Winter 2019 IMDb Movie Data Project Report

Shan Zhong, University of California, Los Angeles
Jiarui Jiang, University of California, Los Angeles
Mengtong Pan, University of California, Los Angeles

Abstract

Film industries are by nature competitive in terms of limited opportunities and large amount of actors and actresses. However, since the nature of opportunity in front of camera has dramatically changed over the past decades, it is necessary to explore the criterion of becoming an influential actor or actress. In regards of measuring the influence of a person, his or her social network is the primary factor to investigate. In our study, pagerank algorithm and networkx graphs were used to represent the importance of an actor or actress.

Introduction

The film industry has been increasingly attracting attention from data analysts as more and more online streaming services are booming, and with the development of more advanced analysis tools we now have the ability to extract millions of data and analyze them. In order to obtain a better understanding of the modern film industry and how the people working in the film industry interact, we gathered data from IMDb, the world's most popular and authoritative source for movie, TV and celebrity content. The datasets that were used in this analysis contained various information such as the title of a movie, the names of the contributors, the average rating , and the year of the release.

Method

1. Data Preprocessing

The data provided on IMDb has eight separated gz files: title.akas.tsv.gz, title.basics.tsv.gz, title.crew.tsv.gz, title.episode.tsv.gz, title.principals.tsv.gz, title.ratings.tsv.gz and name.basics.tsv.gz. Thereinto, title.crew.tsv.gz, title.episode.tsv.gz and title.principals.tsv.gz were excluded since the information about directors, writers, episode and cast/crew for titles are irrelevant to the measurement of actor's influence. The rest of all title-related datasets were merged into one named title.csv which contains 23 variables and 16.87 millions data entries. To reduce the noise of the dataset, columns that contains more than 70% of missing values were dropped, which are: endYear_x, endYear_y, attributes, language and job. The number of variable in use are the rest of 18 variables. *Figure 1* in the appendix describes all missing values in the entire dataset.

2.Feature selection

To ease the process of data analysis, features with high relevance were selected and used as filters to further reduce noises.

After columns that include large amount of inconsequential missing values were dropped, the amount of repetitions of title identifiers were calculated to evaluate the popularity of each film works. The number of appearances for each title varies from 1 to 1210. Since the popularity of film works to some extent reflects the popularity of their associated actors/actresses, to select the most influential actors/actresses, titles that repeated more than 100 times were selected from title.csv for next step analysis. Inside title.csv, titles whose character names are unknown were filtered; titles typed 'movie', associated people whose primary occupation is actor/actress and average ratings higher than 7 were selected to generate a list of popular movies.

For name.basics.tsv file, similar process was taken to to generate a list of important actors/actress named 'names_select.csv': names with actor/actress as primary profession that appear more than 700 times were included.

3. Application of Networkx & Pagerank

Graph of networkx was used to present the connections between the selected important actors/actresses. Movie titles were set as edges to connect different actors/actresses that participated in the same movie. To produce lists of movies and actors that are compatible for networkx graphing algorithms, in names_select.csv file, birth year, deathyear and primary profession variables were dropped; the column that contains 4 titles in each row were separated into 4 columns: one title per row; then reshape the data frame into 3 columns: edges, source nodes and target nodes. Based on the number of nodes included, 'Football' graph was adopted. Pseudocode for Football networkx graph are listed below:

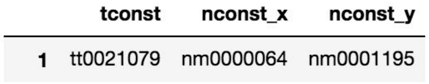
```
 import networkx as nx
 import matplotlib.pyplot as plt
 g = nx.from_pandas_edgelist(df, source = x, target
 = y)

 for n, d in g.degree():
         print('%s %d' % (n, d))
options = {
         'node_color': 'color',
         'node_size': num,
         'line_color': 'color',
         'linewidths': num ,
         'width': num,
 }
 nx.draw(g, **options)
 plt.savefig("nx.png", format="PNG")
 plt.show()
```

After generating the networkx plot, pagerank method was used to represent the level of importance of actors/actresses.

4. Data Visualization

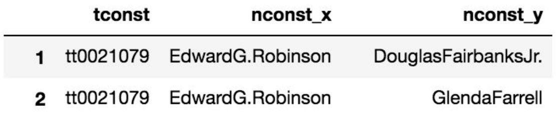
4.1 Actor Networks

The main focuses are on two visualizations, one is to show the relation among actors and actresses (whether they participated in the same movie), the other is to show the pagerank result, based on network of relations. Firstly, we start with the *nx_merge* dataframe, with the format of *tconst, nconst_x, nconst_y*,



	tconst	nconst_x	nconst_y
1	tt0021079	nm0000064	nm0001195

showing which two stars participated in the same movie. Then columns from *name_basic* are used to create a dictionary matching *nconst* to the name of actor, and replace all *nconst* by real names.



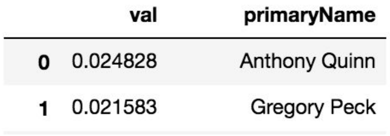
	tconst	nconst_x	nconst_y
1	tt0021079	EdwardG.Robinson	DouglasFairbanksJr.
2	tt0021079	EdwardG.Robinson	GlendaFarrell

Subsequently, we used the data generated by the above data frame to plot the network plots.

Using *nx.from_pandas_edgelist*, we drew each name as a node, and connected the nodes for which names appear at the same row. It is preferable to use a module called *bokeh* to draw the graph, which is capable of being scaling up and down by user to get more detailed views of each network, without being constrained by pixel or resolution limitations.
The resulting network has seven(7) groups, each disconnected from the rest. So the seven graphs are posted separately:
See graphs in **Appendix_Networks**

4.2 Actor Ranking

Based on the network just created, we used standard pagerank function to get ranking value for all actors in our network. The resulting dictionary was converted into pandas dataframe, called:
For_bar_panda1



	val	primaryName
0	0.024828	Anthony Quinn
1	0.021583	Gregory Peck

With *val* as the pagerank value, and *primaryName* as the actor name.

After plotting bar chart of 225 actors, we found that the graph *Bar_chart_1* could not present meaningful information with such a large dataset. The solution was to filter out 19 actors/actresses whose pagerank value is greater than 0.02, and drew another bar chart *Bar_chart_2*. This new graph is visually pleasing and capable of showing meaningful information of pageraking result.
See **Appendix_BarChart**

Results

Results on data

By looking at both Bar charts and Network graphs, we confirm that the center of each group of Network (except some centers of small networks, like Brad Pitt) appears in the Bar chart for highest pagerank actors. Some of the examples are Clint Eastwood, Toshiro Mifune, Michael Caine, Jean Gabin, Robert De Niro, Henry Fonda, Spencer Tracy, etc.

Another intuitive result is that actor/actress in a larger connected network are more likely to have higher pagerank, even if they are not in obvious center position. Viewing in the context of real-world filming industry, it is reasonable to think that's exactly how social networks among actor/actress work. Film stars participated in various films enjoy not only his/her own network, but also the relation and opportunities brought by partners. It may be tempting to assume that people in the center of a network would have larger pagerank scores, but in fact based on our analysis, people who are not in the obviously significant position in a big network actually have higher pagerank scores than those located at the center of smaller networks.

Improvement

The excess of data is definitely an obstacle for processing and analyzing this project, and so the first step of data cleaning excluded all individuals whose primary profession is not actor/actress. However, there might exist some popular films and outstanding stars who have other professions related to the film industry, but are not actors/actresses. In order to handle the dataset and facilitate further exploration, this way of filtering actor/actress is a necessary compromise, but could be improved, if possible.

Also, since the data after cleaning and processing still contain 225 individuals, ordinary plotting method taught in class can no longer generate visually pleasing and meaningful graphs. Thus a new module with capability of scaling up and down as specified by user is introduced in order to plot networks with enough visible details. One feature of our original plan was to display nodes with different sizes, with more influential individual showing larger nodes. However, due to unfamiliarity with new method, our best result is to ensure the clarity of networks, with relations and names of each individual being presented in a readily recognizable way. Had the dataset be more manageable, or our understanding of module functions more comprehensible, it could be possible to reach all expectation in the ideal plan.

