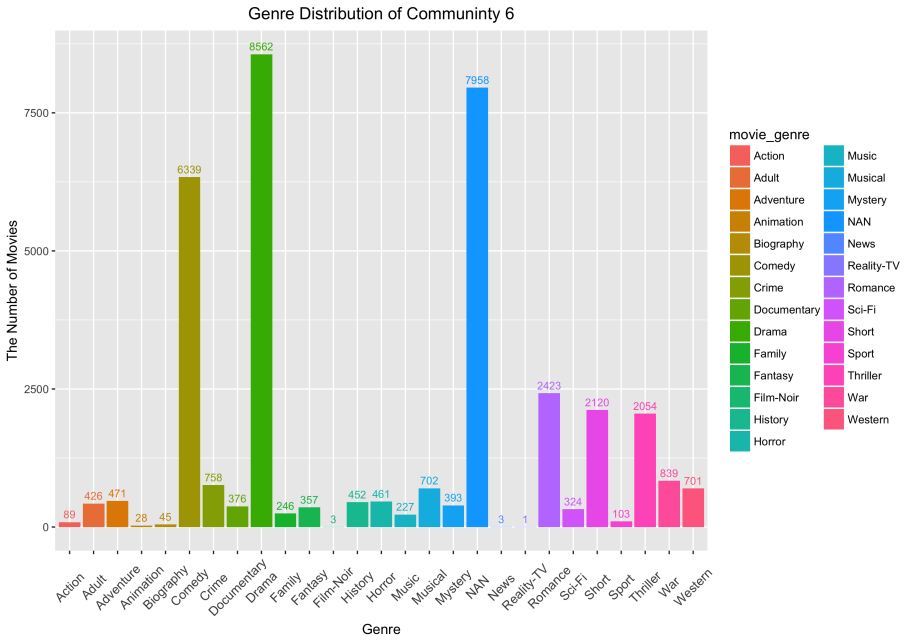

Figure 10: Genre Distribution of Community 6

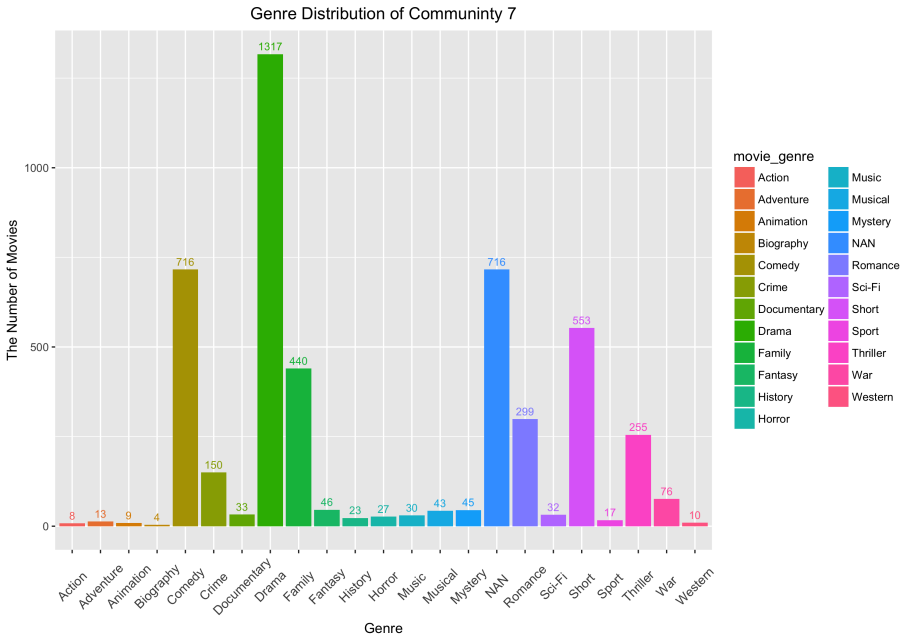

Figure 11: Genre Distribution of Community 7

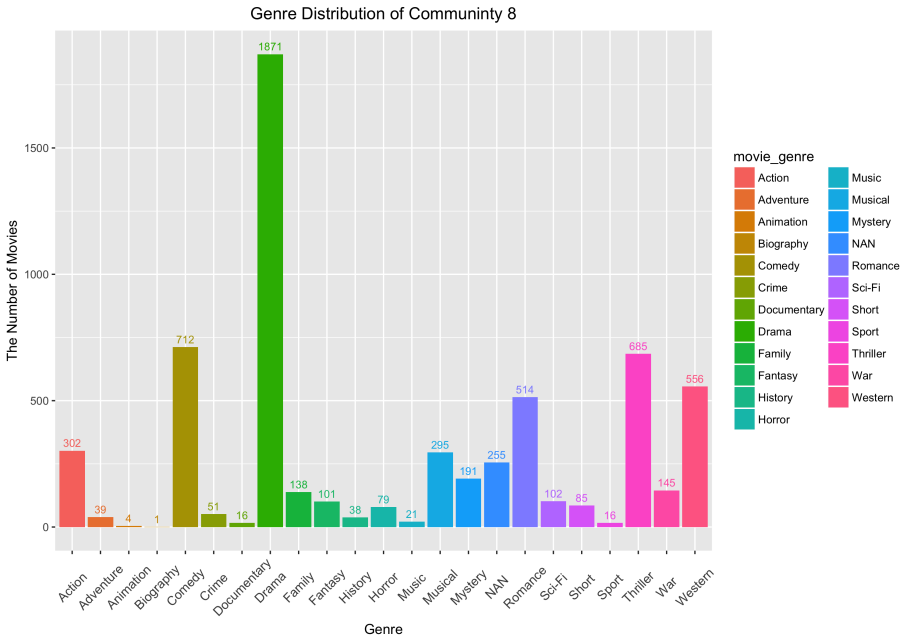

Figure 12: Genre Distribution of Community 8

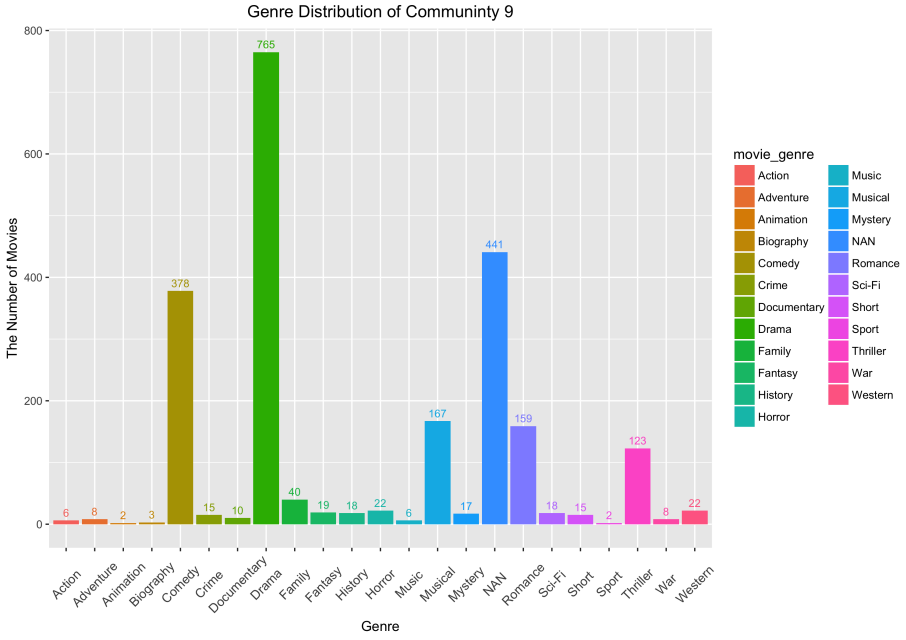

Figure 13: Genre Distribution of Community 9

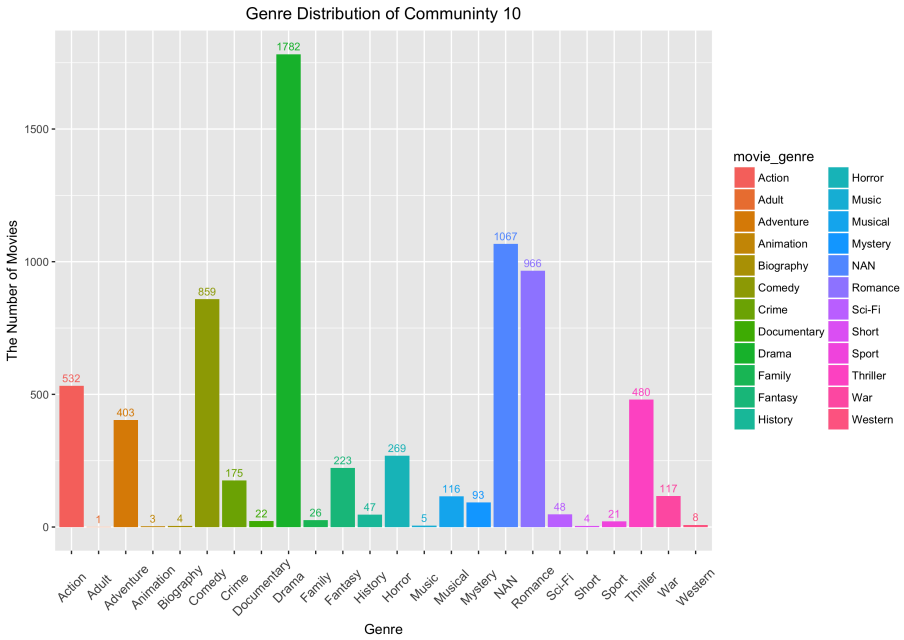

Figure 14: Genre Distribution of Community 10

Based on simple frequency counts, the most dominant genre of each community we got is shown in Table 3. As we can see, the genre "Drama" is the dominant genre of 11 communities among 28 communities. This in fact can be easily speculated, because in the movie network, the number of moives marked "Drama" take a high proportion among all the movies. What is worth to mention is that there are some movies without genre mark, which in my charts marked "NAN", and this kind of movie in few communities actually become the category with most number of movies, such as community 28 (the distribution shown in Figure 15), where only one movie belongs to "Short" and the rest of movies marked "NAN"; however we need to ignore "NAN" movies, so the donimant genre should be "Short" in this community.

Table 3: The dominant genre of each community(according to frequency count)

Community ID	**the dominant genre**
1	Thriller
2	Short
3	Drama
4	Drama
5	Drama
6	Drama
7	Drama
8	Drama
9	Drama
10	Drama
11	Drama
12	Drama
13	Drama
14	Drama
15	Drama
16	Drama
17	Drama
18	Drama
19	Drama
20	Drama
21	Drama
22	Drama
23	Drama
24	Adult
25	Thriller
26	Short
27	Short
28	Short

