Network Science Project Report
Demographic Analysis of the Academy Award Nominations

Sara Kartalovic

sara.kartalovic@studenti.unipd.it

Bogdan Milovanovic

bogdan.milovanovic@studenti.unipd.it

Tarja Savonen

tarjatuulikki.savonen@studenti.unipd.it

February 2022

1 Introduction

The Academy Awards, which are the most recognized trophies in the world, are presented to the greatest filmmakers. Best known as the Oscars, they are given annually by the Academy of Motion Picture Arts and Sciences (AMPAS) for the outstanding achievement in artistic and technical aspects in the film industry. Based on the Academy member's voting results, this awards serve as the international recognition of excellence in cinematic accomplishments. The awards were first presented in 1929, and winners receive a gold-plated statuette commonly called Oscar. Winners are chosen from the following 24 categories, even though categories have been added and discontinued over the years.

Using two different datasets, 'The Oscar Award' and 'Academy Awards Demographics', three networks were created in order to see how movies, actors, and directors are related to each other. In the first network, movies were used as the nodes, while in the other two networks people were used as nodes. All three networks include only 5 categories where the first dataset contains all the nominees from 1927 until 2020, while the second dataset includes only the Academy Award winners from 1927 until 2016. After creating the networks, different properties of the networks were investigated such as degree rage, page rank, different centralities, giant components and clustering coefficients. All data analysis and manipulation was done in Python, while for the network creation and analysis besides Python, we also used Gephi.

2 Datasets (Tarja Savonen)

In this section, two datasets will be described along with data examination. Both datasets provide a large, reliable, and comprehensive collection of data that is meaningful and good for network analysis.

2.1 The Oscar Award Dataset

The Oscar Award dataset consists of records scraped from the Academy Award Database and is freely available on Kaggle website. It contains all nominees and award winners since the first ceremony in 1927 until 92nd ceremony in 2020. Original dataset, before preprocessing, has 10395 entries which include numerous nominees sorted by the year of the ceremony in the ascending order. Also, dataset contains 7 features which are:

- 'year_film' - the year when the movie was released
- 'year_ceremony' - the year when the ceremony was held
- 'ceremony' - the number of the ceremony
- 'category' - the category in which person/movie/organization was nominated
- 'name' - the full name of the nominee
- 'film' - the name of the movie
- 'winner' - holds value 'True' if nominee won an award, and 'False' otherwise

The 'name' column has 6666 unique terms, whereas the column 'winner' indicates that there were 2357 (23%) winners and 8038 (77%) nominees overall. Only the column 'film' has 304 (about 3%) missing values.

Figure 1 shows the distributions of the number of main categories, nominees in different categories, and nominated movies through time. The first plot on the left "Number Of Main Category Change" shows that over time there has been a fluctuation in the number of main categories but with a constant trend of growth. When the academy was founded in 1927 there were presented 12 main categories, while today when winners are chosen from the 24 main categories. Some categories that were additionally considered are "Actor/Actress in supporting role", "Sound effect", "Film editing", etc. The other two distribution plots show the highest number of nominees and nominated films was around the 1940s.

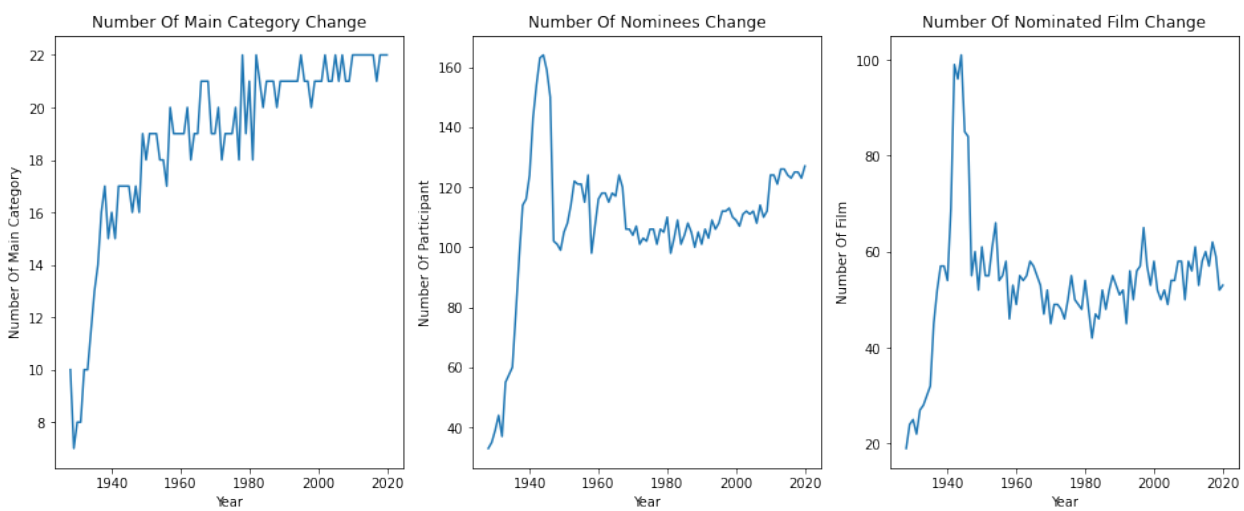


Figure 1: Distribution of the Oscar nominees from the first ceremony to the last one.

Figure 2 shows the probability has been changed through time. The plot on the left side shows the number of wins and only nominated categories, while the right plot represents the probabilities of wins. We can conclude that as the number of nominees grows, they have less chance of winning the award, which leads to conclude that nominating and probability of winning are disproportionate.

Figure 3 shows the distributions of the top 10 nominated and awarded movies. It can be concluded that even though some movies have a lot of nominations (plot on the left), it does not imply that they also got the highest number of awards (plot on the right). As an example the movie "A Star is Born" received 24 major nominations in 2019, winning 3 awards. On the other side the movie that is on both lists, holding second place for the number of nominations and in the first place when it comes to the number of awards won, is "Titanic" with 16 nominations and 12 wins.

