**ECE232E - Project 4**

**IMDb Mining**

*Zhechen Xu (805030074), Yufei Hu (404944367), Qi Zeng (204946904), Liangkun Zhao (204947003)*

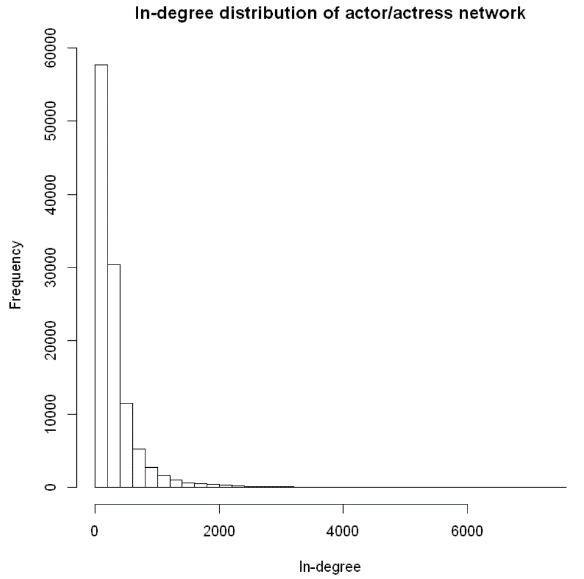
*June 2, 2018*

# **Part 1: Actor/Actress network**

**Q1:**

After cleaning and merging the two text files, 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 **468150**. All actors or actresses who have acted in less than 10 movies are removed.

**Q2:**



**Figure 1:** In-degree distribution of actor/actress network

This directed network is built based on the equation below:



where *Si* is the set of movies in which actor/actress *vi* has acted in.

The in-degree distribution is shown in the figure above. It indicates that most actors or actresses have a low in-degree while only a few of them own a high in-degree. This observation matches the intuition in the sense that only a limited amount of old and hard-working actors or actresses would have the chance of cooperating with hundreds or even thousands of other actors or actresses.

**Q3:**

**Table 1:** Actor parings

|  |  |  |
| --- | --- | --- |
| **Input Actor** | **Output Actor** | **Edge Weight** |
| Hanks, Tom | Allen, Tim (I) | 0.1013 |
| Depp, Johnny | Bonham Carter, Helena | 0.0816 |
| Streep, Meryl | De Niro, Robert | 0.0619 |
| Clooney, George | Damon, Matt | 0.1194 |
| DiCaprio, Leonardo | Scorsese, Martin | 0.1020 |
| Johnson, Dwayne (I) | Austin, Steve (IV) | 0.2051 |
| Pitt, Brad | Clooney, George | 0.0986 |
| Cruise, Tom | Kidman, Nicole | 0.1746 |
| Smith, Will (I) | Foster, Darrell | 0.1224 |
| Watson, Emma (II) | Radcliffe, Daniel | 0.5200 |

The table above surprisingly makes sense as Emma Watson has indeed cooperated with Daniel Radcliffe in the world-famous movie series *Harry Potter*. The ex-wife of Tom Cruise is exactly Nicole Kidman and they also have acted in many movies before. Similar connections can also be found in all other 8 pairs after some googling. All these facts verified the soundness of our paring finding algorithm.

**Q4:**

**Table 2:** Top 10 actor/actress based on pagerank algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| **Actor/Actress** | **Pagerank Scores** | **Number of Movies** | **In-Degree** |
| Flowers, Bess | 0.000235 | 828 | 7537 |
| Tatasciore, Fred | 0.000199 | 353 | 3954 |
| Harris, Sam (II) | 0.000197 | 600 | 6960 |
| Blum, Steve (IX) | 0.000195 | 373 | 3316 |
| Miller, Harold (I) | 0.000173 | 561 | 6587 |
| Jeremy, Ron | 0.000164 | 637 | 3177 |
| Phelps, Lee (I) | 0.000158 | 647 | 5563 |
| Lowenthal, Yuri | 0.000157 | 317 | 2662 |
| Downes, Robin Atkin | 0.000152 | 267 | 2953 |
| O'Connor, Frank (I) | 0.000147 | 623 | 5502 |

None of the top 10 actors and actresses appear in the list before. According to the table above, I find that these actors and actresses have participated in many movies. For example, Bess Flowers was actually born in the 19th century, and she was “best known for her work as an extra in hundreds of films” according to Wikipedia. She has connections with over 800 movies. This pagerank algorithm successfully returns the actors and actresses with roughly most connections with other actors or actresses. However, most of them are not well known to the public compared to those actors and actresses in Table 1 due to the fact that the movies they participated in are not so famous or they hid behind the scene as voice actors or so.

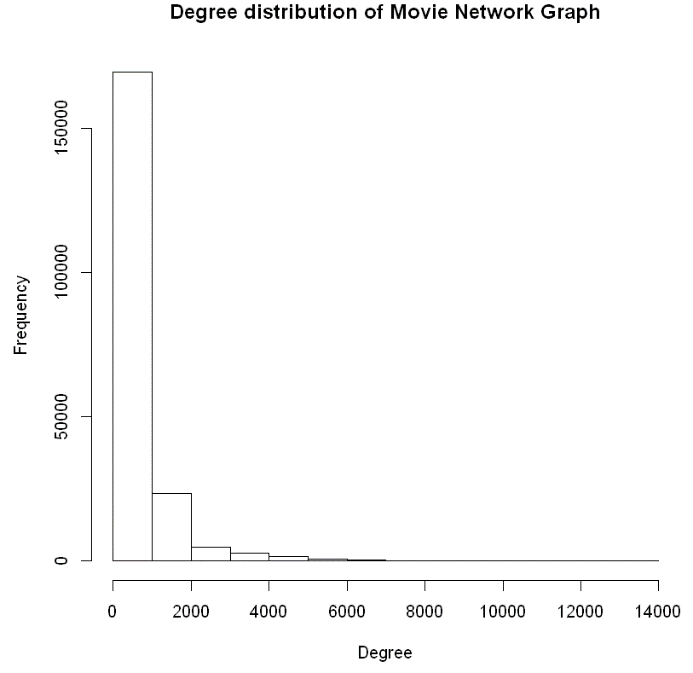
**Q5:**

**Table 3:** Pagerank scores of the specific actors and actresses

|  |  |  |  |
| --- | --- | --- | --- |
| **Actor/Actress** | **Pagerank Scores** | **Number of Movies** | **In-Degree** |
| Depp, Johnny | 0.000054 | 98 | 2144 |
| Hanks, Tom | 0.000051 | 79 | 2064 |
| Pitt, Brad | 0.000043 | 71 | 1739 |
| Johnson, Dwayne (I) | 0.000042 | 78 | 1357 |
| Clooney, George | 0.000040 | 67 | 1573 |
| Cruise, Tom | 0.000040 | 63 | 1651 |
| Streep, Meryl | 0.000040 | 97 | 1594 |
| Smith, Will (I) | 0.000032 | 49 | 1319 |
| DiCaprio, Leonardo | 0.000032 | 49 | 1301 |
| Watson, Emma (II) | 0.000017 | 25 | 453 |

# **Part 2: Movie network**

**Q6:**



**Figure 2:** Degree distribution of movie network

Similarly, an undirected movie network is built based on the equation below:



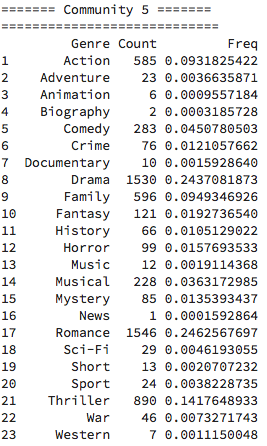
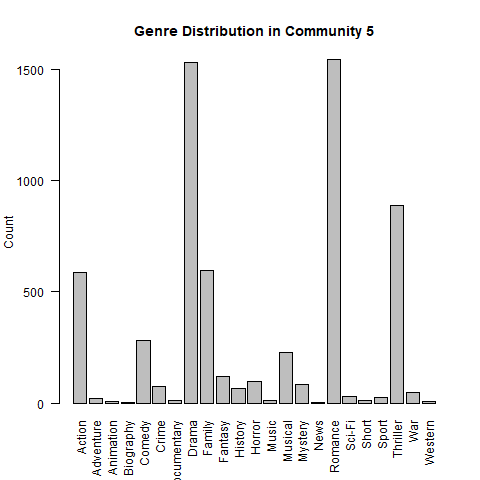
where *Ai* is the set of actors in movie *vi*.