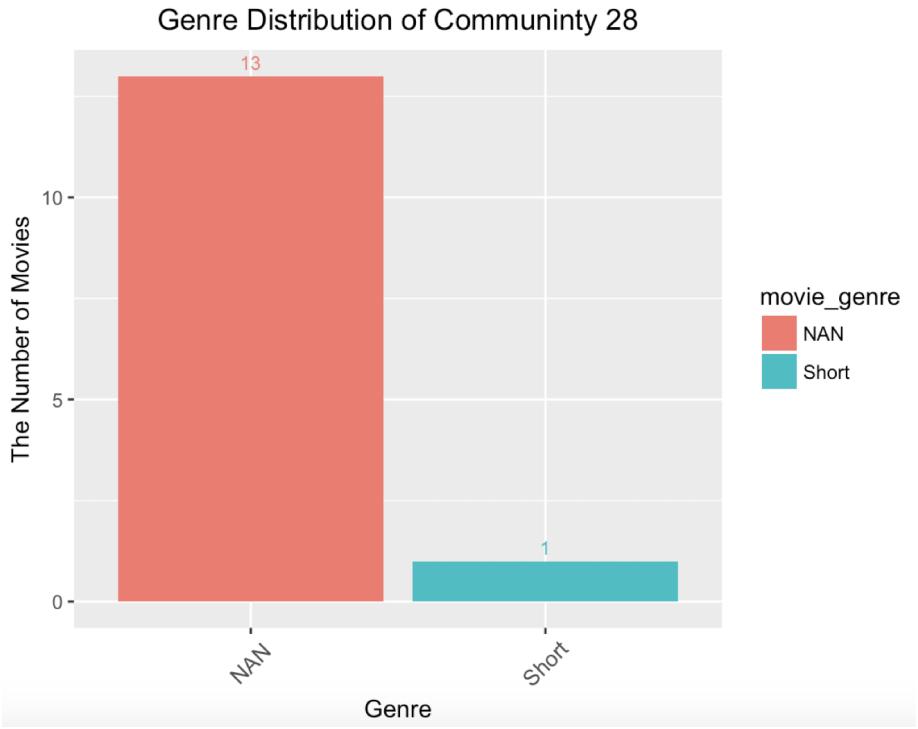

Figure 15: Genre Distribution of Community 28

Based on the measurement of modified score, the dominant genre of each community changes a lot, and the result is shown in Table 4. It illustrates that most of the dominant genres we got according to the modified scores are different from the genres we got by simply couting the frequency of genre within the community. Interestingly, there do exist some communities hold the same dominant genre and all of them have comparably fewer movies than other communities. To be sepecific, the number of movies in these communities is in the range of (10,20). In addition, the frequency of the dominant genre in these communities is far more higer than other genres. For instance, we also plotted community with ID=24 and ID=25, shown in Figure 16 and 17. In fact, this is reasonable because in these kinds of communities, $p(i)$ of the dominant genre(the genre has the highest frequency) is much bigger than other genres within the community, which can play a dominant role of the score function. Moreover, some of the genres in these kinds of communities have only one occurency, which makes $c(i)$ equals to 1 and then $ln(c(i))$ equals to 0, so that the score will be 0. On the contrary, for those communities whose dominant genre differ from different measurement, they generally include large number of movies, and may have genres holding the similar frequency wihthin the communiy. For instance, we could observe the genre distribution of community with ID=6 (Figure 10), the dominant genre changes from "Drama"(the most frequent genre) to "Comedy"(the third frequent genre). This is easy

to understand since the score function consider the fraction of genre within community and also in whole dataset, which is equvalent to multiply some coefficients to the exact frequency of the genre, this somehow makes the score bigger or smaller comparing to the exact frequency. Therefore, the dominant genre of these kinds of community changes.

Table 4: The dominant genre of each community(according to modified score)

Community ID	**the dominant genre**
1	Adult
2	Film-Noir
3	War
4	Crime
5	Family
6	Comedy
7	Family
8	Musical
9	Musical
10	Adventure
11	Family
12	Romance
13	War
14	Adventure
15	Comedy
16	Musical
17	Action
18	Drama
19	Fantasy
20	Comedy
21	Action
22	Romance
23	Short
24	Adult
25	Thriller
26	Short
27	Short
28	Short

We chose community(ID=24) who includes 12 movies to further study the correlation among actors, movies and genres. Table 5 shows the basic informaiton of this community, while Table 6 and Table 7 list the ID and Name mapping correlation in movies and actors respectively.

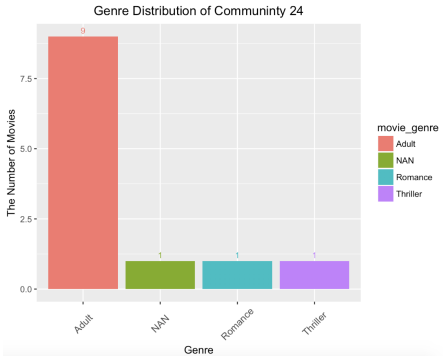

Figure 16: Genre Distribution of Community 24

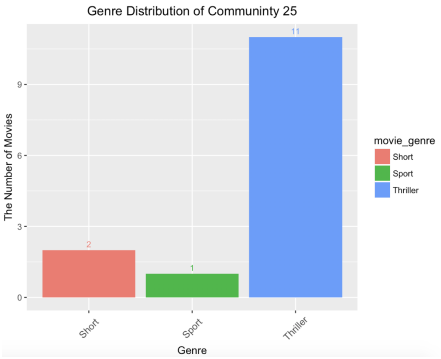

Figure 17: Genre Distribution of Community 25

Table 5: Basic Information of Community 24

Movie ID	Genre	Actor ID
93381	Thriller	4196,19653,26390,35604,45645,56587,60399,66920
151826	Adult	8242,20204,21948,60399,89726
151830	Adult	8242,21948,60399,93363,111639
151831	Adult	8242,21948,32885,60399,68271,83742
151832	Adult	8242,21948,32885,60399,83742
151838	Adult	8242,24483,65193,68271,102098
151840	Adult	8242,20204,21948,60399,89726,11639
151841	Romance	8242,21948,32885,68271,83742
151842	Adult	8242,20204,21948,60399,83742,111639
151845	NAN	8242,21948,32885,60399,69507
251442	Adult	20204,32885,60399,68271,93363
251446	Adult	20204,21948,32885,89726,102098