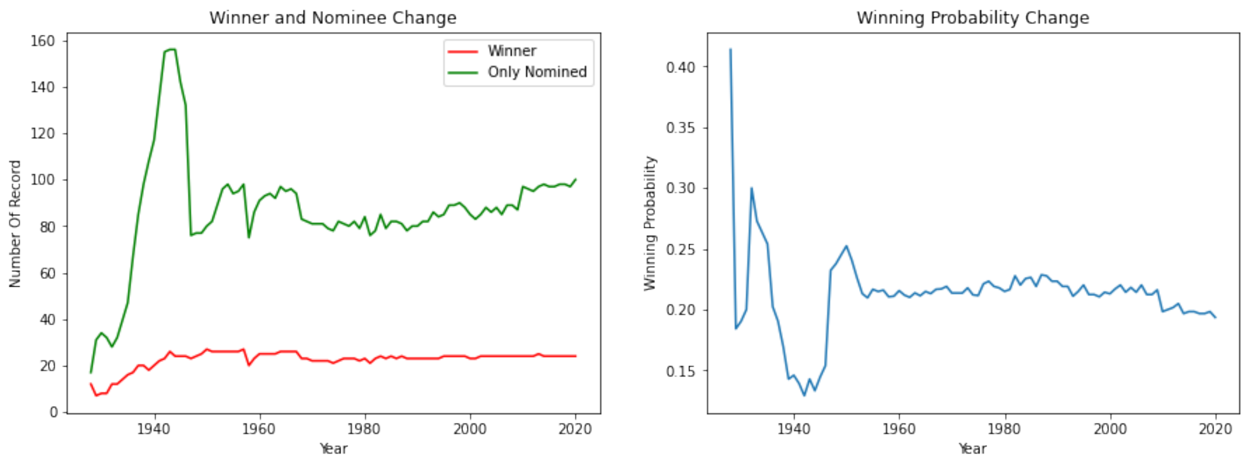


Figure 2: Change in award-winning rate per year.

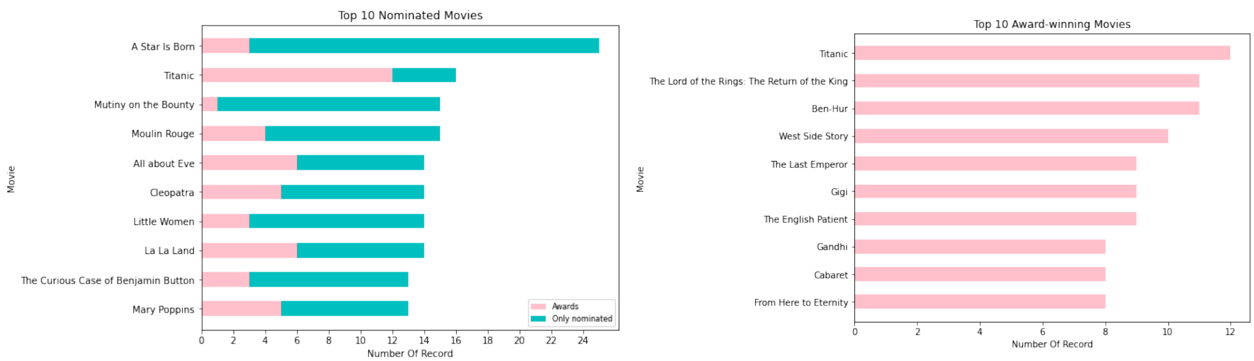


Figure 3: On the left, distribution of nominations and award wins of top 10 nominated movies. On the right, distribution of the top 10 award-winning movies.

The two plots on the Figure 4 represent the top 10 nominees and award winners at all times in all categories. We can see that the first two places in both cases are taken up by "Walt-Disney" and "Metro-Goldwyn-Mayer Studio Sound Department", and based on the data Disney holds the record for most Academy Awards won by an individual as a film producer, having won 22 Oscars from 59 nominations.

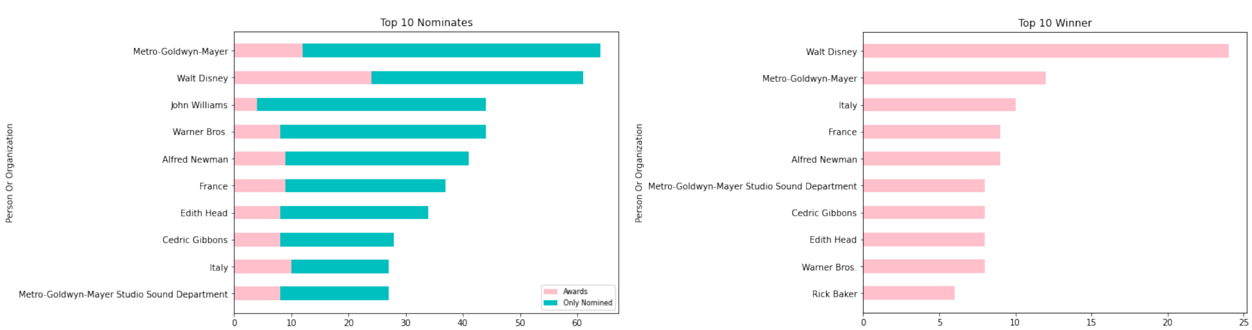


Figure 4: On the left, distribution of nominations and award wins of top 10 nominations overall. On the right, distribution of the top 10 award winners.

2.2 Academy Awards Demographics Dataset

The Academy Awards Demographics dataset is freely available on data.world website; moreover, it contains demographic information of the winners of Academy Award in various categories since 1928. It has 27 variables and some of them include information about religion, race, age, sexual orientation, etc.The winners that are considered are the winners in the following categories: Best Actor, Best Actress, Best Supporting Actor, Best Supporting Actress, and Best Director.

In this dataset, Figure 5 shows the demographic statistics on the Academy Award winners in the five main categories. Plot on the left shows that the Oscar winners are mostly Born in the United States of America (USA) and England. Also, the plot in the middle shows that winners are mostly 'white', while the plot on the right shows that winners are also usually 'straight' when it comes to the sexual orientation.

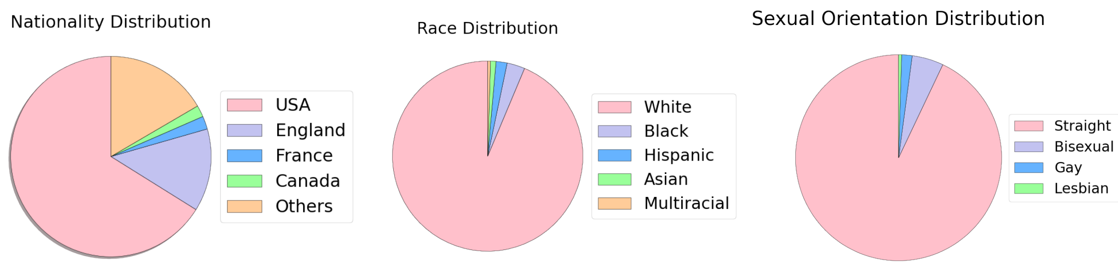


Figure 5: The nationality, race, and sexual orientation distributions among Academy Award Winners.(From left to right)

3 Data Preprocessing

Data preprocessing was done using Python; moreover, Pandas, NumPy, and Math libraries were used to handle the data, while MatPlotLib and SeaBorn were used for data visualization. At the beginning the Oscar Award dataset had 304 missing values which were immediately removed. We were left with 1092 unique values for all 7 columns. Next, all parentheses in the columns 'name' and 'film' were deleted; moreover, all categories were capitalized. Out of 7 variables, only 3 were used for the given tasks.

