**ECE232E - Project 4**

**IMDb Mining**

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

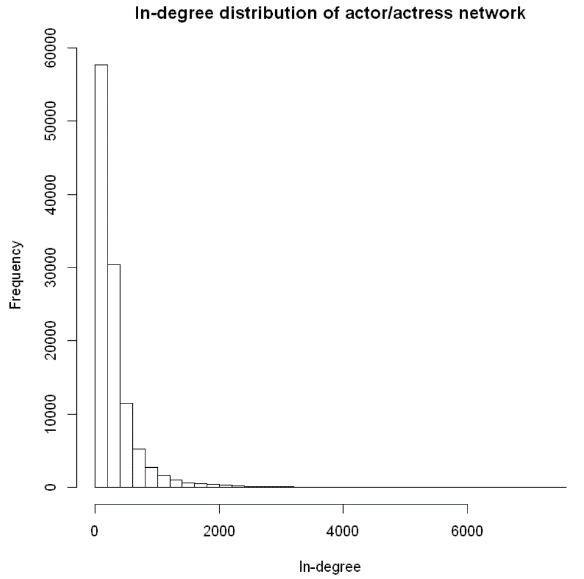
*June 2, 2018*

# **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.

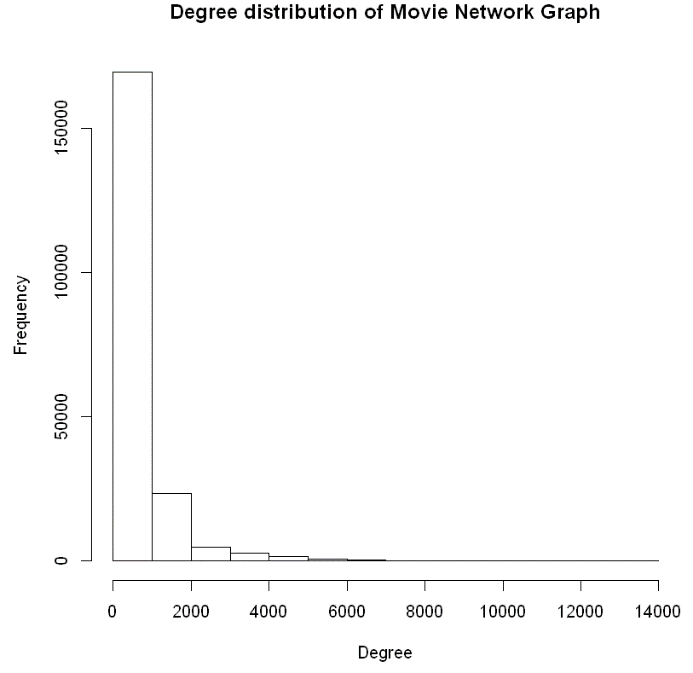
**Q5:**

**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:**

**Q12:**

For this task, we 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:

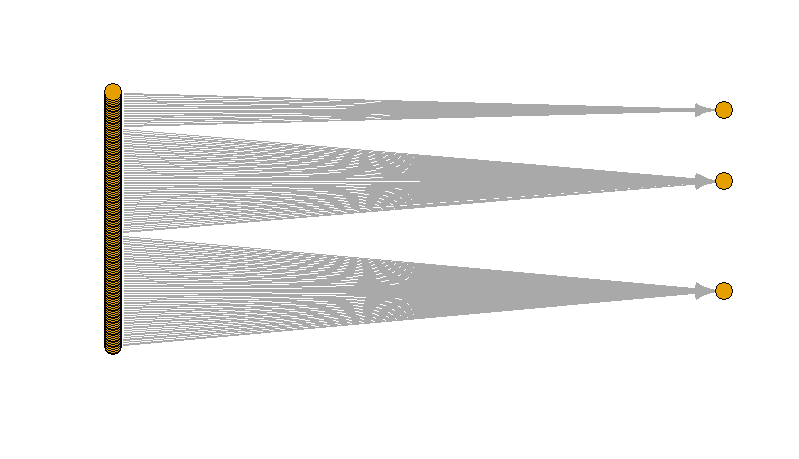
|  |  |
| --- | --- |
| **Movie** | **Rating** |
| Batman v Superman: Dawn of Justice (2016) | 6.12 |
| Mission: Impossible - Rogue Nation (2015) | 6.18 |
| Minions (2015) | 6.15 |

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)**.