Discussion

Due to the large population size, a random sample of 30,000 data were extracted from the cleaned title dataframe for the analysis purpose instead of analyzing all the existing data. After removing missing values, selecting for the title type to be movies only, and selecting for the movie data that included actors, actresses and directors,we had 9598 observations for further analysis. A scatterplot of movies (Appendix A) by start year and genres showed a clear trend that more movies were produced after 2000. Since many movies have more than one category, we simplified the problem by using the first name in the string as its main category. Genres such as Music, Sci, and Horor showed low average ratings (mean = 4.45,5.52,5.25, respectively) whereas genres such as Biography, Documentary, and Crime showed high average ratings (mean = 6.93,7.09,6.42, respectively). Drama was the largest category out of a total of 23 categories, and 29.2% of the movies belonged to this category. The second largest category was Comedy, which consisted of 24.5% of the movies. The third largest category was Action, and it consisted of 15.0% of the movies.

Another scatterplot comparing runtime and average ratings showed that there was no strong association between runtime and average ratings (Appendix B), but it could be concluded from the plot that the average of the runtime equals to 98.41 minutes. It may suggest that the ratings of a movie does not depend on the runtime and there are a large amount of variations in terms of the ratings within a particular runtime.

In order to gain more insights into how the popularity of a movie was determined by other factors, we created another scatterplot to analyse the number of votes and the average ratings. We consider movies over 500,000 votes to be popular, and based on this criteria, it seems that among all the movies that are popular, the more votes one received, the higher the average ratings became. From the plot we could infer that most of the highly rated movies clustered within the year range 2000 to 2018, and more movies in general were produced after year 2000 (Appendix C). This association can be viewed more clearly in (Appendix D). A simple linear regression showed that for movies that had more than 500,000 votes, the predicted average rating equals to 1.076e-06*number of votes + 7.245, ($p < 2e-16$, r-squared = 0.3072).

Next, we explored the relationship among the popularity of a movie, the rating of a movie, and its director. We first filtered 202,212 directors out of the whole dataset, and selected those with movies that had more than 500,000 votes. This gave us 7,526 directors with information on the title, genre, start year, average rating of the movie, etc. We then filtered movies based on the average ratings and selected the top ten rated movies (Appendix E). Since we were interested in identifying top directors, we used the directors who had directed the top rated movies and searched for all the popular movies that were directed by them. From there we developed a chart of top directors and their movies. (Appendix F) To gain a better understanding of audience feedback, we subsequently calculated the mean ratings of each top directors' movies combined and arranged them in descending order. (Appendix G)

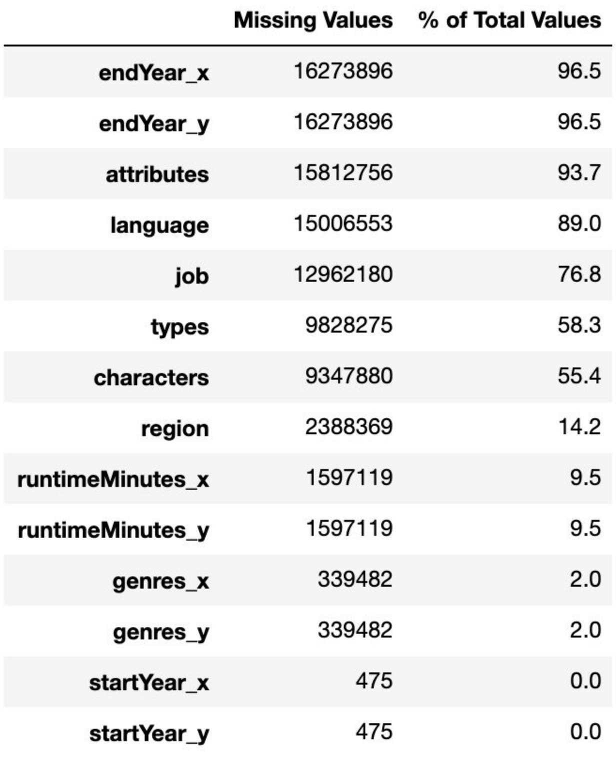
As a conclusion, we could see that several factors affect whether a movie would receive high popularity and ratings. First of all, audience in general prefer genres such as Biography, Crime, and Documentary, but the top rated movies are typically Drama or Crime, indicating that mediocre Drama movies may not have high ratings, but the ones that are outstanding do receive high ratings. Runtime does not influence how a movie is rated, but most of the movies last approximately 100 minutes. It is interesting to see that more votes predict higher ratings, but this pattern only exists among movies that already have a large amount of votes. Therefore it can be concluded that these two factors are influenced by each other. For example, people may be more likely to watch a movie that has a good rating, and they may also be influenced by the high rating to vote and give a high rating themselves. Several directors such as Quentin Tarantino, Steven Spielberg, and Christopher Nolan manifest a consistent popularity and productivity among all the directors, but we can also identify directors such as Frank Darabont who only had one famous movie and did not have other successful ones. There are many factors that could potentially influence a rating of a movie, and the trend that we observe demonstrates a cluster-like relationship: movies that tend to have the best directors and actors are more popular than the others. In other words, they have better resources from the beginning and thus are more likely to succeed in the market.

Reference

Football network algorithm
https://networkx.github.io/documentation/stable/auto_examples/graph/plot_football.html#sphx-glr-auto-examples-graph-plot-football-py