Table 6: Movie ID mapping to Movie Name (for community 24)

Movie ID	Movie Name
93381	Baise-moi (2000)
151826	Aleska & Angelika: Pornochic 21 (2011)
151830	Dorcel Airlines: Paris/New York (2010)
151831	Initiation of Lou Charmelle (2010)
151832	Jade, Secretaire de Luxe (2011)
151838	Maximum Orgy, spÈcial pin-up (2012)
151840	Russian Institute: Anal Lesson (2010)
151841	Soubrettes Services (2010)
151842	Soubrettes Services: Special Stars (2010)
151845	Story of MÈgane (2008)
251442	Hard Intrusion (2007)
251446	Russian Institute Lesson 16: Lolitas (2011)

Table 7: Actor ID mapping to Actor Name (for community 24)

Actor ID	Actor Name
4196	Barrio, Sebastian (I)
8242	Brossman, James
19653	Embarek, Ouassini
20204	Evil, Leny
21948	Forte, Alex
24483	Giotto, Lauro
26390	Gustave, HervÈ P.
32885	JPX
35604	Kodjo Topou, Patrick
45645	MinÈo, Jean-Marc
56587	Rioufol, Marc
60399	Scott, Ian (III)
65193	SX, Bruno
66920	Titof
68271	Uhl, George
69507	Vidal, Nacho (I)
83742	Dollar, Cindy
89726	Hope, Cindy
93363	Lafitte, Yasmine
102098	Polina, Anna
111639	White, Tarra

The bipartite graph we got is shown in Figure 18, where red vertices represen the actors and blue vertices represent the movies.

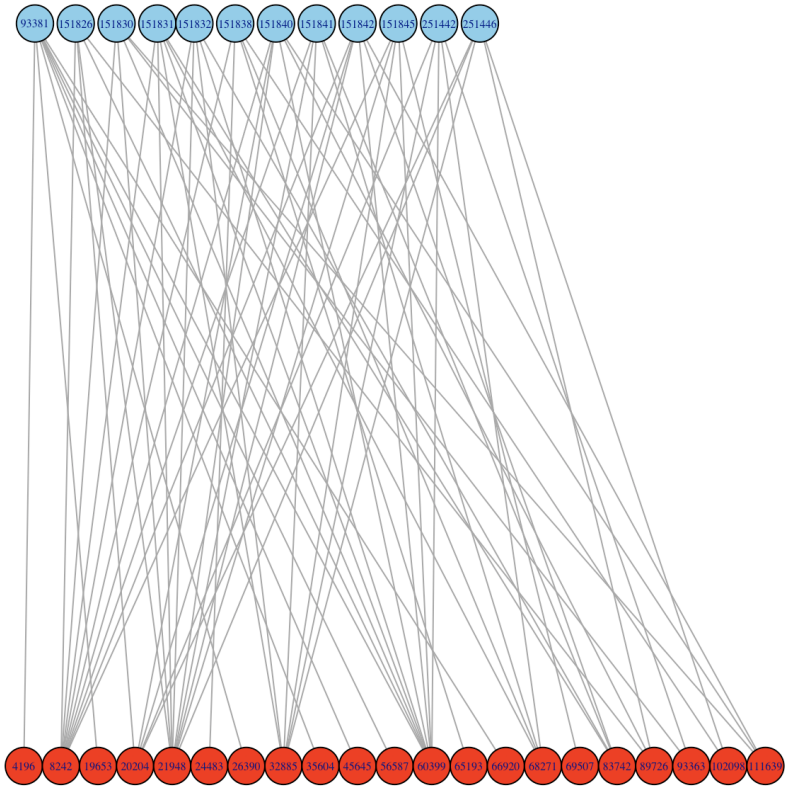

Figure 18: Bipartite Graph of Movies and Actors in Community 24

After statistical analysis, we found that the three most important actors'ID are 8242 21948 60399. Then, we searched all the movies these three actors have ever acted. Not surprisingly, the dominant genre of these movies is also "Adult", the same as the dominant genre of the community 24 that we got in we got in 8(a) and 8(b). That is to say, these three most important actors involved in the same kind of movies, which determines the community and the dominant genre of the community. In other words, the community we got by the Fast Greedy algorithm helps us find movies in same(or similar) genre.