Regarding the Academy Awards Demographics dataset, firstly we replaced all strings 'Na' by numpy NA values. Then, instead of having many different countries in the 'birthplace' column, we wanted to focus only on 3 largest classes. To achieve this, we changed all country names that are different from 'USA' and "England' to 'Other.' The similar task was done in the 'sexual_orientation' column where all 'Gay' and 'Lesbian' values were united together under the label 'Gay/Lesbian'. Since 'religion' column had 269 missing values out of 441, this column was removed from the dataframe. We were left with 441 unique values and 8 variables out of which we used only 5 for creating the network.

4 Networks (Bogdan Milovanovic, Sara Kartalovic)

To build, explore, and analyze graphs, we used the combination of the Python's NetworkX libarary and Gephy interface. The network properties which were analyzed and are included in the Table 1 are following:

- **Graph type**:
  - *Directed graphs* are a class of graphs that don't presume symmetry or reciprocity in the edges established between vertices.
  - *Undirected graphs* are a class of graphs that imply an extra assumption regarding the reciprocity in the relationship between pairs of vertices connected by an edge.
- **Number of nodes**: Number of nodes in the network.
- **Number of edges**: Number of the edges in the network.

- **Clustering/Modularity**: Measures how well a network decomposes into modular communities. This algorithm will calculate an overall modularity score for the network, as well as assign each node to a separate cluster that will appear as a variable suitable for partion coloring.
- **Degree**: This is a numerical node variable. It is the number of edges connected to a particular node.
- **Node Size**: The range used for better visualising nodes.
- **Density**: This is a property of the whole network. It is the ratio between the number of edges and the total number of possible edges between the nodes.
- **Centrality**: In network analysis, measures of the importance of nodes are referred to as centrality measures. In Gephi we get centrality measurements using degree. Degree is a property of the whole network, not a single node or edge that represents the number of links between the two nodes in the network that are the farthest apart.
  - *Betweenness Centrality*: This is a numerical node variable. It is a measure of how often a node appears on the shortest paths between nodes in the network. Betweenness centrality, which is also expressed on a scale of 0 to 1, is fairly good at finding nodes that connect two otherwise disparate parts of a network. If you're the only thing connecting two clusters, every communication between those clusters has to pass through you. In contrast to a hub, this sort of node is often referred to as a broker. Betweenness centrality is not the only way of finding brokers, but it's a quick way of giving you a sense of which nodes are important not because they have lots of connections themselves but because they stand between groups, giving the network connectivity and cohesion.
  - *Page Rank*: This algorithm measures the importance of each node within the graph, based on the number incoming relationships and the importance of the corresponding source nodes. The underlying assumption roughly speaking is that a node is only as important as the nodes that link to it.
  - *Closeness Centrality*: This is a numerical node variable. It is the average distance from a given node to all other nodes in the network. Nodes with a high closeness score have the shortest distances to all other nodes. So, a node with high closeness centrality is literally close to other nodes.
  - *Eigenvector centrality*: This is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.
- **Clustering coefficient**: The clustering coefficients measure the average probability that two neighbors of a vertex are themselves neighbors (a measure of the density of triangles in a network).
- **Giant Components**: A giant component is a connected component of a network that contains a significant proportion of the entire nodes in the network.
- **Transitivity**: It is represented by using the ratio of all triangles over all possible triangles, and provides a way of assuming about potential relationships in the graph that are not presented currently.

When visualising the networks, different layouts were used in Gephy. Most of the networks are primarily represented using 'Fruchterman reingold' layout with additional application of the 'Label adjust' layout for easier representation of the nodes. Two networks made with the Oscar Award dataset, were also represented using 'Force atlas' layout, so that the communities differentiation is the main focus of the graph.

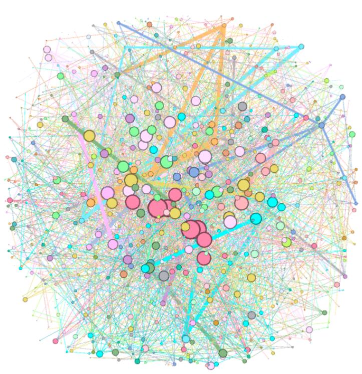
4.1 Film Network

Network science was applied on the processed data where the main focus were the similarities between the films. In the produced network, hubs indicate more Oscar nominations of the film, while the connecting

Network Property	Film Network	Person Network	Winners Network
Graph type	Undirected	Undirected	Undirected
Number of nodes	1253	1160	338
Number of edges	2097	1332	10023
Number of Communities	430	417	3
Node Size	1-70	1-70	1-30
Degree range	0-31	0-25	0-217
Density	0.003	0.002	0.173
Diameter	14	13	2
Clustering coefficient	0.682	0.569	0.319
Giant Components (Number of Nodes)	763 (60.99%)	687(59.22%)	219(65.57%)
Giant Components (Number of Edges)	1996(95.41%)	1253(94.07%)	8577(87.24%)

Table 1: List of the properties for each network.

edges show similarities between those films. For this network nodes are connected based on whether they have a common person (e.g. same actor, actress, director). The more people two films have in common, the larger is the weight of the edge. Most of the edges are weighted only by 1, while 48 out of 2097 edges are weighted more. The three edges, with the largest weight of 5, are between the following pairs of movies: "Silver Linings Playbook" and "American Hustle", "American Hustle" and "The Fighter", and "A Streetcar Named Desire" and "On the Waterfront".