Bokeh module
https://bokeh.pydata.org/en/latest/docs/user_guide.html#userguide

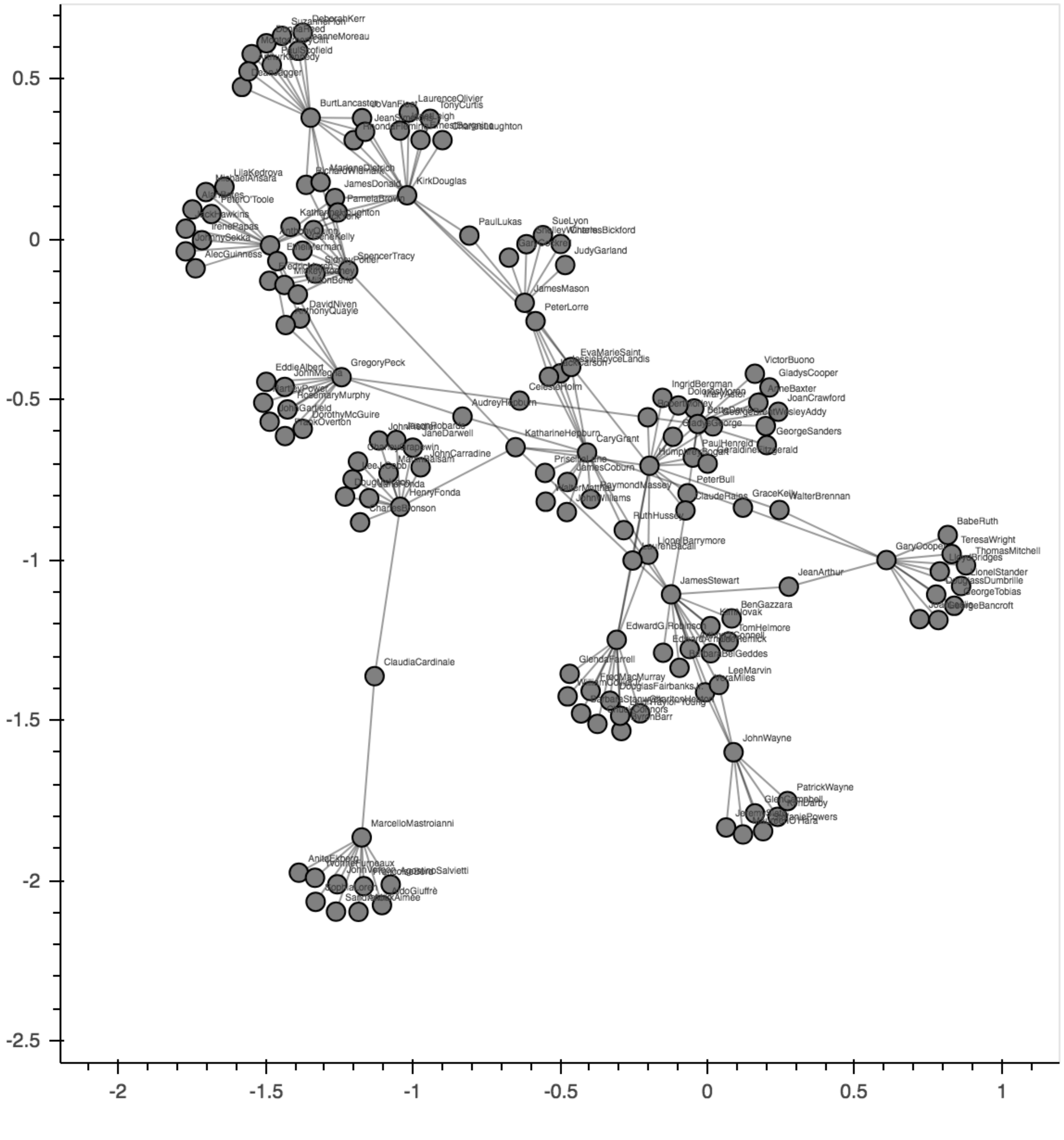
Appendix_Missing_Values

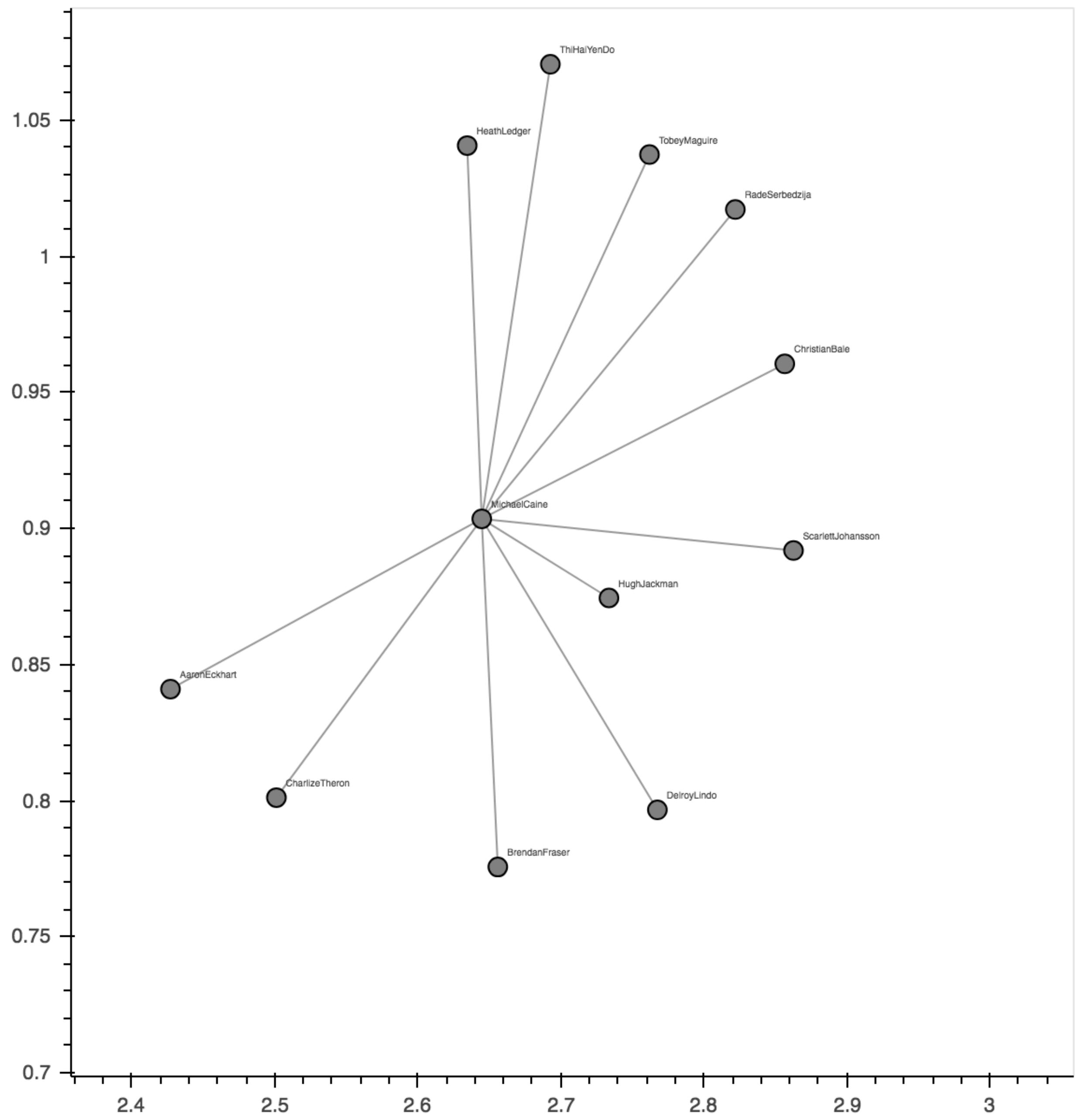


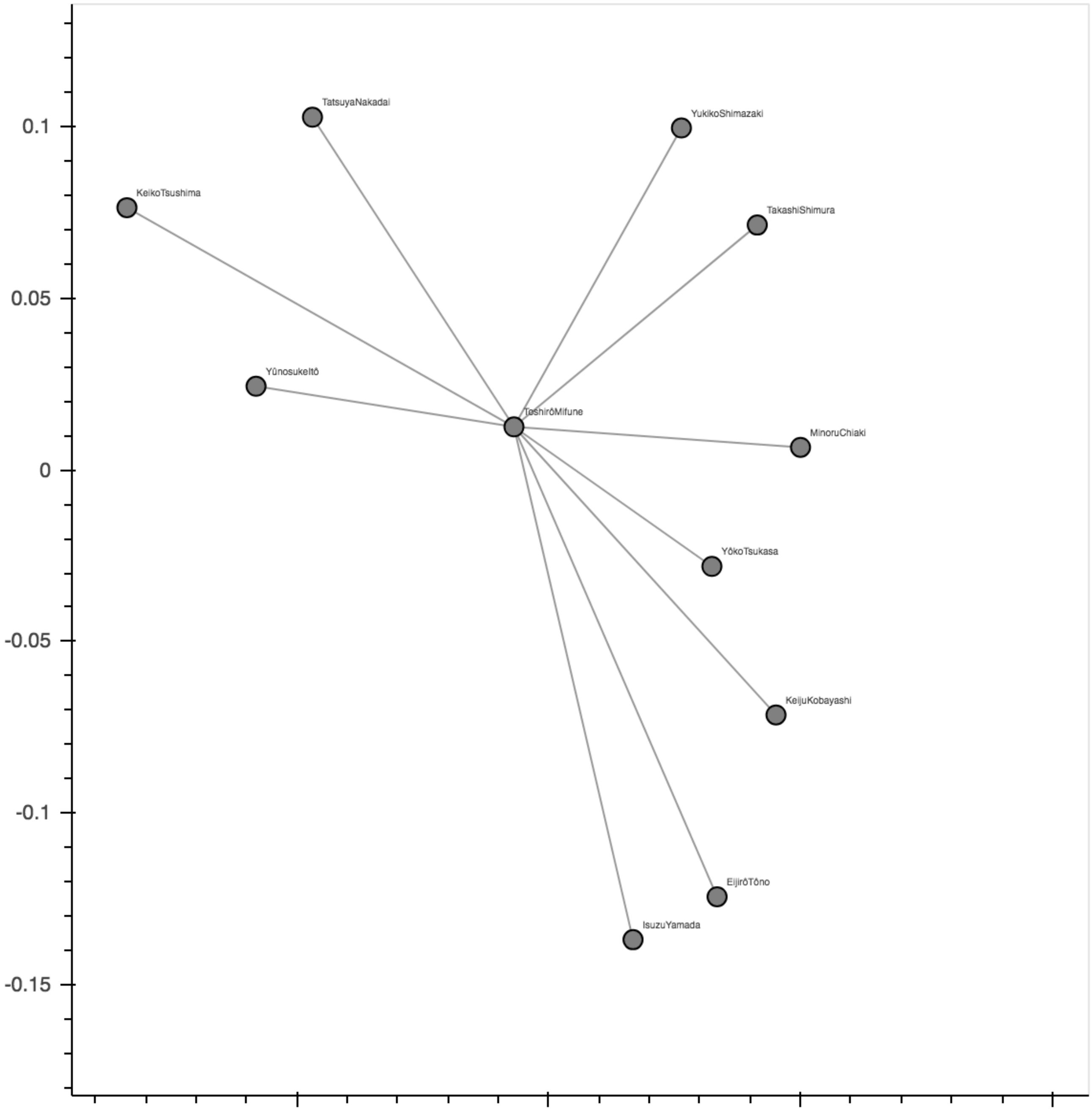
	Missing Values	% of Total Values
endYear_x	16273896	96.5
endYear_y	16273896	96.5
attributes	15812756	93.7
language	15006553	89.0
job	12962180	76.8
types	9828275	58.3
characters	9347880	55.4
region	2388369	14.2
runtimeMinutes_x	1597119	9.5
runtimeMinutes_y	1597119	9.5
genres_x	339482	2.0
genres_y	339482	2.0
startYear_x	475	0.0
startYear_y	475	0.0