2.3 Neighborhood analysis of movies

In this section, we analyzed the relationship between the rating of a movie and its neighbors. Specifically, we mainly investigated the following three movies.

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

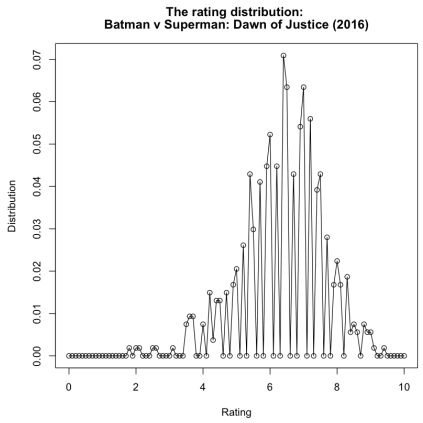

Figure 19: The rating distribution of the neighbor network
Batman v Spiderman: Dawn of Justice (2016)

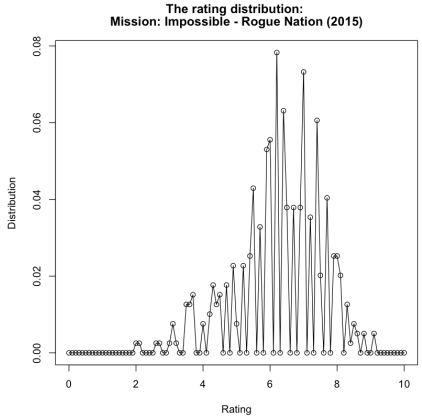

Figure 20: The rating distribution of the neighbor network
Mission: Impossible - Rogue Nation (2015)

- Minions (2015); Rating 6.4.

Question 9: For each of the movie listed above, extract it's 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.

After extracting the neighbors of the above three movies and computing the mean of the available ratings in the neighbor vertices, we obtained the Fig. 19, 20, and 21. The average ratings of the neighbor network of the above three movies are shown in Table. 8. It can be seen that, with this scheme, the average rating in the neighbor network is not very similar to the actual rating of the movie.

Table 8: Average rating vs. actual rating of the three movies

Movie Name	Actual Rating	Average Rating
Batman v Spiderman: Dawn of Justice (2016)	6.6	6.4
Mission: Impossible - Rogue Nation (2015)	7.4	6.3
Minions (2015)	6.4	6.9

Question 10: Repeat question 9, 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.

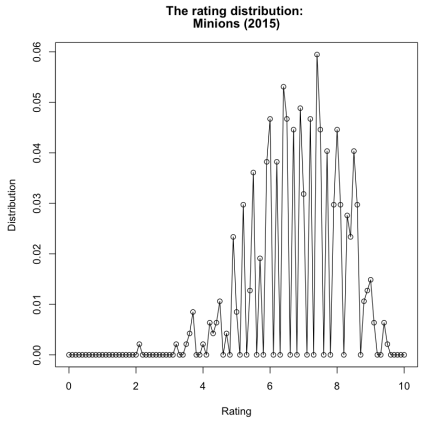

Figure 21: The rating distribution of the neighbor network
Minions (2015)

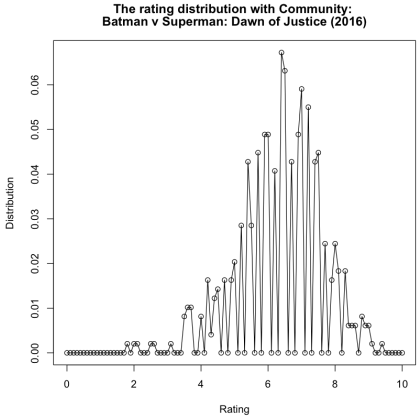

Figure 22: The rating distribution of the neighbor network with the same community
Batman v Spiderman: Dawn of Justice (2016)