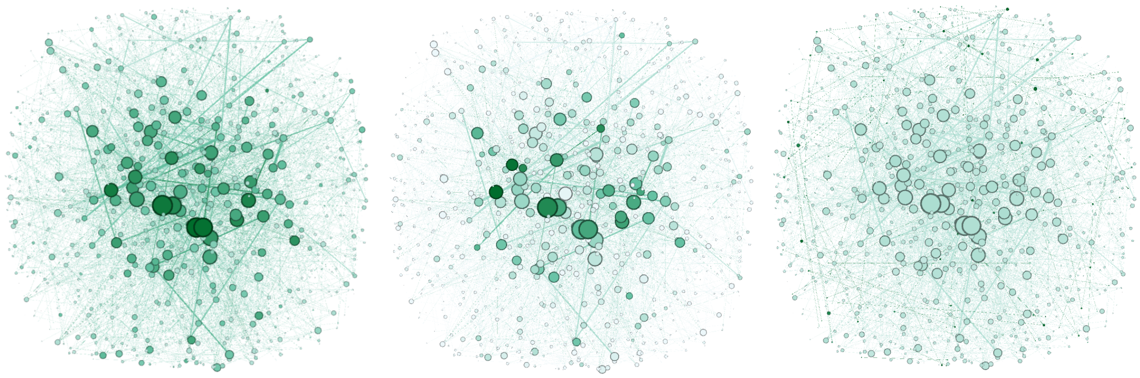
(a) Fruchterman reingold layout

(b) Force Atlas Layout

Figure 6: Different Film Network communities layouts

Total of 430 communities were discovered within the network through the use of Modularity on Gephi platform, then separated and arranged using the 'Force Atlas' and 'Fruchterman reingold' algorithms. These communities are shown in Figure 6 in the two different layouts. Force Atlas affects the distances, while the node sizes stay the same.

The network was not directed, therefore centrality aspects Page Rank and Degree were essentially the same for our networks. In Figure 7, Page Rank, Betweenness, and Closeness of the Film network are shown with the green color gradient. Darker color indicates higher value of the centrality aspect in question.



(a) Page Rank (b) Betweenness (c) Closeness

Figure 7: Film network centrality.

Nodes with highest page rank values are correlated to the hubs in the Film network as shown on the Figure 7a, and they correspond to the films with the most Oscar nominations. Similarly, betweenness seems to be high among the nodes with high degrees as shown on the Figure 7b, although there are some smaller nodes which have the darkest color indicating the highest betweenness. These films with highest betweenness create a bridge between biggest compartments in the network. Closeness is relatively low throughout the Film network which can be seen on the Figure 7c. This means that none of the films stand out as highly connected to all of the films in the network.

When we focus more at the distribution of different centrality measurements on the Figure 8, we find that top 5 nodes of each measurement are diverse. Some of the movies rank highly consistently such as "Julia" and "Ironweed". On the other hand, "Prizzi's Honor" and "A Star is born" have only high betweenness (if closeness is not taken into account since it was consistently low throughout the network), whereas "The Godfather" has only high page rank. According to these data, the most centralized films are "Kramer vs. Kramer" and "Doubt" with the highest page rank values and the second and third highest eigenvectors.

Additionally, based on the Film Network Centrality's Distribution on the Figure 8, we can conclude that nodes with the highest page rank are considered most important since their degree is the highest. Next, for the betweenness centrality nodes such as "Kramer vs. Kramer", or "The Godfather" connect two disparate parts of the network. We can also call them brokers since every communication has to pass through these nodes. It means that those nodes are important not because they have lots of connections themselves but because they stand between groups, giving the network connectivity and cohesion.

Interestingly, most of the highest centralized films belong to the same community; "Kramer vs. Kramer", "Doubt", "Ironweed", "The Deer Hunter", and "Silkwood" are all grouped in the same community by the algorithm. Therefore, "Julia", "Godfather, "Prizzi's Honor", and the "Africal Queen" are all in separate communities to each other.

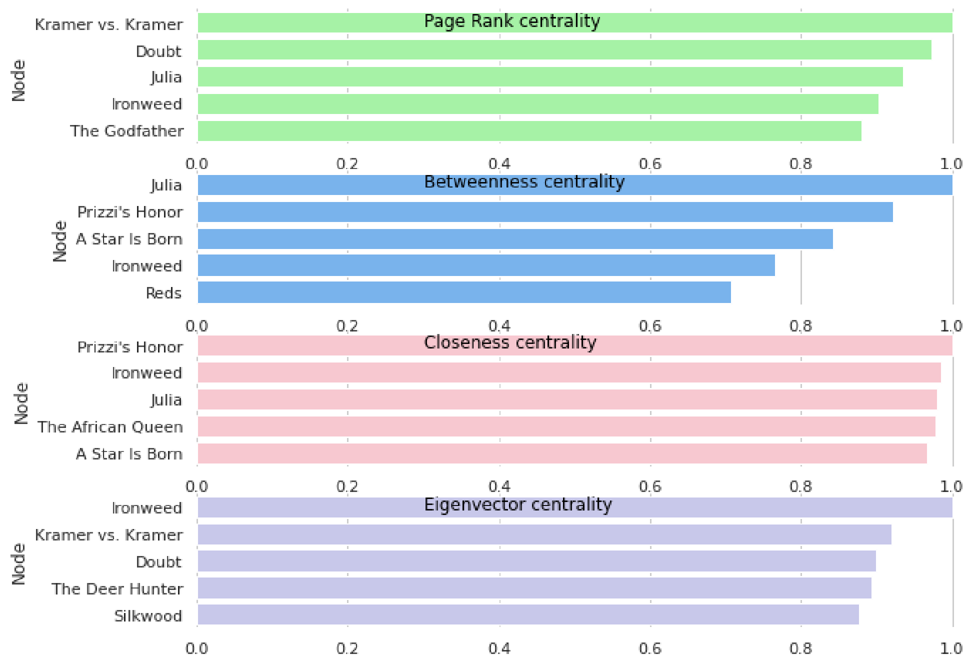
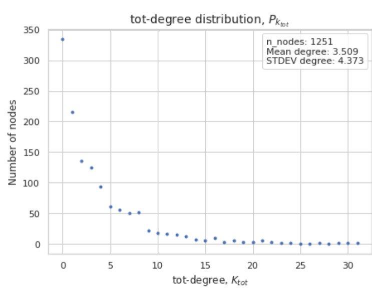
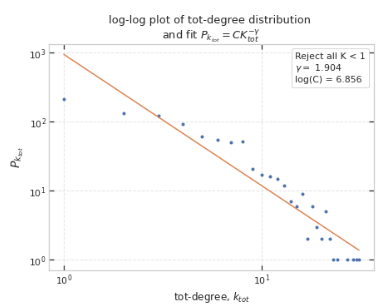


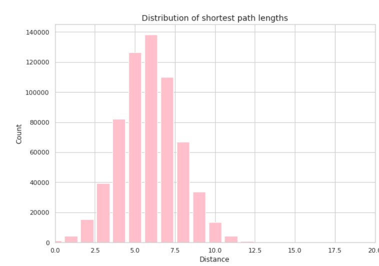
Figure 8: Film Network centralities distribution.