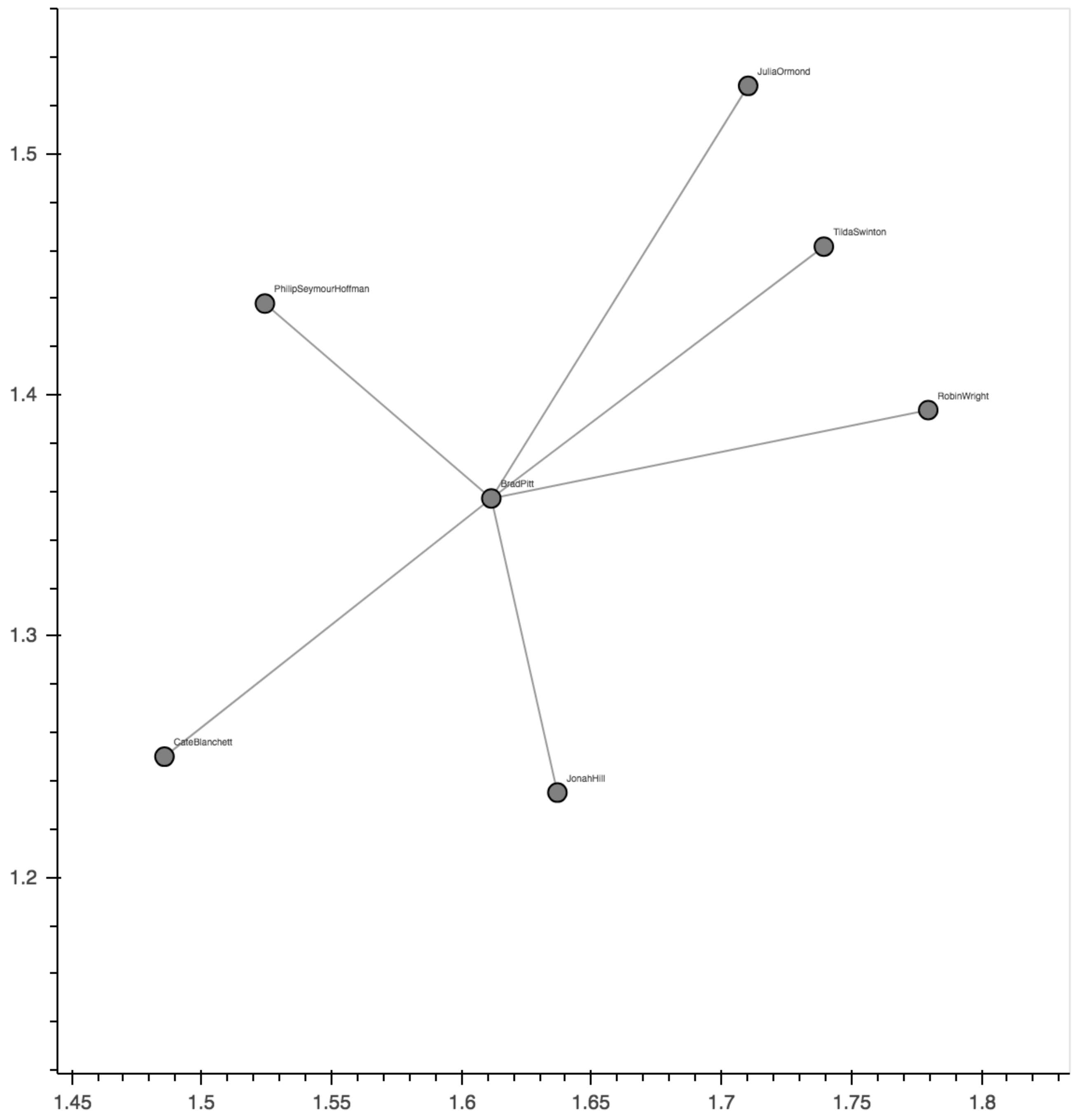
Figure 1

Appendix_Networks_1

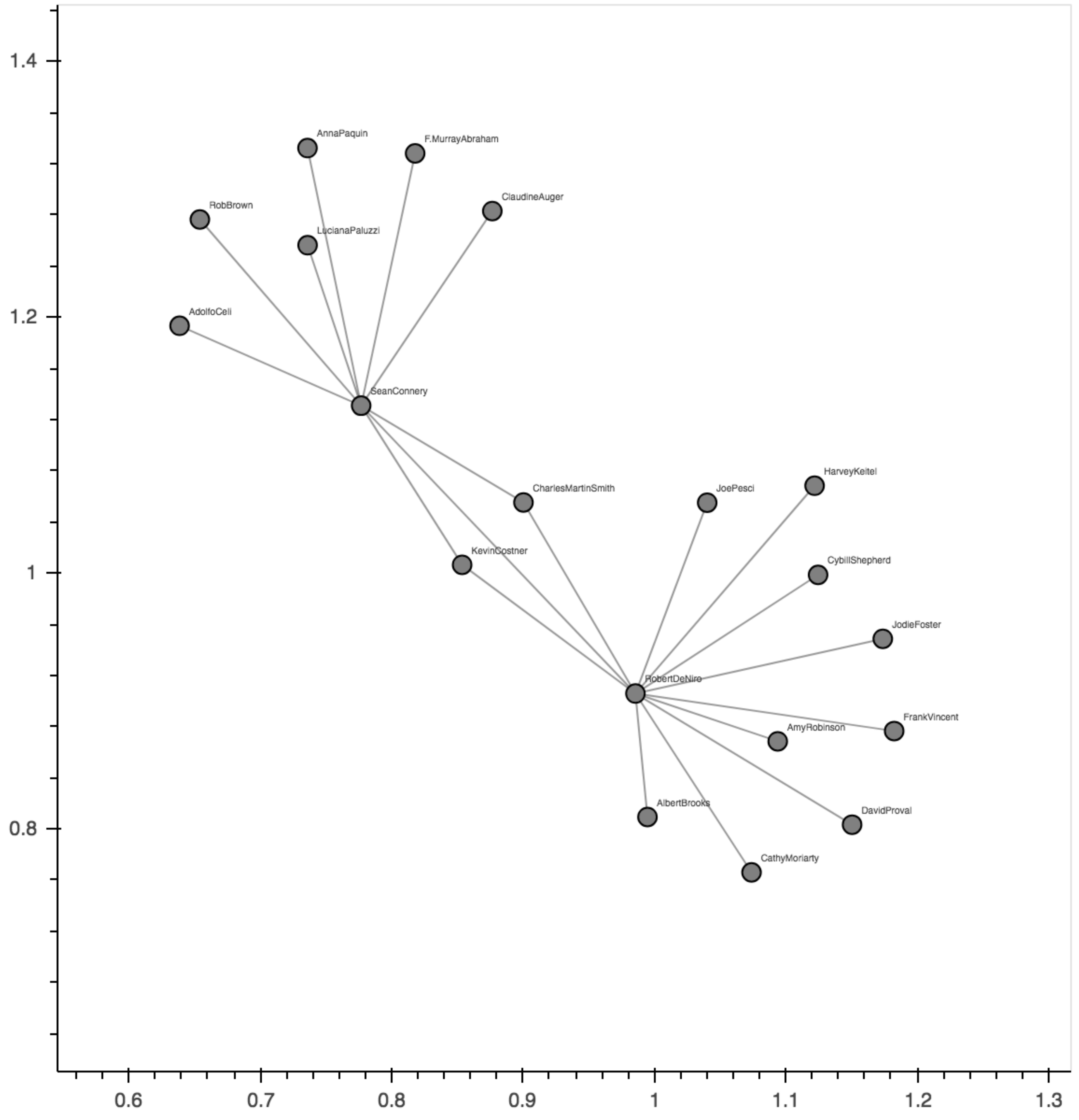


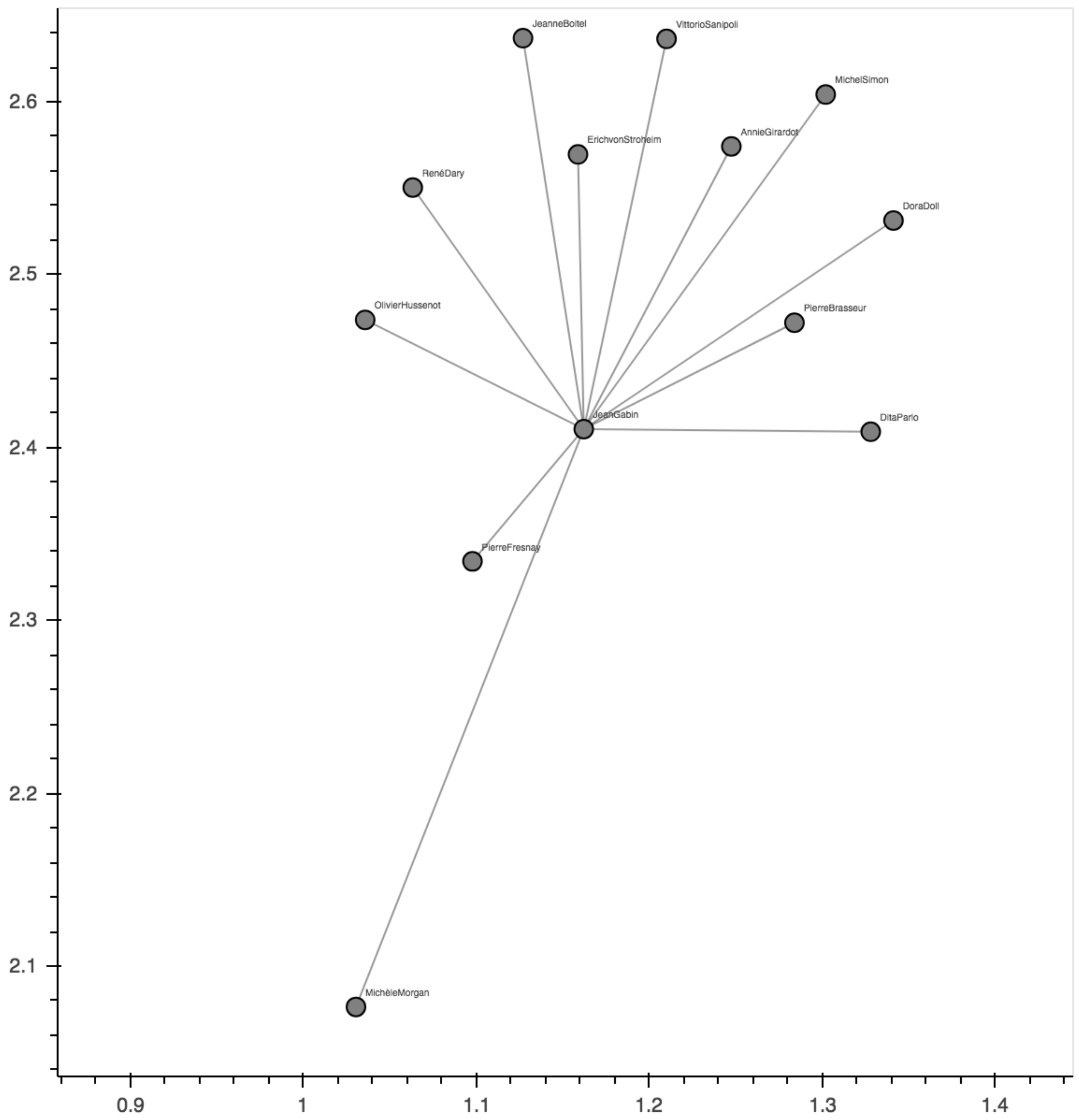

ThiHaiYenDo
HeathLedger
TobeyMaguire
RadeSerbedzija
ChristianBale
MichaelCaine
ScarlettJohansson
HughJackman
AaronEckhart
CharlizeTheron
DelroyLindo
BrendanFraser
1.05
1
0.95
0.9
0.85
0.8
0.75
0.7
2.4
2.5
2.6
2.7
2.8
2.9
3