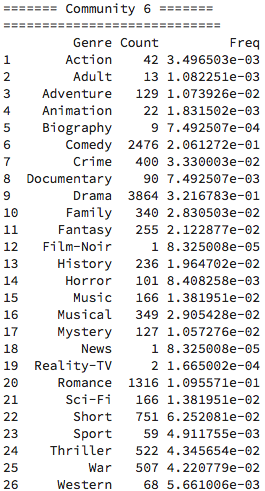
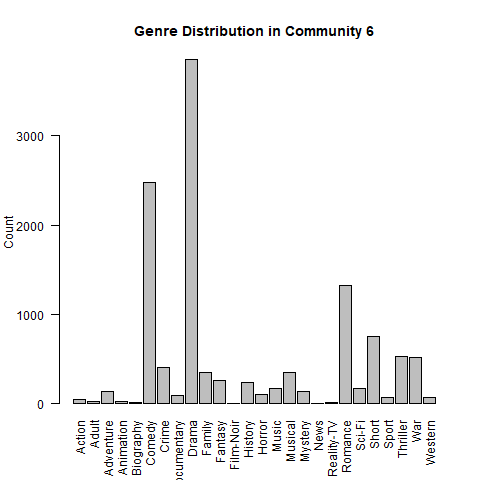
The figure above looks similar to the in-degree distribution of the actor/actress network. However, it seems that most vertices in the movie network own a smaller degree compared to those in the actor/actress network.

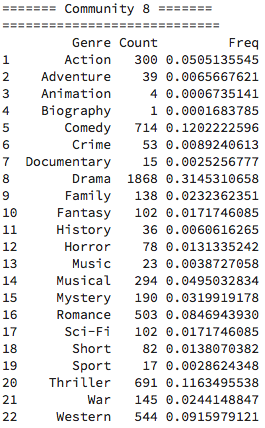
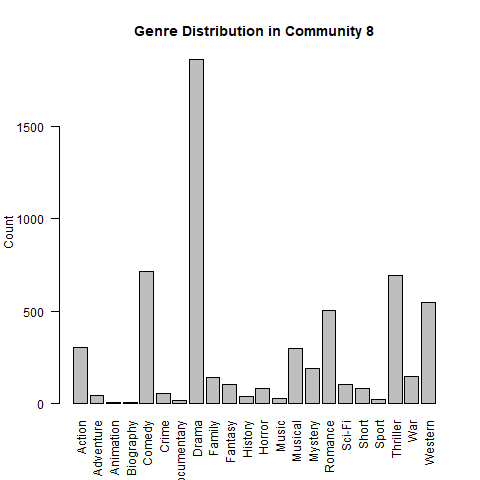
**Q7:**

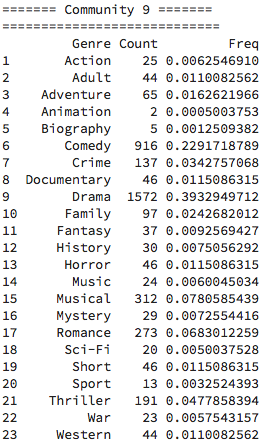
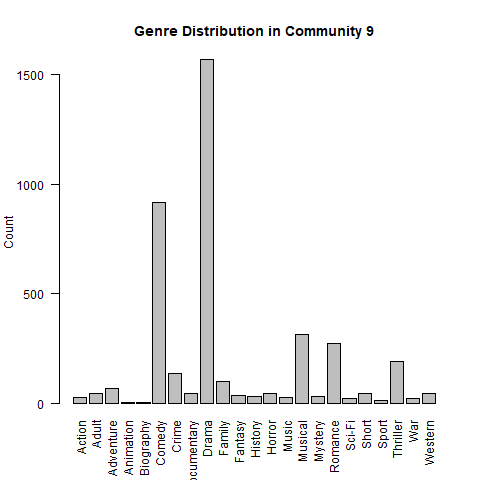
In Question 7, we used the fast greed community detection algorithms to find the communities in the movie network. There are 28 communities in the network. We labeled “None” as a genre to movie which has no genre label. To keep the original network structure, we found communities in the network first, then ignored the “None” genre movies after finding.

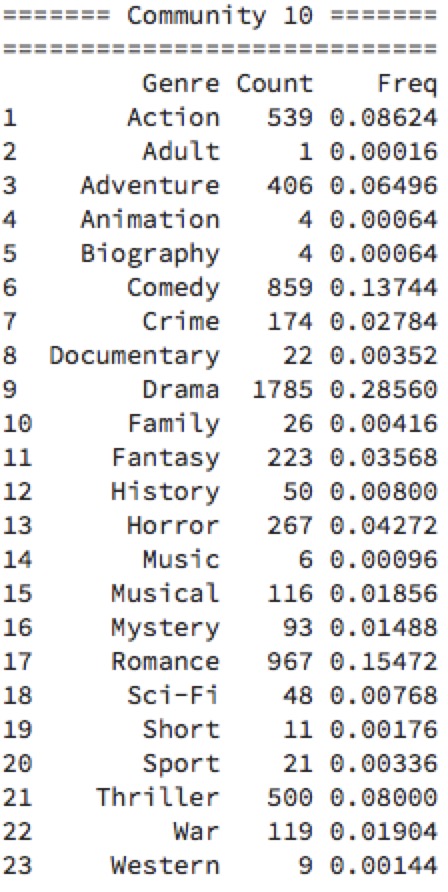
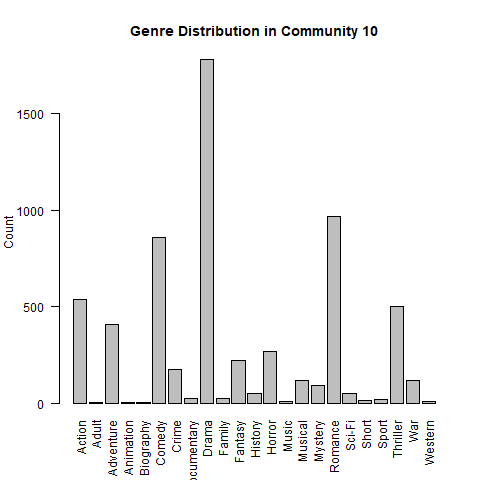
We randomly picked 10 communities, and plotted the distribution of the genres of the movies in these community.