(a) Degree Distribution



(b) LogLog degree distribution



(c) Shorthest Path Distribution

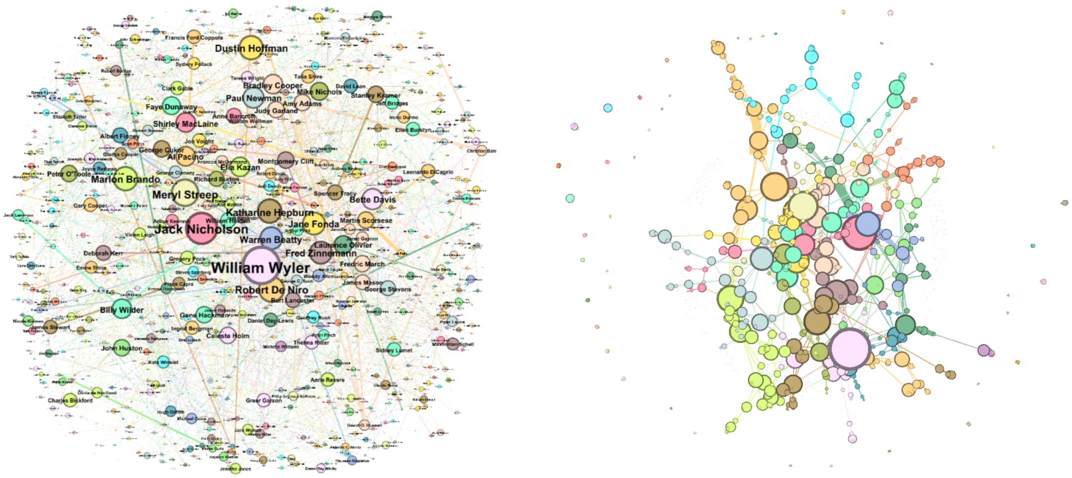
Figure 9: Film network statistics.

Figure 9 shows the statistics of the Film network. The values of degree distribution are spread out through the 1251 nodes with a mean of 3.509 and standard deviation(STDEV) of 4.373 as shown on the Figure 9a. Most of the films, approximately 300 nodes, have the degree of 0. The total degree distribution then decreases rapidly; for example, there are only around 20 nodes with degree of 10. The distribution shows that most of the films are not highly connected to each other in the network. In the Figure 9b degree distribution was presented on logarithmic scale, creating somewhat linear correlation between the two. The Log-Log plot of degree distributions gives us tailed distribution following a power law, and based on parameter γ, whose value is in the range $2 < \gamma < 3$, we can conclude that this network is scaled-free. Distribution of the shortest paths on the Figure 9c shows normally distributed counts of the distances. The most common distance is approximately diameter 6.

4.2 Person Network

The second network was built from the first Oscar Award Dataset, where each node represents the name of the nominee including 5 categories: actor, actress, actor in a supporting role, actress in a supporting role, and directing. The nodes are connected based on whether they have a common film, with the weight of their connection based on how many nominated films they have in common, while the node degree depends on the number of Oscar nominations.

The network which is represented on the Figure 10 consists of 1160 nodes and 1332 edges, and the node size used for representing this graph is in the range of 1-70. It can be seen that people in the center represented by the biggest nodes are big names who were nominated numerous times. Some of the names worth mentioning are William Wyler, Jack Nicholson, Maryl Streep, Marlon Brando, etc. Out of 1332 edges, 1291 have the weight 1 while 29 have the highest weight which is 3. Additionally, we found 417 communities represented with different colors.



(a) Fruchterman reingold layout

(b) Force Atlas Layout

Figure 10: Different Person Network communities layouts

First way the network was analyzed was by defining the structure of the communities, and this network can clearly reveal to us what are the relations between the nodes within the network. After merging the different datasets in Gephi, through the use of Modularity different groups have been identified (assigned random colors), separated, and arranged using the Force Atlas algorithm as shown in the Figure 10b. The large number of communities in this network can be attributed to the huge diversity among nominee, since they have been nominated always for different movies, as well as in various categories.

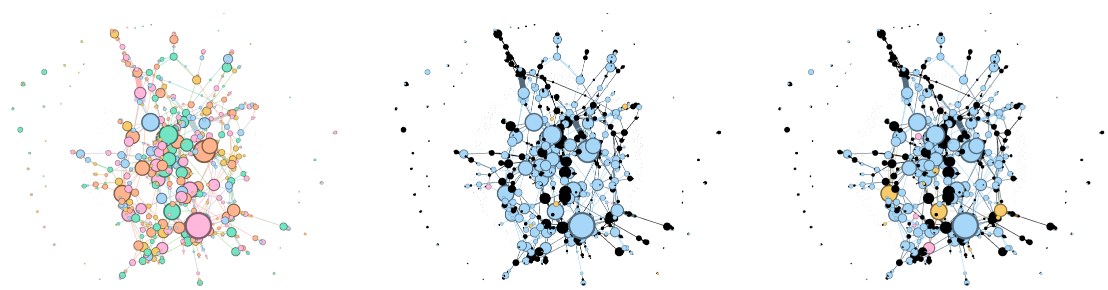
All shown networks so far have been build using the The Oscar Award dataset that contains all nominees and award winners. In addition to this, for our network analysis we included another dataset - The Academy Awards Demographics dataset, that contains demographic information of the winners of Academy Award in the top 5 categories including different factors and demographic data closely related to them, such as religion, race, age, sexual orientation, etc.