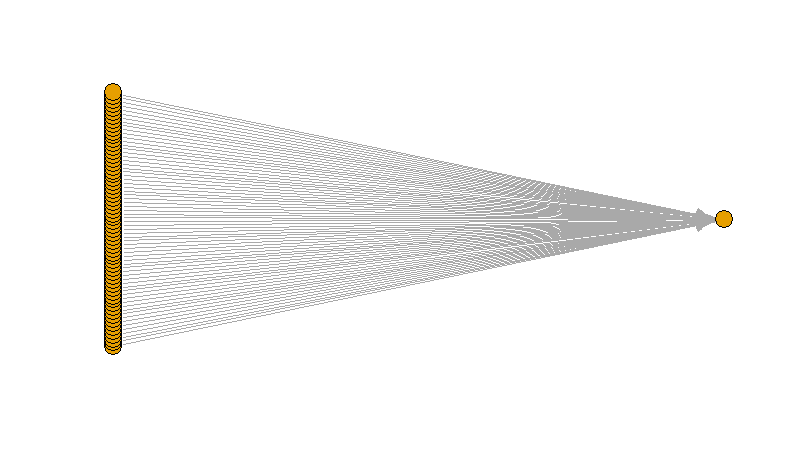
**Q13:**

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.

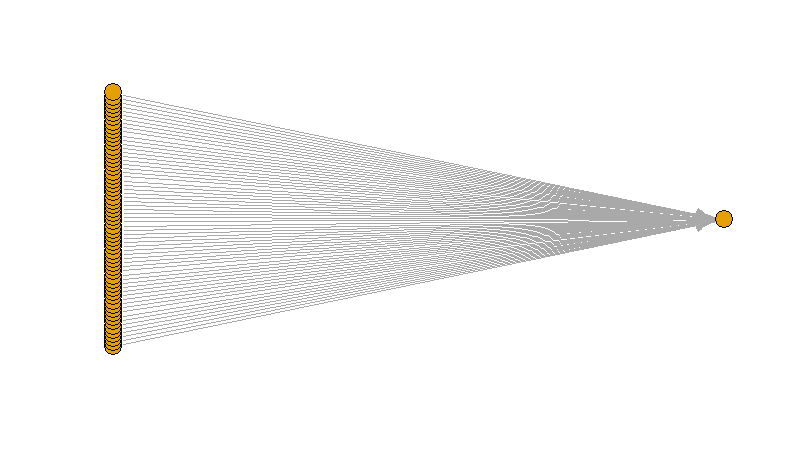
The bipartite graph shown as follows.



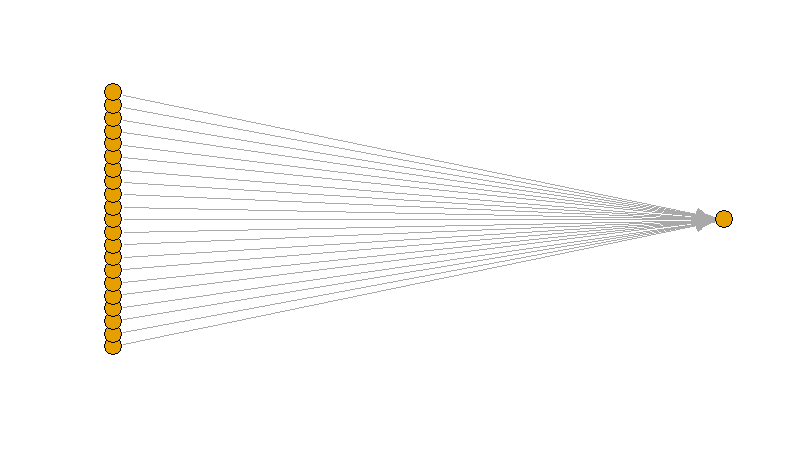
**Figure 3:** All: (actors are on the left while movies are on the right)



**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)