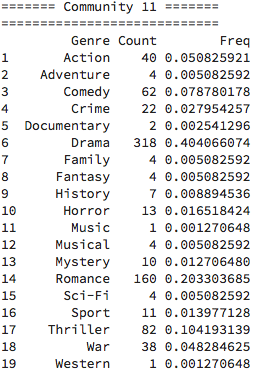
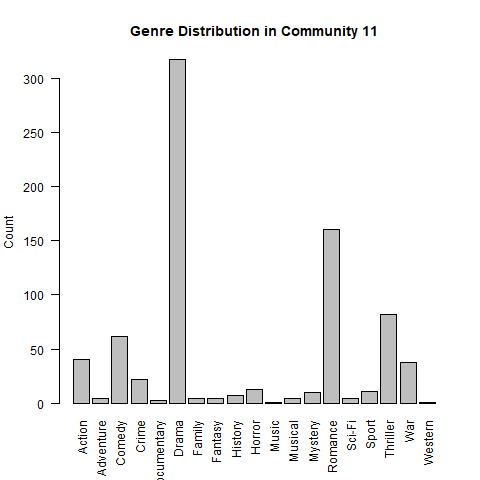


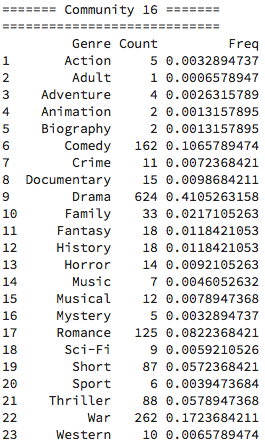
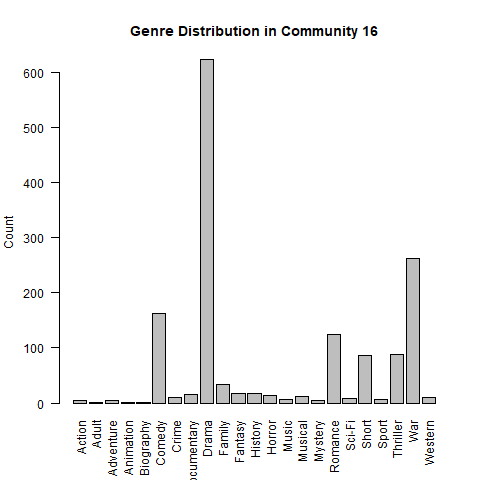


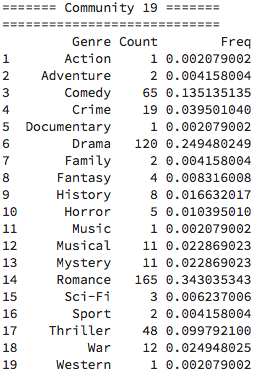
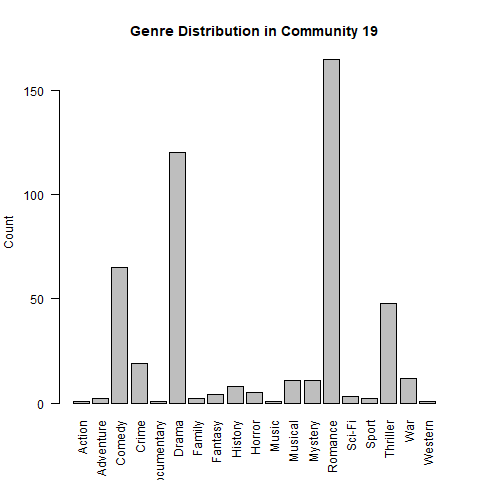


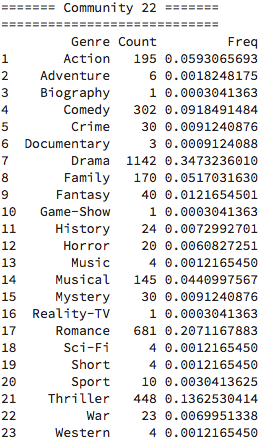
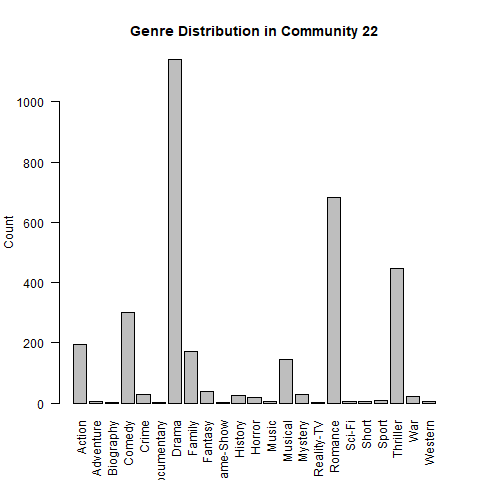


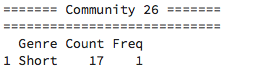
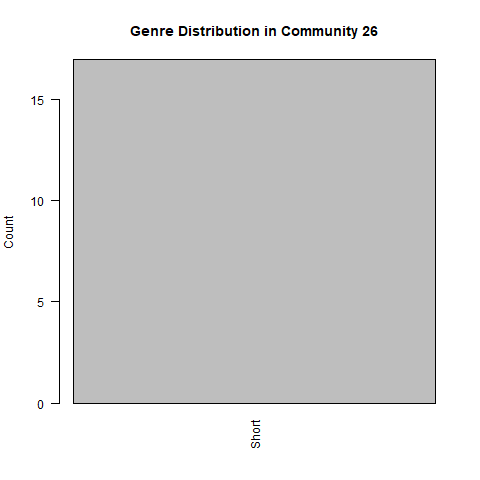












**Figure 3:** Genre distribution of 10 randomly selected communities

Q8(a):

Based on frequency counts, the following genres are the most dominant genre in these communities.

|  |  |
| --- | --- |
| Community | The Most Dominant Genre |
| 5 | Romance |
| 6 | Drama |
| 8 | Drama |
| 9 | Drama |
| 10 | Drama |
| 11 | Drama |
| 16 | Drama |
| 19 | Romance |
| 22 | Drama |
| 26 | Short |

The number of romance-genre movies is only slightly more than the number of drama-genre movies in community 5. Because of the large amount of drama-genre movies and actors who act in such kind of movies in the whole dataset, drama is the most dominant genre in most community based on frequency counts.

Q8(b):

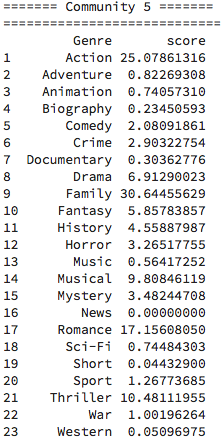
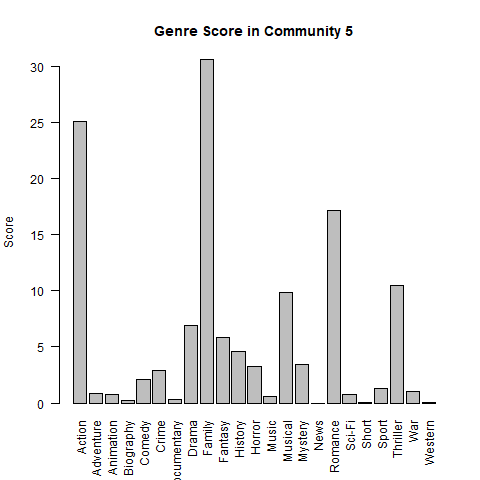
In question 8b, we used an another rule to value the most dominant genre:

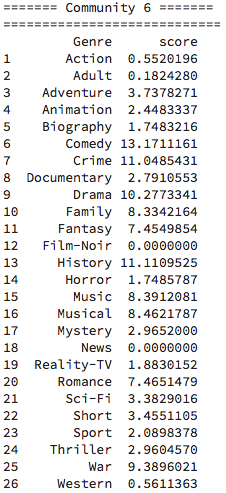
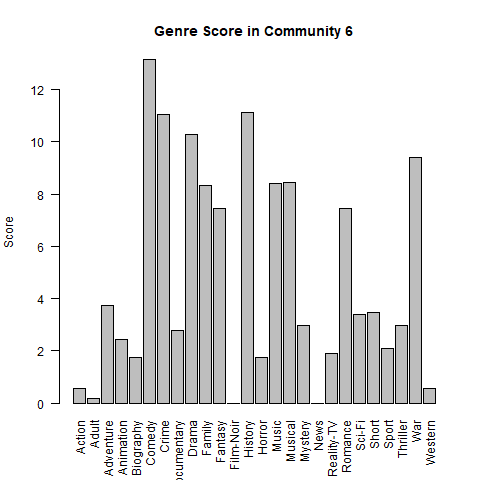
c(i): The number of movies belonging to genre i in the community.