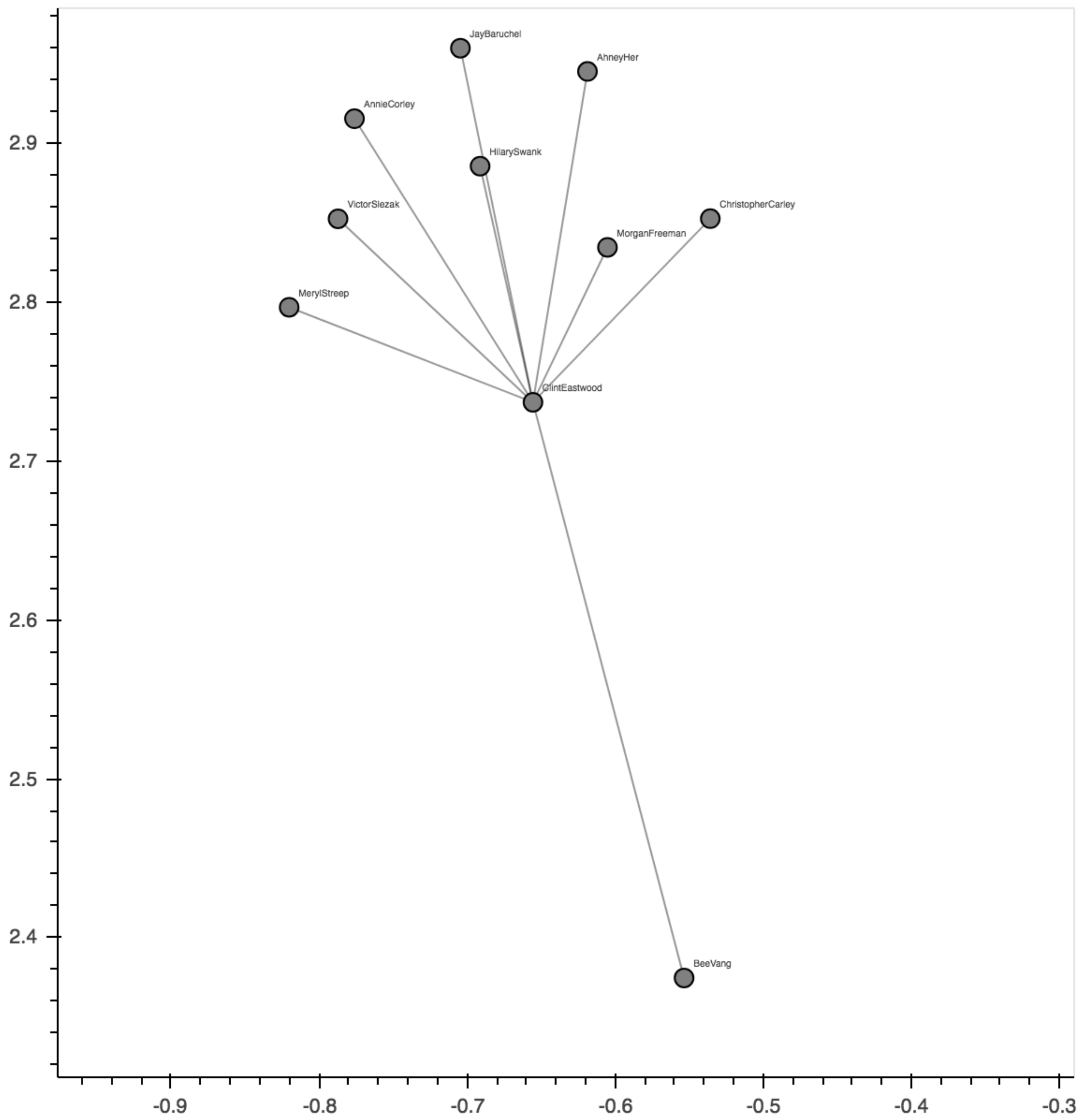

TatsuyaNakadai
YukikoShimazaki
KeikoTsushima
TakashiShimura
YûnosukeItô
ToshirôMifune
MinoruChiaki
YôkoTsukasa
KeijuKobayashi
EijirôTôno
IsuzuYamada
0.1
0.05
0
-0.05
-0.1
-0.15
2.6
2.8
3
3.2


JuliaOrmond
TildaSwinton
PhilipSeymourHoffman
RobinWright
BradPitt
CateBlanchett
JonahHill
1.5
1.4
1.3
1.2
1.45
1.5
1.55
1.6
1.65
1.7
1.75
1.8