Because these two datasets did not include the same kind of information about nominees and winners, our idea was to join them and make networks that contain more attributes, so we can improve their analysis. In general, attributes are associated values belonging to a graph, vertices or edges. These can represent some property, like data about how the graph was constructed, the color of the vertices when the graph is plotted, or simply the weights of the edges in a weighted graph. Regarding this, the Figure 11 shows networks based on three different attributes:

- The Category:
  - Blue: Actor in a supporting role (23.53%)
  - Orange: Actress in a supporting role (23.36%)
  - Pink: Directing (19.91%)
  - Green: Actress (16.64%)
  - Dark orange: Actor (16.55%)

- The Ethnicity:
  - Blue: White (93.92%)
  - Orange: Black (2.74%)
  - Pink: Hispanic (1.52%)
  - Green: Asian (0.91%)
  - Black: Missing data (0.92%)
- The Sexual Orientation:
  - Blue: Straight (92.11%)
  - Orange: Bisexual (3.95%)
  - Pink: Gay/Lesbian (2.74%)
  - Black: Missing data (0.28%)

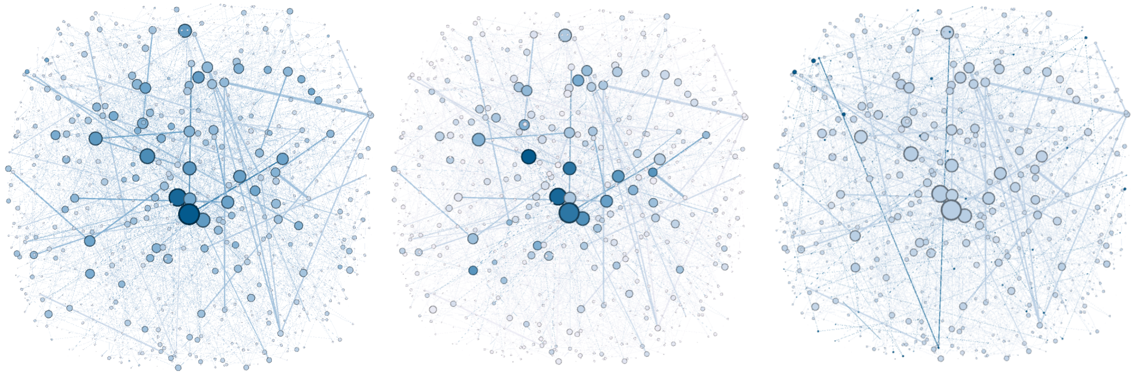
Black nodes represent missing values, due to the difference between the two datasets, since the Demographics dataset contains winners' data until 2014. Also, some of the nominated people from the first dataset have never won the award, therefore matching was not possible.



(a) Nominee Category (b) Winner ethnicity (c) inner sexual orientation

Figure 11: The Academy Award Demographics dataset united with the Oscar Award dataset to produce additional attributes.

The Figure 12 represents the Page Rank, Betweenness and Closeness centralities calculated for this particular network graph. When considering Page Rank, the underlying assumption roughly speaking is that a page is only as important as the pages that link to it. Since in our case we work with an undirected graph, we can notice that the Page Rank is proportional to the degrees of the vertices of the graph. The first centrality calculation shown is closeness centrality, and the highest closeness centralities are the nominees that have the closest film connections to all the other nominees. Another centrality metric displayed is betweenness centrality which indicates which nominees serve as the most important bridges for nominees to connect based on films they have in common.



(a) Page Rank (b) Betweenness (c) Closeness

Figure 12: Person network properties.

Furthermore, Figure 13 indicates nodes "William Wyler", "Jack Nicholson" and "Meryl Streep" with the once with the highest Page Rank, so we should consider them as most important since their degree is the highest. At the same time, the same three nodes have the highest Betweenness centrality. It means that those nodes are important not because they have lots of connections themselves but because they stand between groups, giving the network connectivity and cohesion. Moreover, high Closeness centrality values for given nodes show that they are directly connected to other nodes in the network, while a high score for Eigenvector centrality stands for nodes that pointed to many nodes.

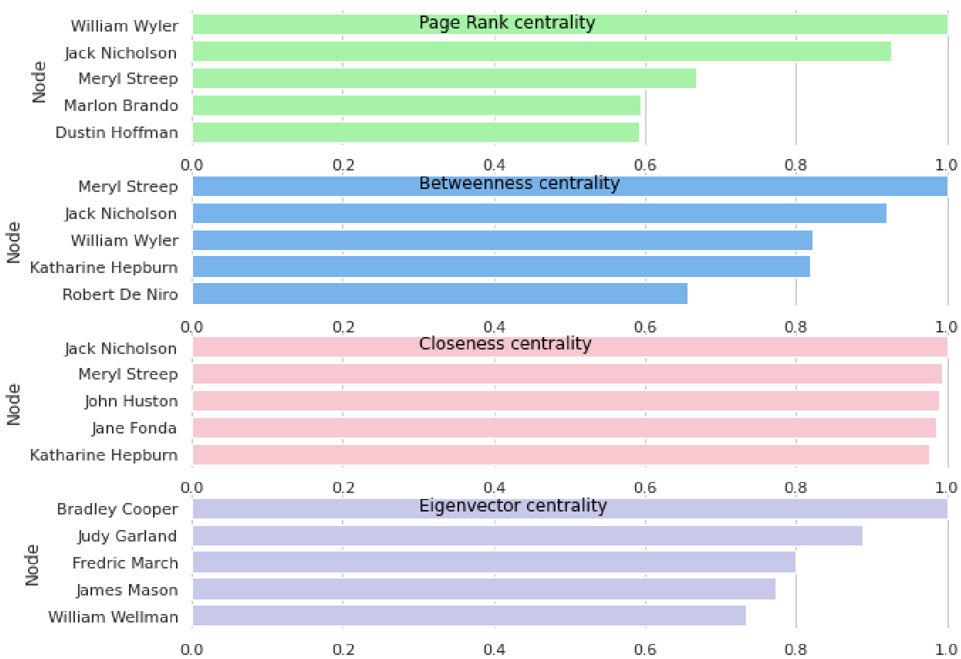
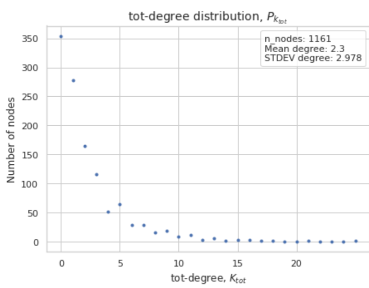
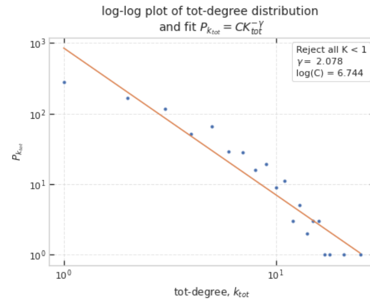


Figure 13: Person Network centralities distribution.