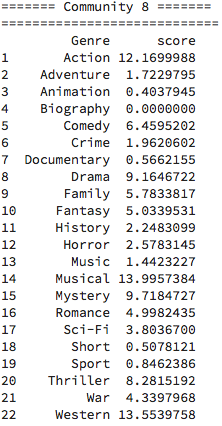
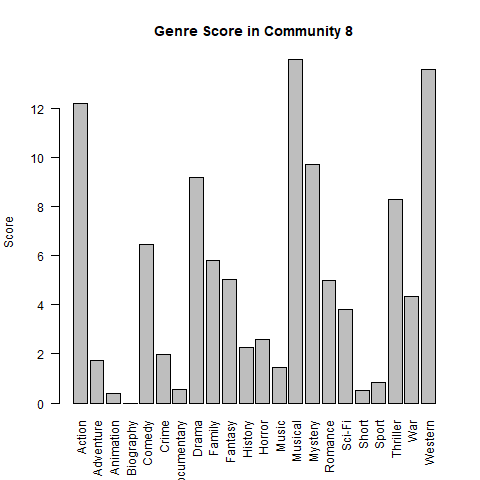
p(i): The fraction of genre i movies in the community

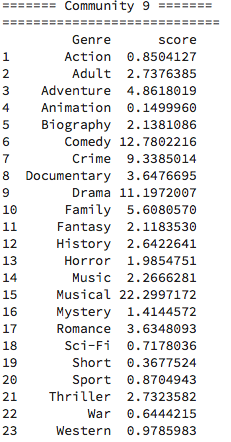
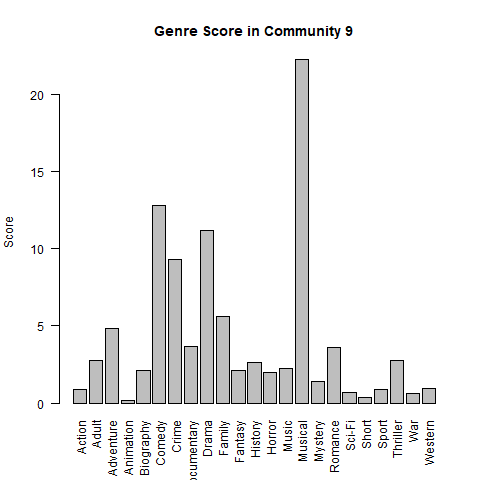
q(i): The fraction of genre i movies in the entire dataset.

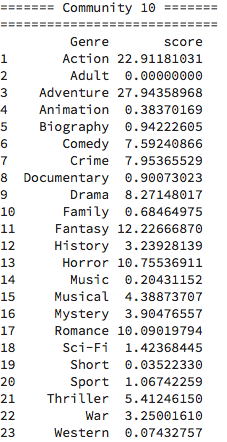
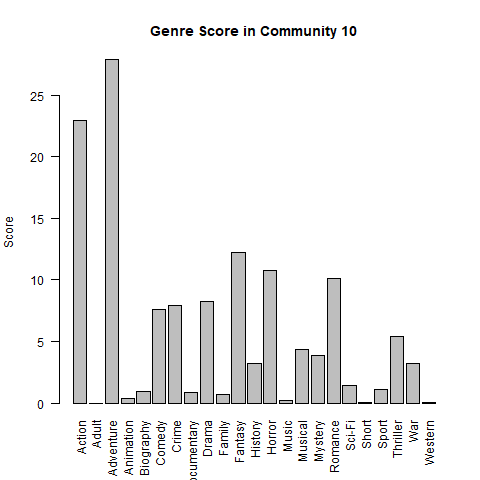
The plots below are the score plots for each genre in the previous 10 communities.

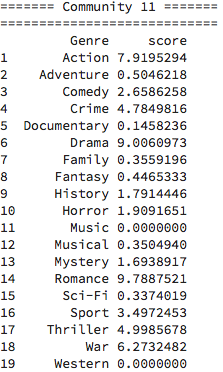
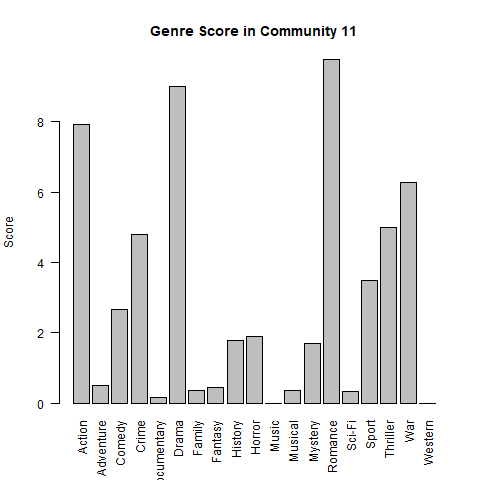


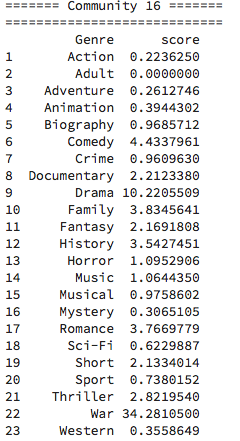
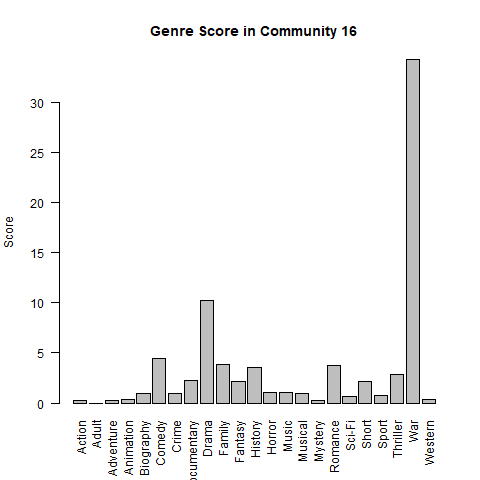


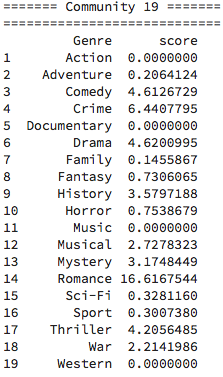
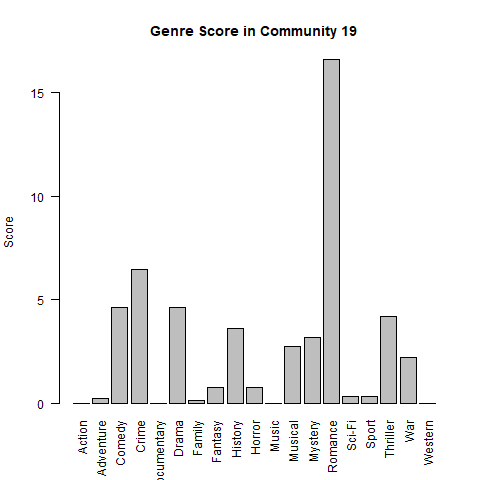


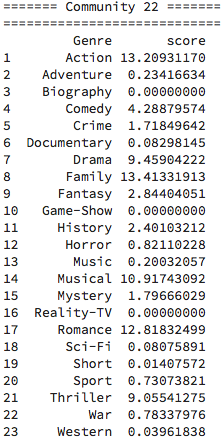
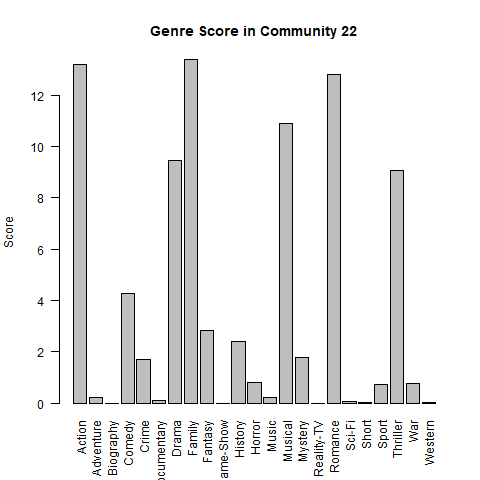