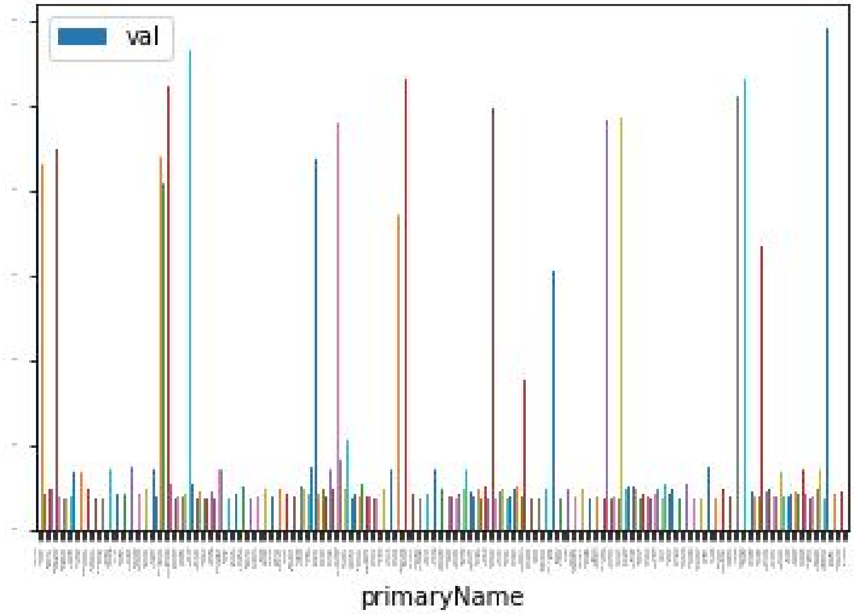
Appendix_Networks_2



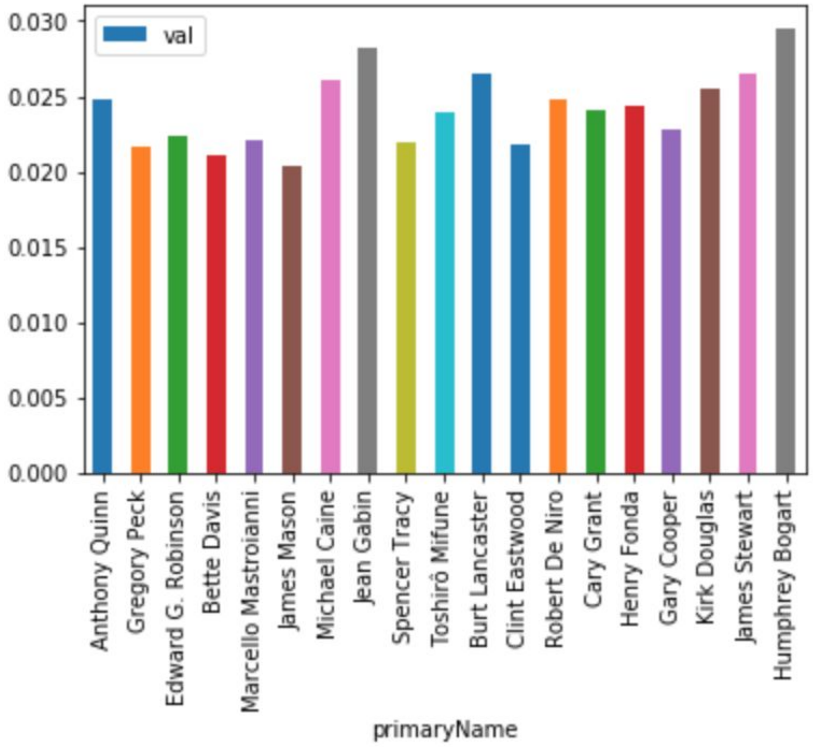

JeanneBoitel
VittorioSanipoli
MichelSimon
ErichvonStroheim
AnnieGirardot
RenéDary
DoraDoll
OlivierHussenot
PierreBrasseur
JeanGabin
DitaParlo
PierreFresnay
MichèleMorgan
2.6
2.5
2.4
2.3
2.2
2.1
0.9
1
1.1
1.2
1.3
1.4


JayBaruchel
AhneyHer
AnnieCorley
HilarySwank
VictorSlezak
ChristopherCarley
MorganFreeman
MerylStreep
ClintEastwood
BeeVang
2.9
2.8
2.7
2.6
2.5
2.4
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3