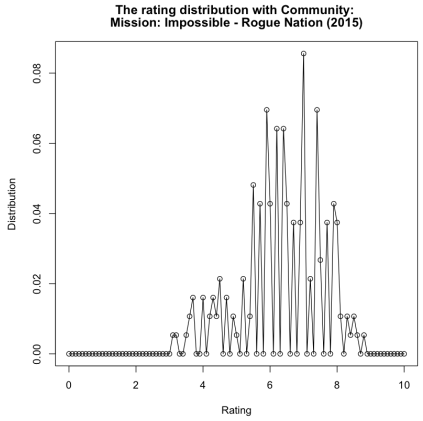

Figure 23: The rating distribution of the neighbor network with the same community
Mission: Impossible - Rogue Nation (2015)

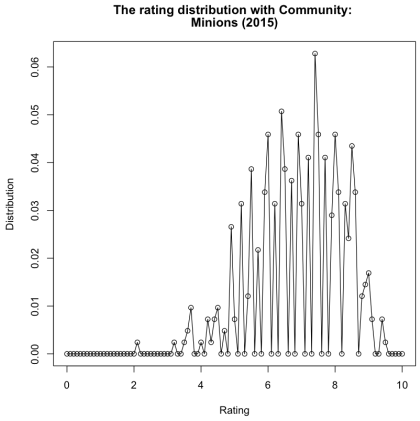

Figure 24: The rating distribution of the neighbor network with the same community
Minions (2015)

After constraining the community, the rating distributions we obrtained are shown in Fig. 22, 23, and 24. The average ratings of the neighbor network of the above three movies are shown in Table. 9. It can be seen that, after contraining the neighbor vertices are in the same community, the average rating in the neighbor network is still not very similar to the actual rating of the movie.

Table 9: Average rating vs. actual rating of the three movies, with community considered

Movie Name	Actual Rating	Average Rating
Batman v Spiderman: Dawn of Justice (2016)	6.6	6.4
Mission: Impossible - Rogue Nation (2015)	7.4	6.4
Minions (2015)	6.4	6.9

Question 11: For each of the movies listed above, extract it's 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 edge values in decreasing order in the neighbor network, the result is shown in Table. 10.

Table 10: Top 5 neighbors of the three movies

Movie Name	Top 5 Neighbors	Community ID
Batman v Spiderman: Dawn of Justice (2016)	Eloise (2015)	1
	The Justice League Part One (2017)	1
	Into the Storm (2014)	1
	Love and Honor (2013)	1
	Man of Steel (2013)	1
Mission: Impossible - Rogue Nation (2015)	Fan (2015)	5
	Phantom (2015)	5
	Breaking the Bank (2014)	4
	Suffragette (2015)	4
	Now You See Me: The Second Act (2016)	1
Minions (2015)	The Lorax (2012)	1
	Inside Out (2015)	1
	Up (2009)	1
	Despicable Me 2 (2013)	1
	Surf's Up (2007)	1

2.4 Predicting ratings of movies

In this section, we explore how to use different models to predict the rating of a movie. Still, we try to predict the ratings of the following three movies.

- Batman v Spiderman: Dawn of Justice (2016);

- Mission: Impossible - Rogue Nation (2015);
- Minions (2015).

For regression model, we chosed five features: the top 5 pagerank values of the actors who involved in the movie. We first removed those movies that do not have rating record in our dataset, and then randomly selected 70 percentage of the data as the training data and the rest 30 percentage as the test data. Then, we get RMSE = 1.26431431866 when evaluating our model.

The movies listed above do not have rating record in the dataset, so the predicted ratings of them we got is: XXXXXXXXXXXX

The other way we used to predict ratings for movies is to create a bipartite graph corresponding to actors and movies. The metric we designed for assigning a weight to each actor is to calculate the average rating values of the movies an actor has acted and set this value as the weight of each actor. When implementing this rating mechanism, we found that there are several actors who do not have a weight (score) becasue all the movies he/she involved in do not have rating records. Therefore, we ignore these kind of actors. Then the rating prediction step is just to find all the actors involved in a certain movie and compute the average score of all the actors as the rating of the movie.

Applying this method, RMSE = 1.023422, and the predicted ratings are shown in Table. 11.

Table 11: Average rating vs. actual rating of the three movies, with community considered

Movie Name	Actual Rating	Predicted Rating
Batman v Spiderman: Dawn of Justice (2016)	6.6	6.363
Mission: Impossible - Rogue Nation (2015)	7.4	6.425
Minions (2015)	6.4	6.858