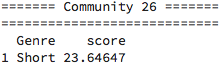
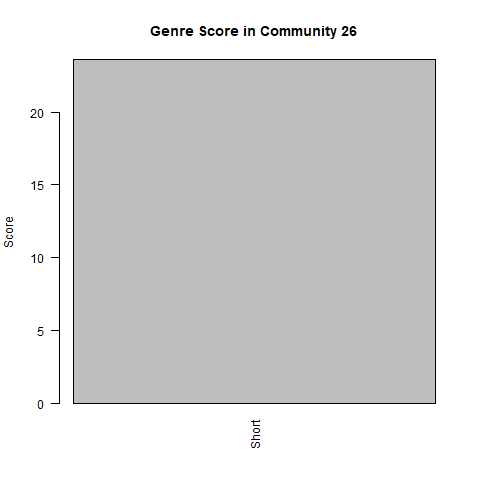












**Figure 4:** Genre score of 10 randomly selected communities

We can tell that the plots of genre score are very different from the plots of genre distribution. Based on score, the following genres are the most dominant genre in these communities.

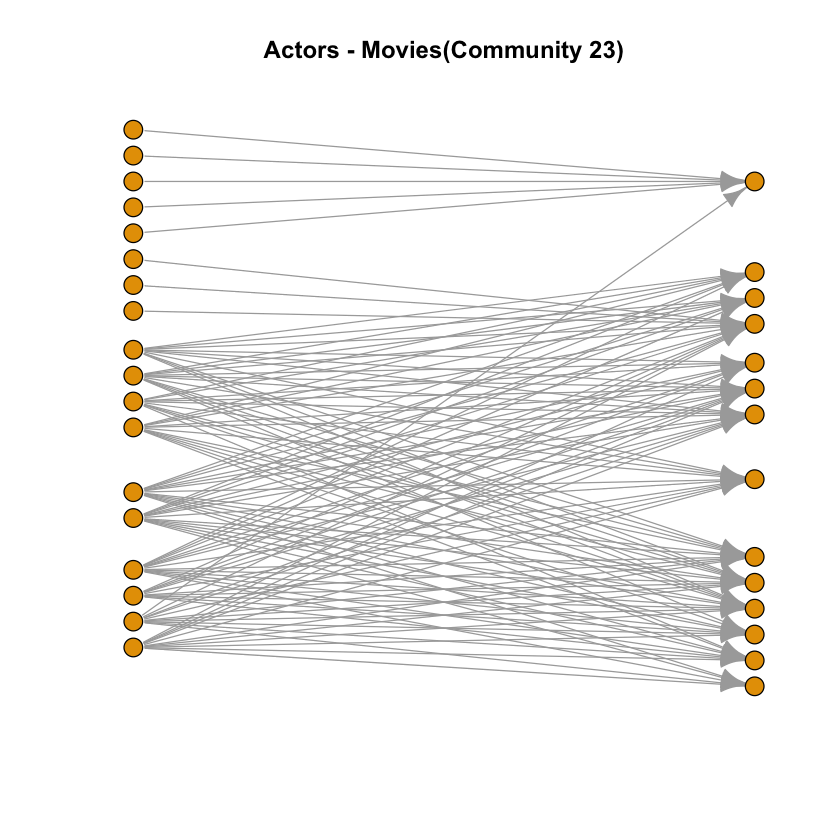
|  |  |
| --- | --- |
| Community | The Most Dominant Genre |
| 5 | Family |
| 6 | Comedy |
| 8 | Musical |
| 9 | Musical |
| 10 | adventure |
| 11 | Romance |
| 16 | War |
| 19 | Romance |
| 22 | Family |
| 26 | Short |

From the table above, it shows that unlike the dominant genre based on frequency counts, in which case there can be a genre that is dominant for most part of communities, the dominant genre based on the score is various from community to community. We can find out the reason why such phenomenon happens by looking into the score function:

The score is proportional to ln(c(i)) \* p(i), which means if the number of a genre and the fraction of this genre in a community is large, the score tends to be large. However, the score does not only depend on the amount of a genre in the dataset. It is also inverse proportional to q(i), which works as a constraint saying “the larger is not always the better”, and that is the most important difference between score and frequency count. Score depends on both of amount of a genre and the “purity” of this genre. The purity measures how dense a genre distributed in communities. That is the reason why drama is not the most dominant genre based on score, because it distributes in many communities. In other word, the purity of drama is low. However, genres such as family and short are densely distributed in specific communities even though the total amount of these genres is not large, the score of these genres is much larger than other low-purity genres.

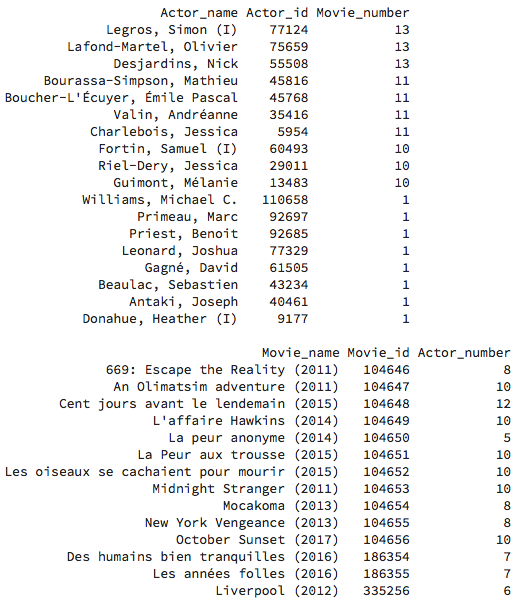
Q8(c):

In this question, we found 3 communities that has size between 10 and 20. Below is the bipartite plot and information of each community.

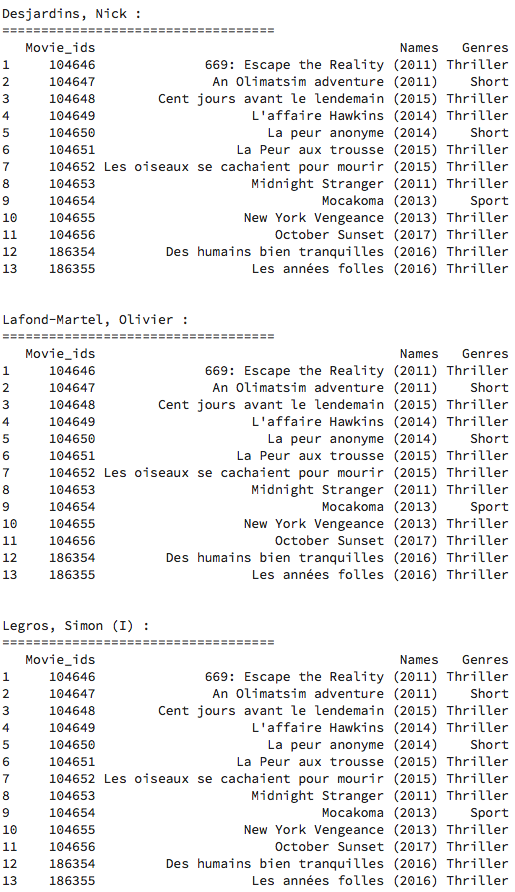


**Figure 5:** Actor – Movie Bipartite Plot of Community 23

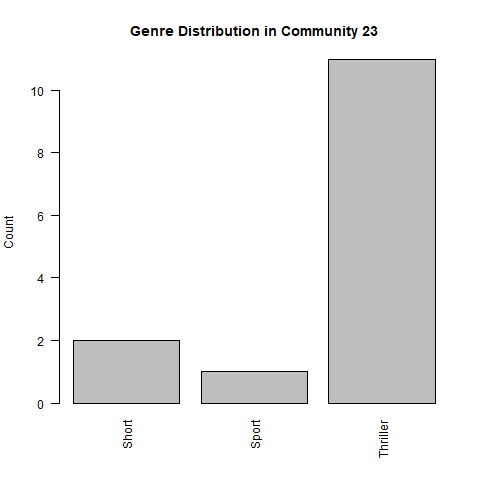
Actors and movies in this community:



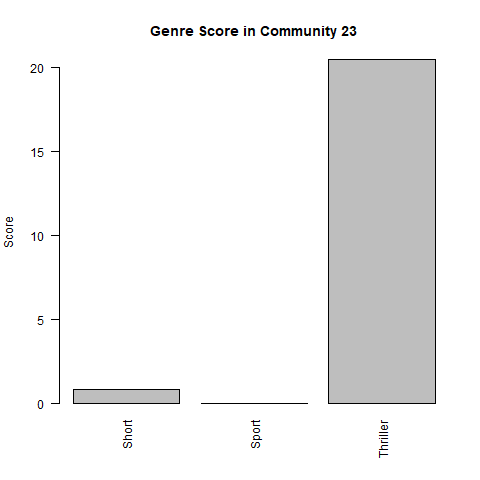
Top three actors information in this community:

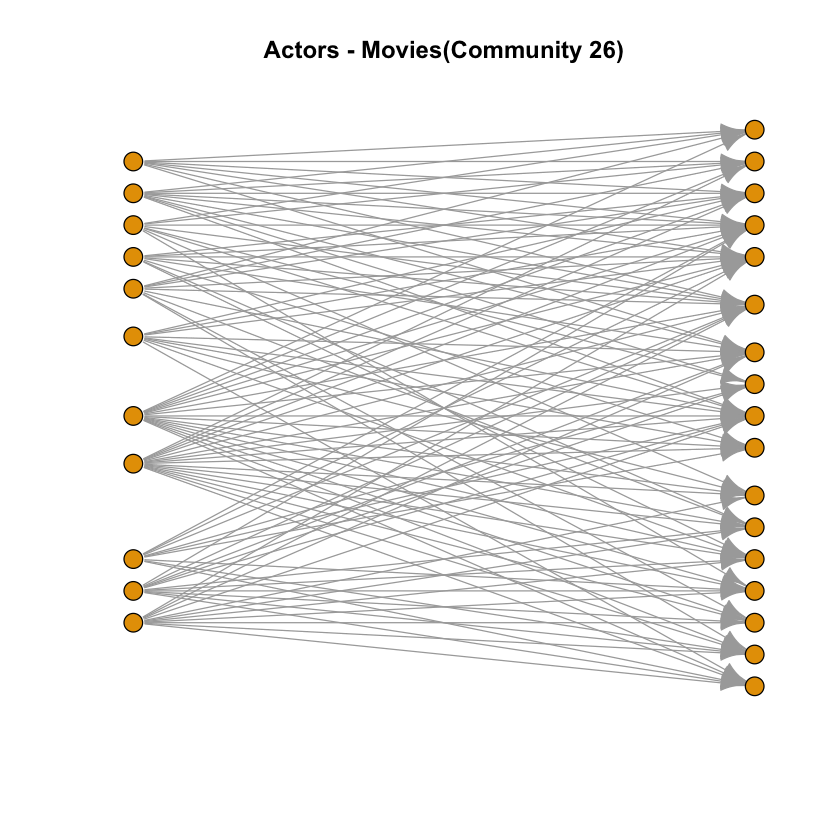


The genre distribution of this community:



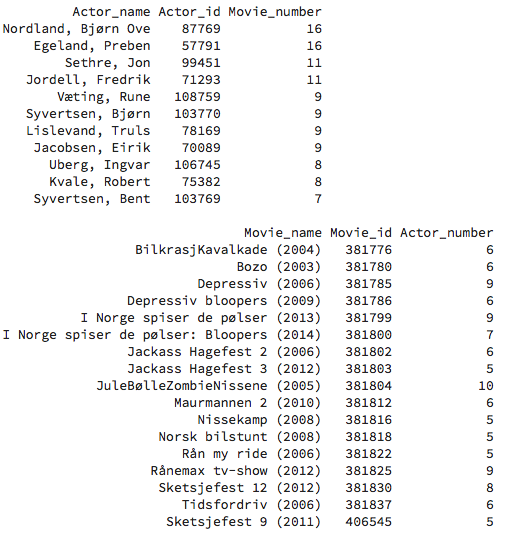
The score of each genre in this community:



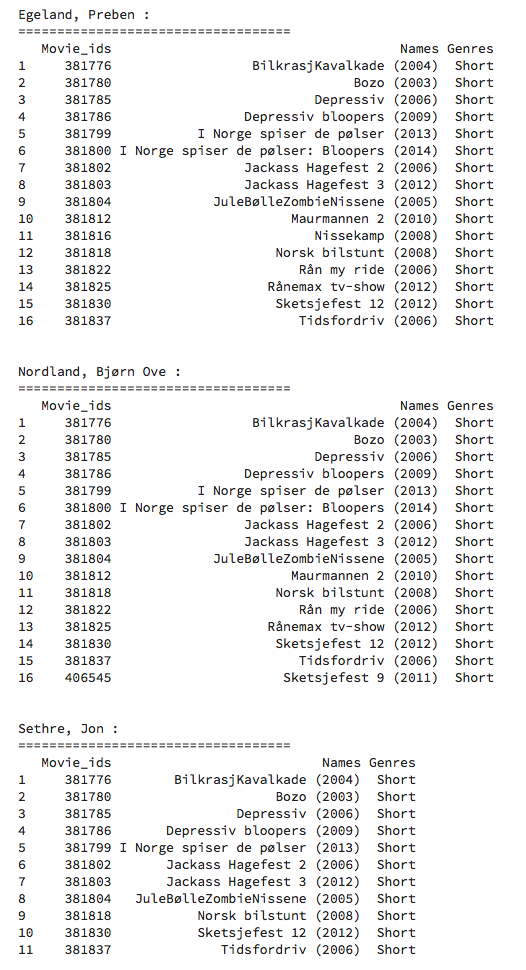


**Figure 6:** Actor – Movie Bipartite Plot of Community 26

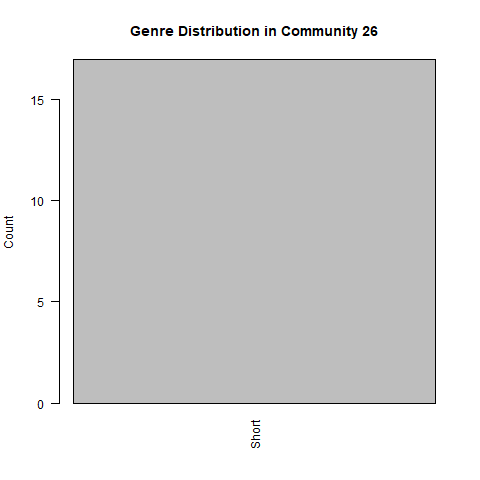
The actors and movies in this community:



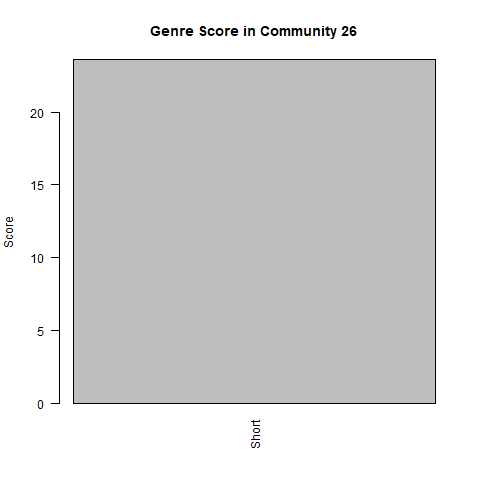
The top three actors information in this community:

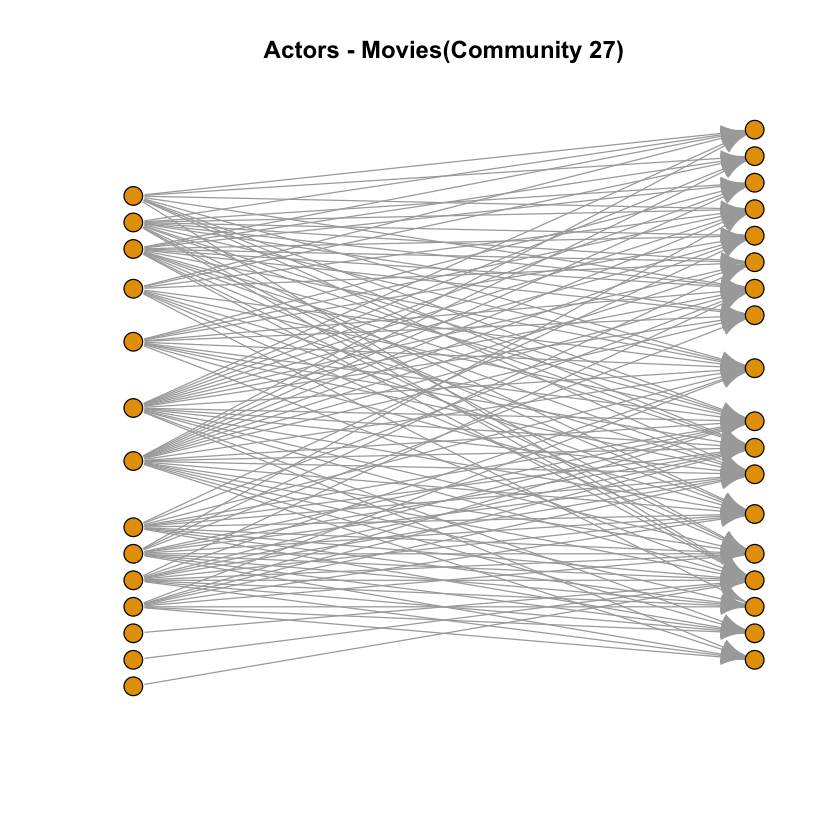


The genre distribution in this community:



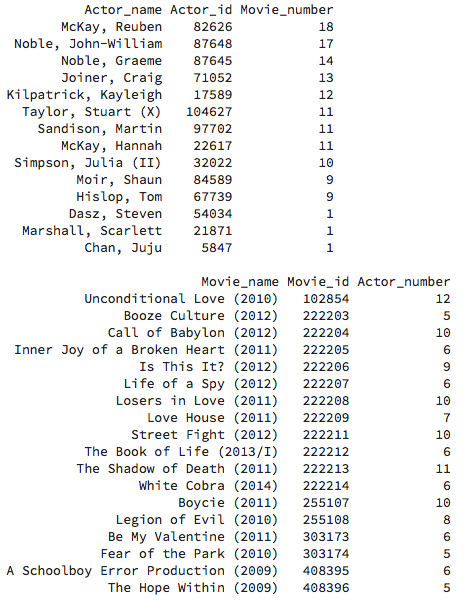
The score of each genre in this community:



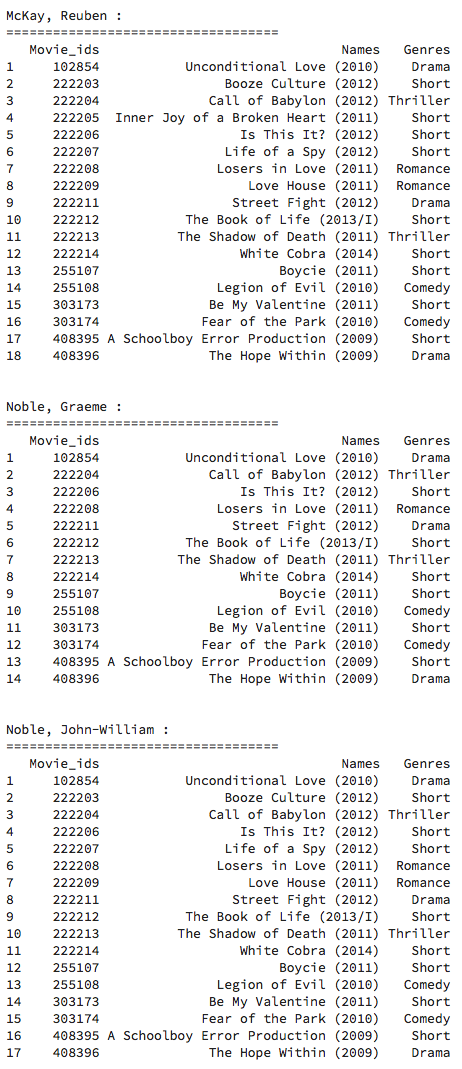


**Figure 7:** Actor – Movie Bipartite Plot of Community 27

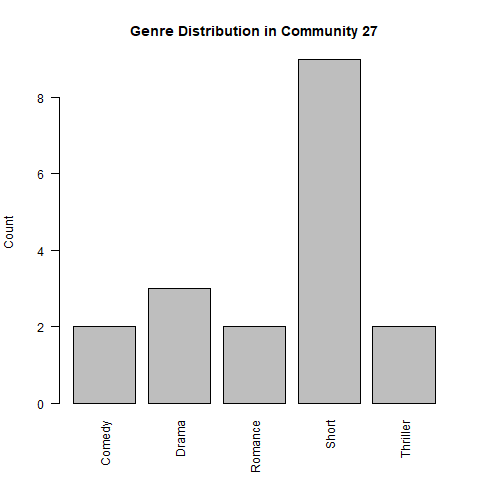
The actors and movies in this community:



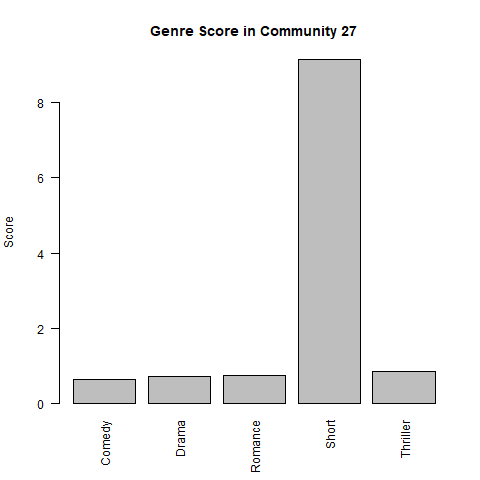
The top three actors information in this community:



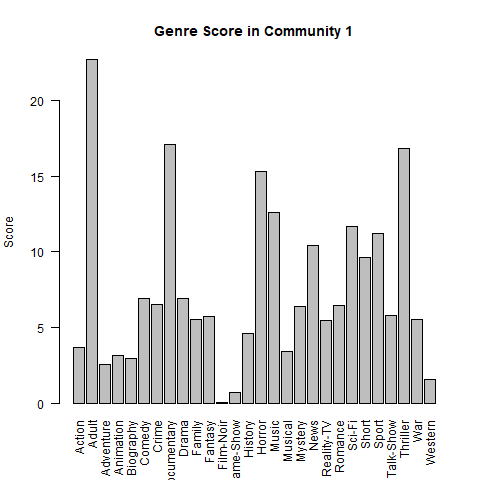
The genre distribution in this community:



The score of each genre in this community:



From plots above, we can tell that the dominant genre based on frequency count and score is same for both cases. Also, the dominant genre in both cases are highly related to the genre of movie which these top three actors acts in. Take community 23 as an example, top three actors majorly act in thriller movies. Thus, the dominant genres are both thriller based on frequency count and score. It is because that an actor tends to act in one specific genre of movie more often than in other genres, and the actor helps movies in same genre gathering together to form a community where has a “pure” dominant genre. For example, in community 1:



The adult genre has a very high score in this community, but has very low scorea in other communities, and the reason is that the most important actors in community 1 are adult movie actors, and they creat edge between movies and gather the adult movies together to form the community 1.