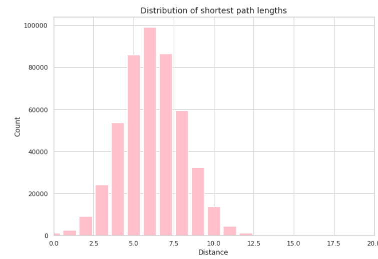
Plots from the Figure 14 reveal a high presence of nodes with a low level of degree, and then a few nodes with a high level of degree. Also it can be seen that that the majority of the nodes in the network have shortest path from 3 to 8, that represents a path between two nodes in a graph is a path with the minimum number of edges. Since our graph is weighted, it is a path with the minimum sum of edge weights.



(a) Degree Distribution



(b) LogLog degree distribution



(c) Shorthest Path Distribution

Figure 14: Person network statistics.

4.3 Winners Network

Firstly, it is important to mention that changes in the demographic background of nominated people significantly changed in the last couple of years. According to NewYork Times, there are new rules and at least one of them must be implemented. Three main rules are:

- At least one actor from an underrepresented racial or ethnic group must be cast in a significant role.
- The story must center on women, L.G.T.B.Q. people, a racial or ethnic group or the disabled.
- At least 30 percent of the cast must be actors from at least two of those four underrepresented categories.

Keeping this in mind, the third network was made using 4 attributes from the Academy Awards Demographics dataset where the nodes represent the name of the winners in the 5 selected categories.

The nodes are connected based on the number of the awards won and birthplace country. The size of the node describes the number of the awards that a specific person won, whereas there are 3 different communities in the original network as shown on the Figure 16 which represent 3 distinguishable groups 'USA', 'England', 'Others' (all other countries which were minority compared to the 'USA' and 'England'). The edge weight between the nodes depends on the number of the awards that person won. So if two people won more Academy awards, the edge weight between them will be higher than between two nodes that represent people who won just once.

This undirected network includes 338 nodes and 10023 edges; moreover, there were 3 communities detected which are represented with different colors (Blue: 'USA', Pink: 'England', Orange: 'Others') as represented on the plot 16a. The node sizes are represented between 1 and 30 while degree range is between 0 and 217. Out of 10023 edges, 7716 have the weight 1 while 18 edges have the highest weight 6. The density of the graph is 0.173 which indicates that out of all connections, there is smaller number of them. This network has the smallest diameter of 2 which means that the longest shortest path between two nodes in this network is 2. As shown on the Figure 15 which shows the distribution of shortest paths, majority of the edges have exactly the diameter of 2.

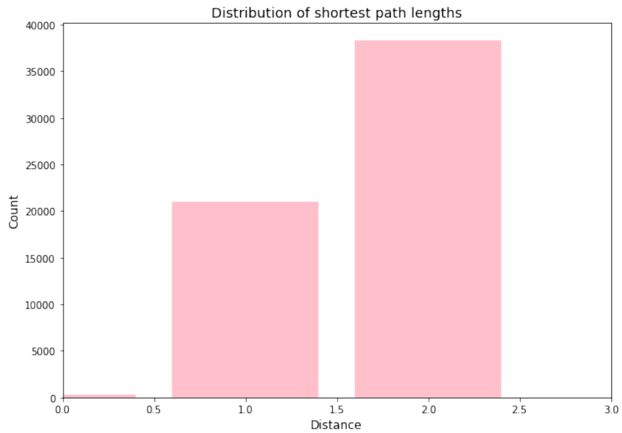
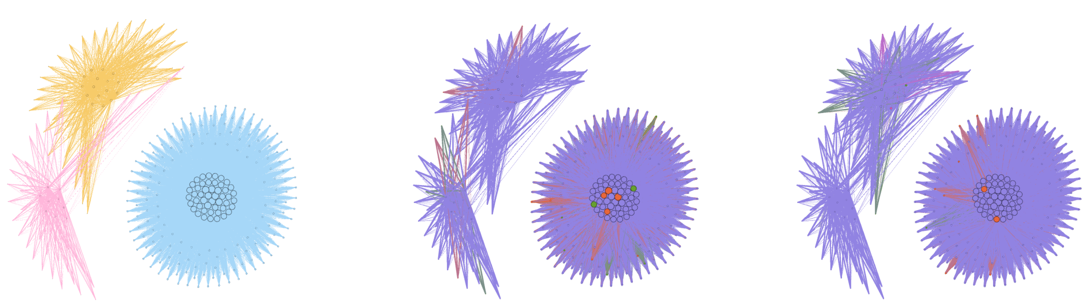


Figure 15: Winners network shortest path distribution

Since only the second dataset that includes demographics was used, in this network the focus was on only the Oscar winners. The Figure 16 specifications of the network attributes are summarized in the following list:

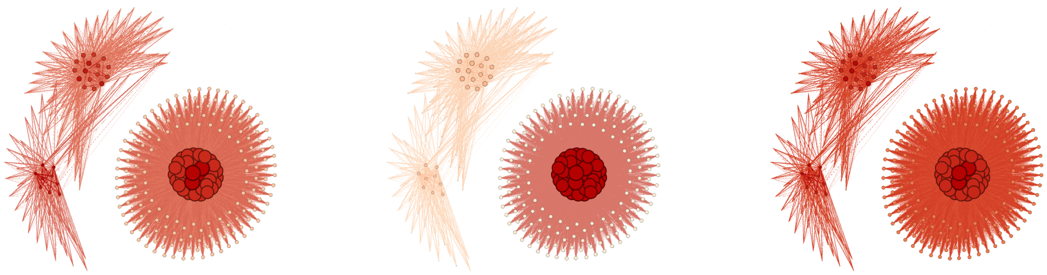
- The Birthplace:
  - Blue: 'USA' (65.67%)
  - Orange: 'England' (13.43%)
  - Pink: 'Others' (20.9%)
- The Ethnicity:
  - Purple: White (93.71%)
  - Orange: Black (2.69%)
  - Green: Hispanic (2.1%)
  - Pink: Asian (0.15%)
- The Sexual Orientation:
  - Purple: Straight (93.11%)
  - Orange: Bisexual (3.95%)
  - Green: Gay/Lesbian (2.99%)



(a) Birthplace (b) Sexual Orientation (c) Race ethnicity

Figure 16: Winners network communities

Figure 17 presents the visualisation of the graph according to its page rank, betweenness, and closeness centralities computed in Gephi and visualized using 'Fruchterman reingoid' algorithm. All three graphs produce similar results where people who won the most awards and were born in the same country are grouped together. It is evident that the closeness centrality has overall higher values for majority of the nodes. On the contrary, when looking at the betweenness centrality we can see that only 'USA' community nodes are high, while the other two groups are low. Page rank centrality produces the middle result, where middle nodes are high, and outer nodes are medium.