Appendix_BarChart

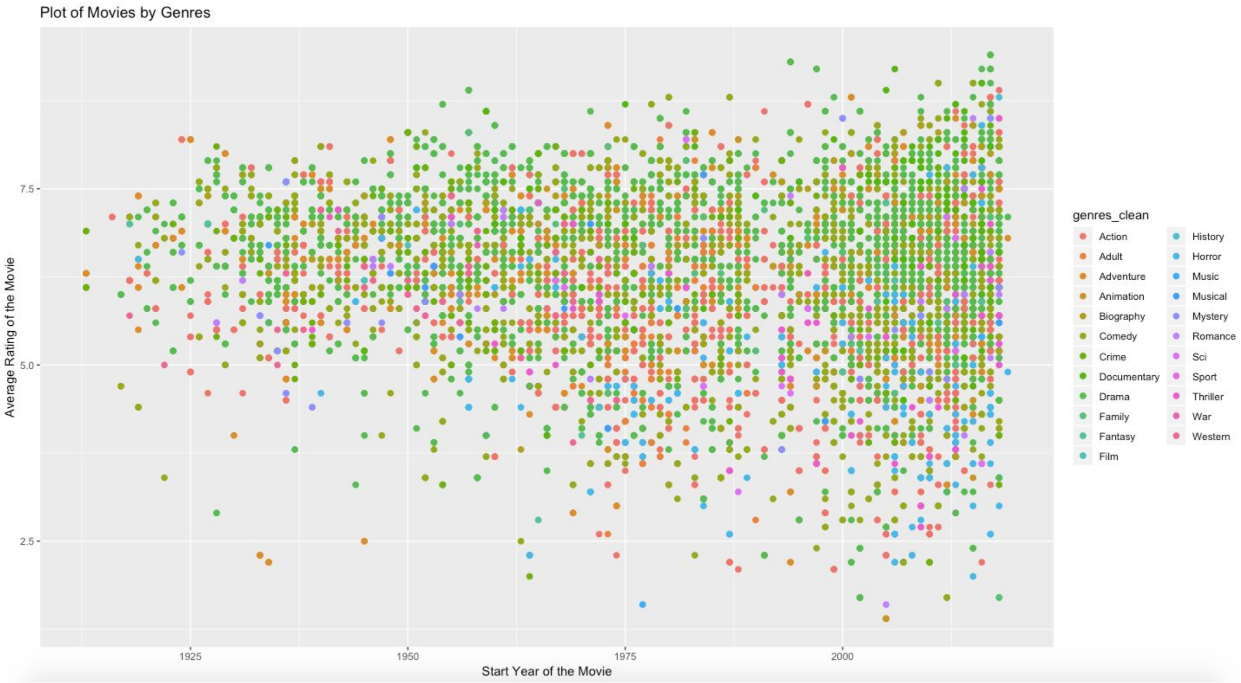


Bar_chart_1



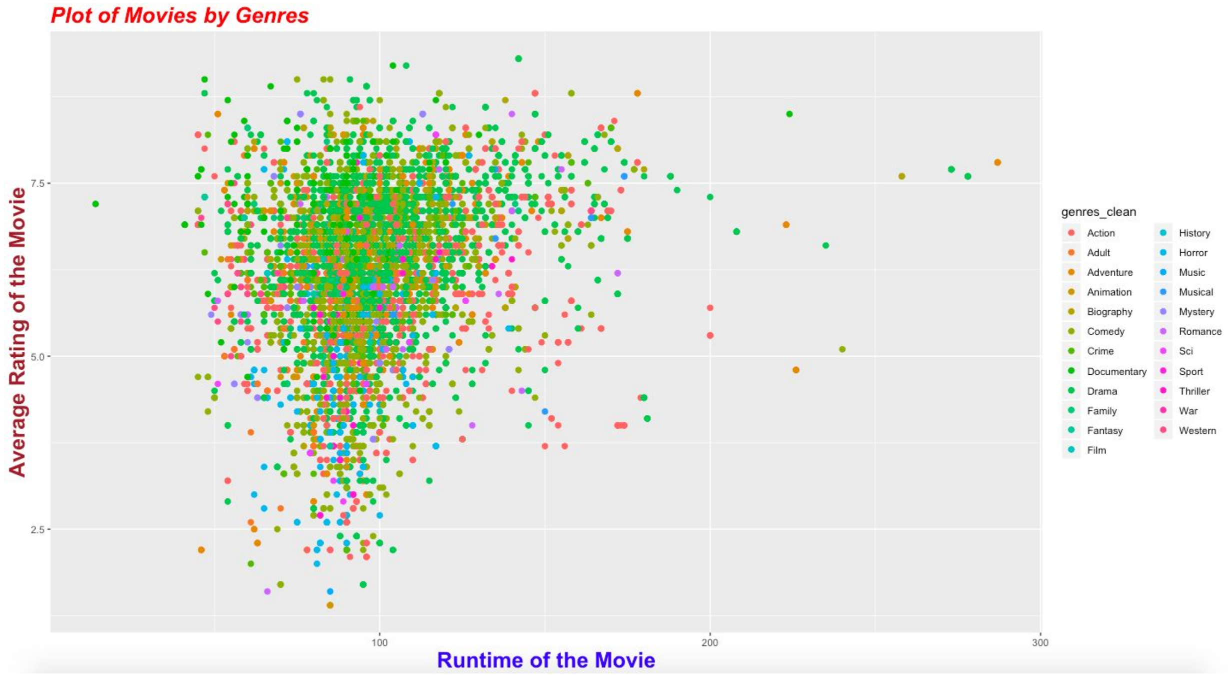
Bar_chart_2

Appendix A
Relationship of the Average Rating and the Start Year of a Movie



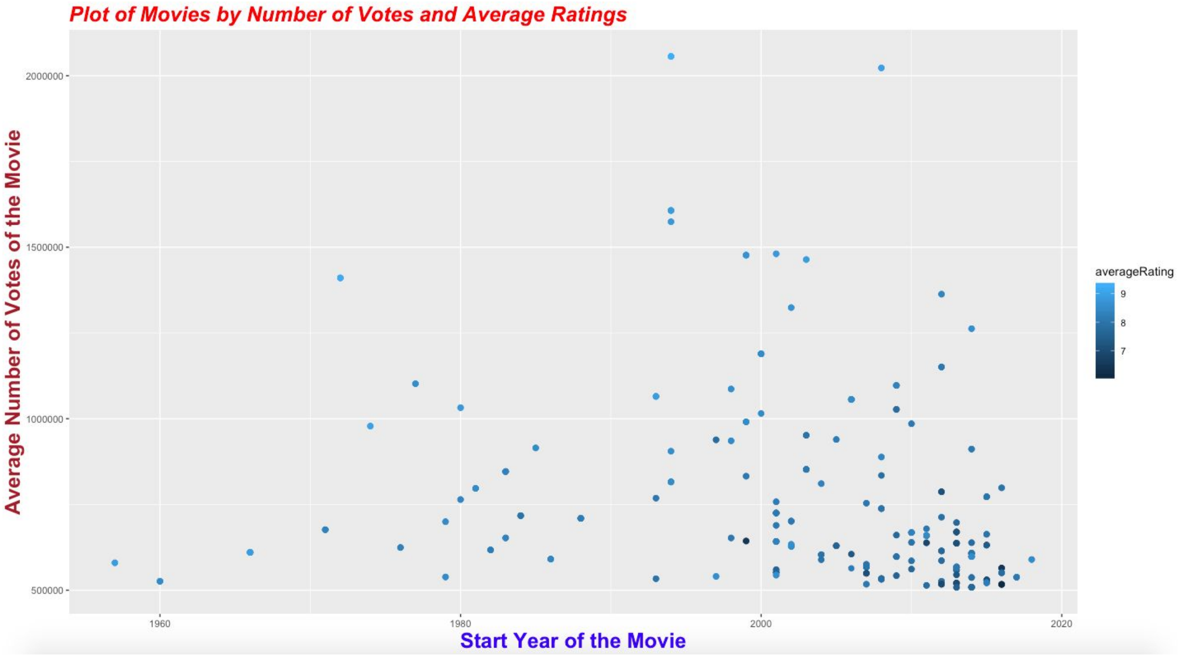
Appendix B

Relationship of the Runtime of the Movie and the Average Rating of the Movie



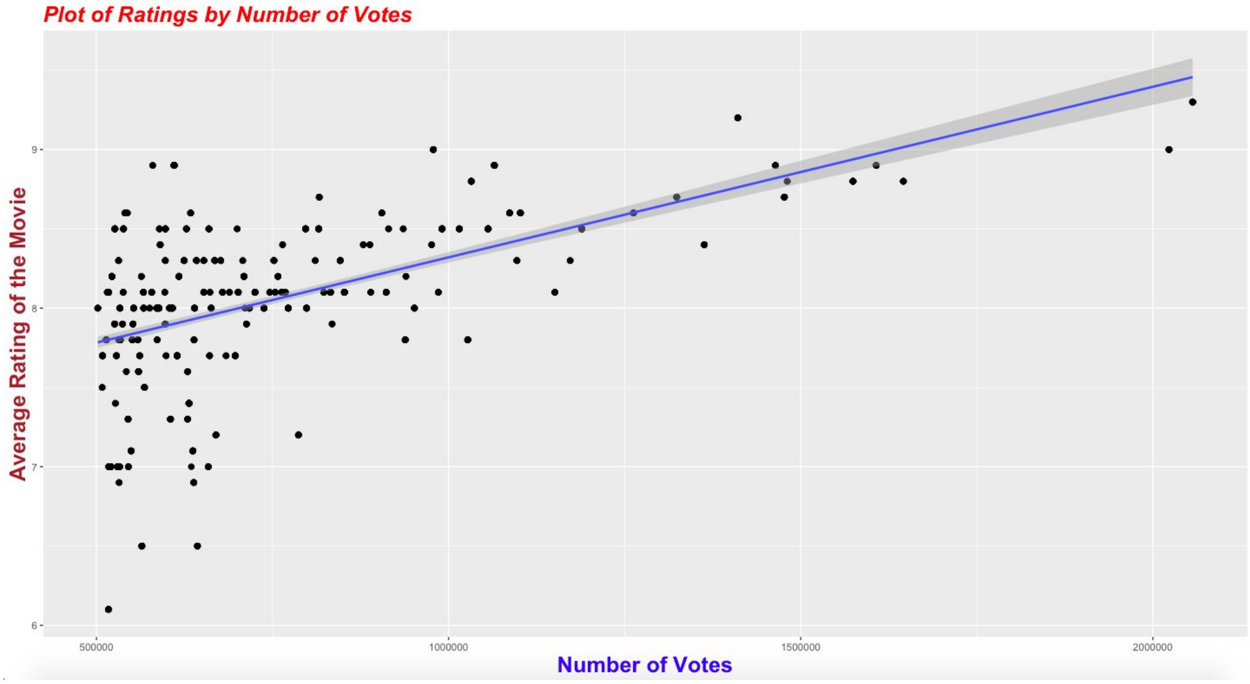
Appendix C

Relationship of the Start Year of the Movie and Average rating of the Movie



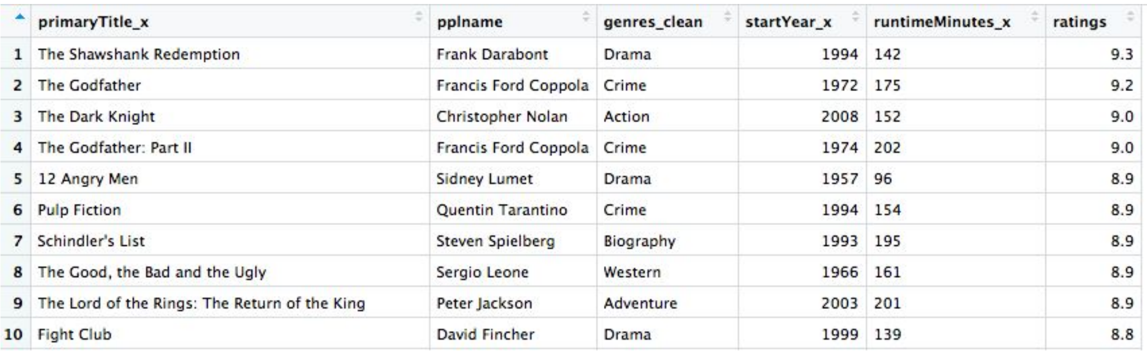
Appendix D

Relationship of the Number of Votes and the Average Rating of the Movie



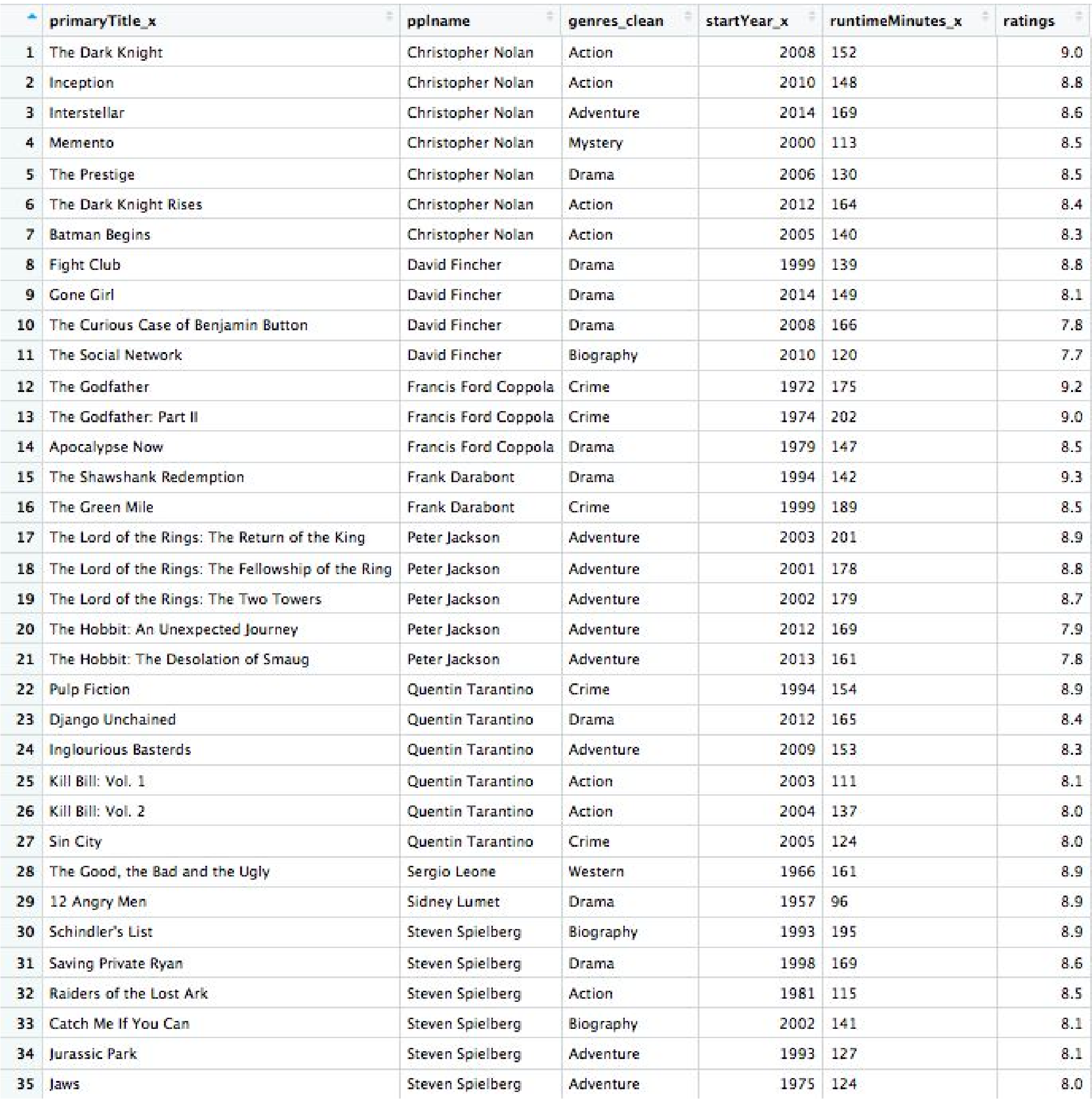
Appendix E

The Top Ten Rated Movies



▲	primaryTitle_x	pplname	genres_clean	startYear_x	runtimeMinutes_x	ratings
1	The Shawshank Redemption	Frank Darabont	Drama	1994	142	9.3
2	The Godfather	Francis Ford Coppola	Crime	1972	175	9.2
3	The Dark Knight	Christopher Nolan	Action	2008	152	9.0
4	The Godfather: Part II	Francis Ford Coppola	Crime	1974	202	9.0
5	12 Angry Men	Sidney Lumet	Drama	1957	96	8.9
6	Pulp Fiction	Quentin Tarantino	Crime	1994	154	8.9
7	Schindler's List	Steven Spielberg	Biography	1993	195	8.9
8	The Good, the Bad and the Ugly	Sergio Leone	Western	1966	161	8.9
9	The Lord of the Rings: The Return of the King	Peter Jackson	Adventure	2003	201	8.9
10	Fight Club	David Fincher	Drama	1999	139	8.8

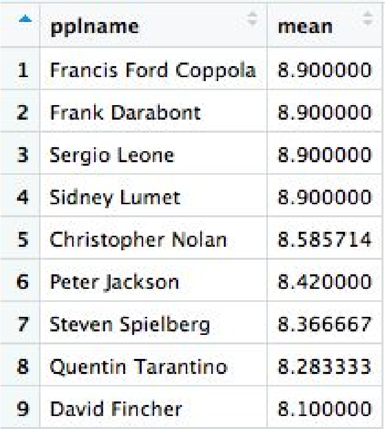
Appendix F
The Top Directors and Their Works



	primaryTitle_x	pplname	genres_clean	startYear_x	runtimeMinutes_x	ratings
1	The Dark Knight	Christopher Nolan	Action	2008	152	9.0
2	Inception	Christopher Nolan	Action	2010	148	8.8
3	Interstellar	Christopher Nolan	Adventure	2014	169	8.6
4	Memento	Christopher Nolan	Mystery	2000	113	8.5
5	The Prestige	Christopher Nolan	Drama	2006	130	8.5
6	The Dark Knight Rises	Christopher Nolan	Action	2012	164	8.4
7	Batman Begins	Christopher Nolan	Action	2005	140	8.3
8	Fight Club	David Fincher	Drama	1999	139	8.8
9	Gone Girl	David Fincher	Drama	2014	149	8.1