**Q12:**

trained a regression model to predict the rating of three movies. When constructing the model, we **randomly picked 5 actors** who get involved into corresponding movies. When rating a movie people usually tend to evaluate the actors’ performance in the movie, so we use it as our features. The reason why we randomly choose the 5 actors is that we want our model to be robust and try to ignore the outlier problems.

For the model itself, we utilized SVR algorithm to fit our model. The training set is derived from the whole movie (90% training, 10% testing) with rating (if the movie given has no rating in the rating file, we discard it).



We used the root mean square error (RMSE) to evaluate the results.



According to the result, the RMSE given by our model is **0.77(testing); 1.13(validation)**. The predictions for those three movies (Batman v Superman: Dawn of Justice (2016), Mission: Impossible - Rogue Nation (2015), Minions (2015)) are shown in the following table:

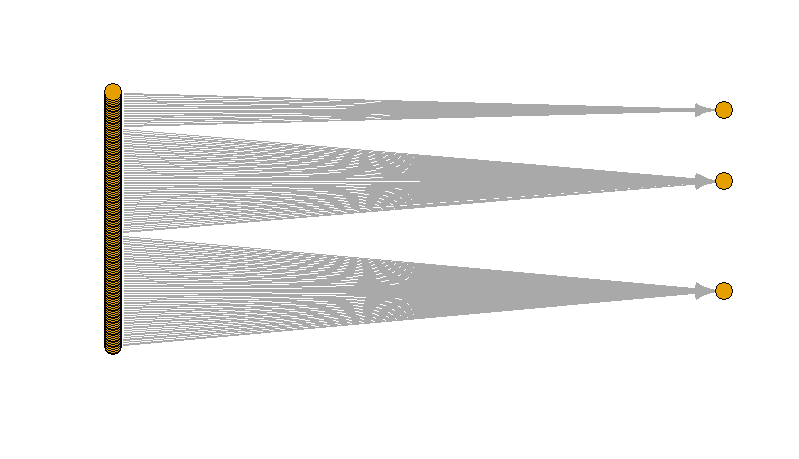
|  |  |
| --- | --- |
| **Movie** | **Rating** |
| Batman v Superman: Dawn of Justice (2016) | 6.12 |
| Mission: Impossible - Rogue Nation (2015) | 6.18 |
| Minions (2015) | 6.15 |

We also tried using top 5 pagerank actors instead of randomly picking to build the same model and we got a better result: **0.69 (testing); 1.08 (validation)**.

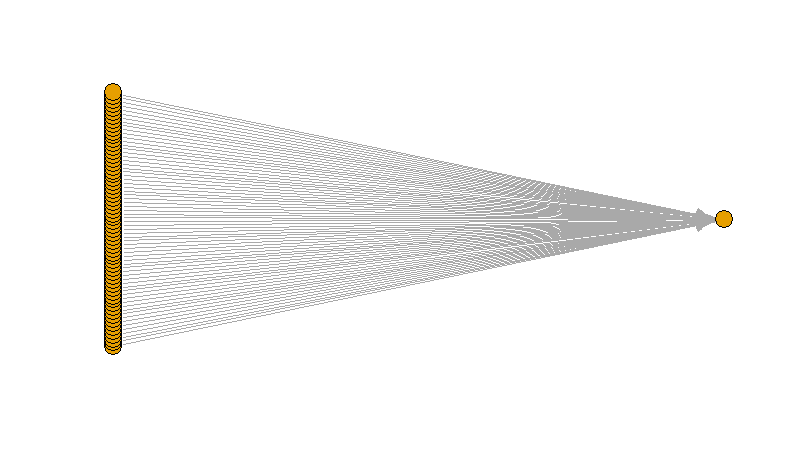
**Q13:**

For this task, we used bipartite graph to predict the rating of each movie. The metric we utilized in this model is mean. To be specific, we investigated one actor’s all relative movies with ratings and calculated the mean of those ratings to get the rating of the corresponding actor. Intuitively, this kind of definition makes sense due to how people tend to rate a movie. People usually tend to rate a movie with high a score if there are some high-quality actors in it.

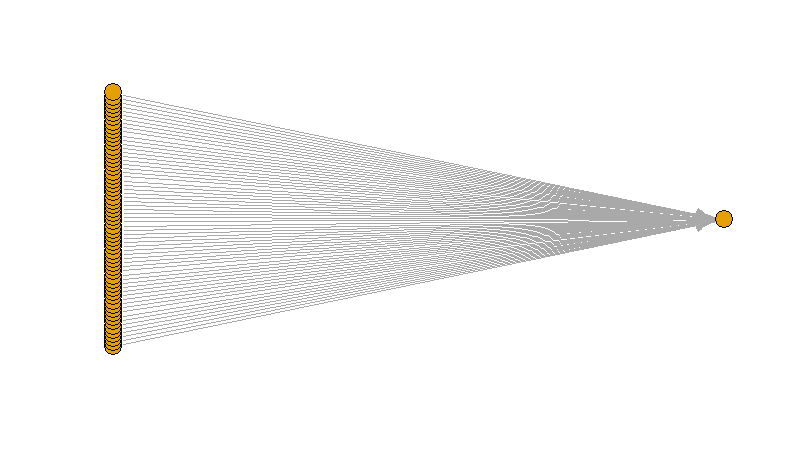
The bipartite graph shown as follows.



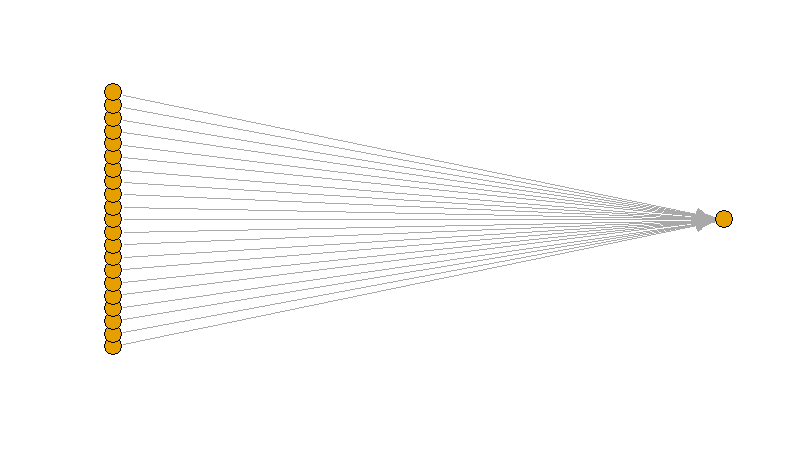
**Figure 3:** All: (actors are on the left while movies are on the right)



**Figure 4:** Batman v Superman: Dawn of Justice (2016)



**Figure 5:** Mission: Impossible - Rogue Nation (2015)



**Figure 6:** Minions (2015)

The RMSE given by our model using bipartite graph is **0.58**. The predictions are:

|  |  |
| --- | --- |
| **Movie** | **Rating** |
| Batman v Superman: Dawn of Justice (2016) | 6.43 |
| Mission: Impossible - Rogue Nation (2015) | 6.54 |
| Minions (2015) | 6.90 |

Clearly, we can see, the bipartite model performs better than the model in problem 12. The reason of that is because we consider more actors in this model for a single movie while the other model only takes 5 actors into consideration when assigning ratings. (This is not always the truth and we need to analyze the parameters to get the results for each situation)