(a) Page Rank (b) Betweenness (c) Closeness

Figure 17: Winners network properties

Figure 18 shows the top five nodes by their correspondent centrality measurement. It is easy to see that top five nodes have almost same importance; furthermore, when we look at betweenness, closeness, and eigenvector centralities, top 5 people are same for each centrality.

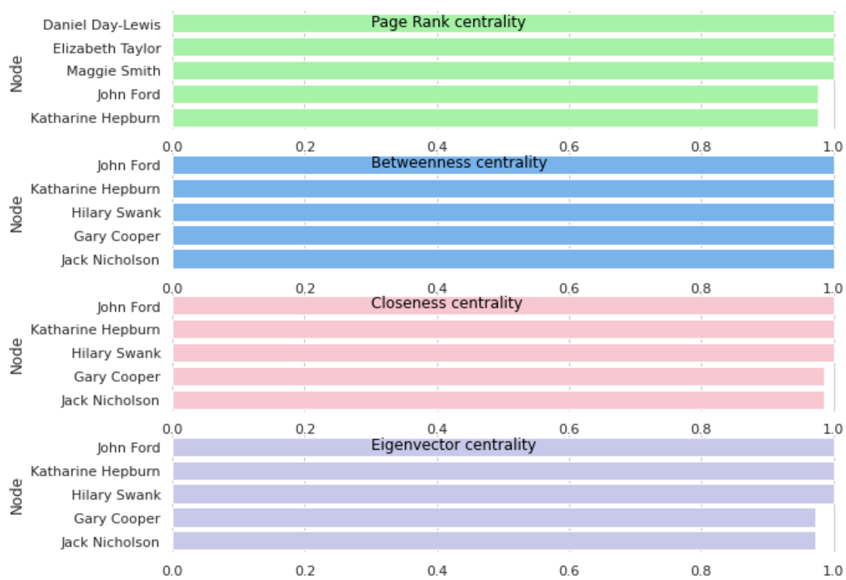
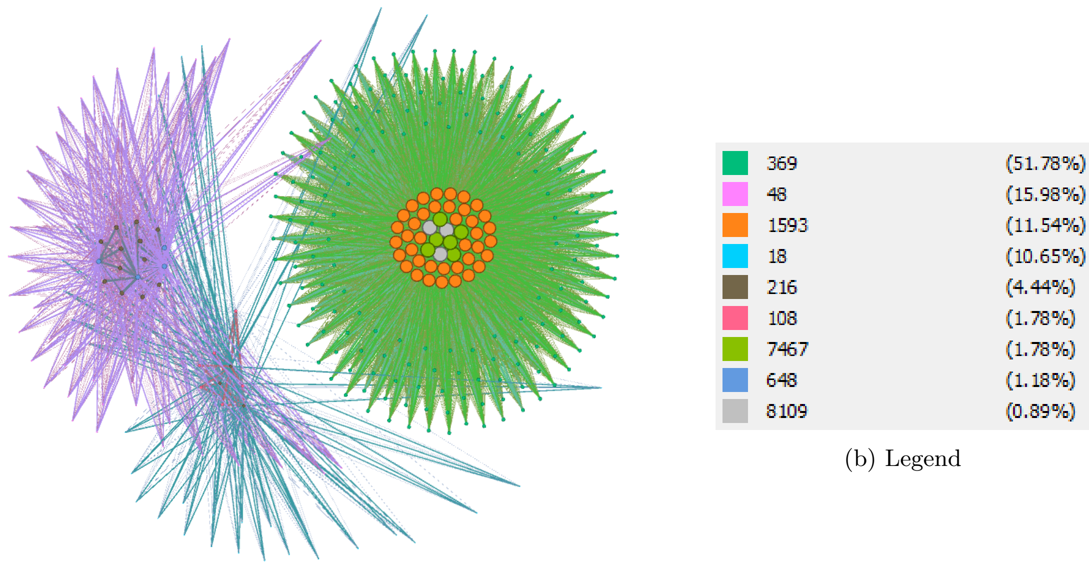


Figure 18: Winner Network centralities distribution.

The clustering coefficient of the network is 0.319 which represents the fraction of all possible triangles that are in the graph indicating connectivity of a particular region to its neighbors.The total number of triangles (cyclic paths of length three) that Winners network has is 68517. Each of these triangles can be seen as the set of exactly 3 nodes where each node has the connection with the other two. Also, these triangles can be used to find communities of actors or directors which worked together more than ones on different movies. Transitivity (concerns the concept of triadic closure which is equal to 0.189) is equal to 0.104, and because our network is not precisely dense, there is a small amount of possible triangles to begin with.



(b) Legend

(a) Number of Triangles

Figure 19: Number of Trianges

5 Conclusion

- Three different networks were created using 2 different datasets. The Academy Awards dataset was used for creating Film network, the Oscar Award dataset was used to create Winners network, and both of them were combined together to create Person network.
- The communities in the graphs were classified based on the films, people nominated and awarded, country of birth of the winners, and different attributes such as sexual orientation, race ethnicity, and category of nomination.
- Since density represents the ratio of actual edges in the network and all possible edges, low density would indicate that number of connections is low. Given that our three Film, Person, and Winner networks have 0.003, 0.002, and 0.173 densities respectively, the previous statement is proven because Winners network has more than 10000 edges and only 338 nodes while the other two networks have small gaps between the number of nodes and number of edges.
- All three networks are not connected because each one of them have more than one component. Because there are some isolated nodes which do not have paths to all other nodes, it is not feasible to find all shortest paths. For example, in Person Network, it is possible to find people who were nominated just once for a given movie and never again. They will be isolated because they do not have connections to other nodes who were nominated multiple times for different movies. Also sometimes, multiple people were nominated for the same film in different categories.
- Both networks related to The Academy Award Dataset are scaled-free, since the they follow a power law and the parameter γ is in the range $2 < \gamma < 3$.
- The Film and Person network have small closeness centrality values, while Winners network has very high closeness values. All three network show similar behaviour when it comes to page rank and betweenness centralities indicating that the highest number of shortest paths are passing exactly through the hubs.
- With all this being said, it is evident that most nominated films and people in different categories demonstrate that hubs are usually connected with each other, meaning that actors and directors worked frequently together on different nominated movies.