10	The Curious Case of Benjamin Button	David Fincher	Drama	2008	166	7.8
11	The Social Network	David Fincher	Biography	2010	120	7.7
12	The Godfather	Francis Ford Coppola	Crime	1972	175	9.2
13	The Godfather: Part II	Francis Ford Coppola	Crime	1974	202	9.0
14	Apocalypse Now	Francis Ford Coppola	Drama	1979	147	8.5
15	The Shawshank Redemption	Frank Darabont	Drama	1994	142	9.3
16	The Green Mile	Frank Darabont	Crime	1999	189	8.5
17	The Lord of the Rings: The Return of the King	Peter Jackson	Adventure	2003	201	8.9
18	The Lord of the Rings: The Fellowship of the Ring	Peter Jackson	Adventure	2001	178	8.8
19	The Lord of the Rings: The Two Towers	Peter Jackson	Adventure	2002	179	8.7
20	The Hobbit: An Unexpected Journey	Peter Jackson	Adventure	2012	169	7.9
21	The Hobbit: The Desolation of Smaug	Peter Jackson	Adventure	2013	161	7.8
22	Pulp Fiction	Quentin Tarantino	Crime	1994	154	8.9
23	Django Unchained	Quentin Tarantino	Drama	2012	165	8.4
24	Inglourious Basterds	Quentin Tarantino	Adventure	2009	153	8.3
25	Kill Bill: Vol. 1	Quentin Tarantino	Action	2003	111	8.1
26	Kill Bill: Vol. 2	Quentin Tarantino	Action	2004	137	8.0
27	Sin City	Quentin Tarantino	Crime	2005	124	8.0
28	The Good, the Bad and the Ugly	Sergio Leone	Western	1966	161	8.9
29	12 Angry Men	Sidney Lumet	Drama	1957	96	8.9
30	Schindler's List	Steven Spielberg	Biography	1993	195	8.9
31	Saving Private Ryan	Steven Spielberg	Drama	1998	169	8.6
32	Raiders of the Lost Ark	Steven Spielberg	Action	1981	115	8.5
33	Catch Me If You Can	Steven Spielberg	Biography	2002	141	8.1
34	Jurassic Park	Steven Spielberg	Adventure	1993	127	8.1
35	Jaws	Steven Spielberg	Adventure	1975	124	8.0

Appendix G

Directors and the Average Ratings for Their Movies Combined



	pplname	mean
1	Francis Ford Coppola	8.900000
2	Frank Darabont	8.900000
3	Sergio Leone	8.900000
4	Sidney Lumet	8.900000
5	Christopher Nolan	8.585714
6	Peter Jackson	8.420000
7	Steven Spielberg	8.366667
8	Quentin Tarantino	8.283333
9	David Fincher	8.100000