Network analysis of gender-based stereotypical representations of movie character relationships

### Abstract

Brief summary of the whole study (around 60-120 words), summarising the salient parts of the sections below.

### Context

The project is part of the field of social analysis and more specifically it is focused on conducting some gender relations studies applied to the cinematographic field.

### Problem and Motivation

The project aims at analysing whether there are stereotypical representations of character relationships in movies of various genres released in the span of time between 1915 and 2010s. Specifically, the hypothesis is that the representation of male-female relationships between characters in movies is subjected to difficulties that concretize in the tendency to shoot scenes in which characters interact mostly with character of the same gender. Moreover, whenever a relationship between a male character and a female one exists, this is usually stronger than the average, mostly of the time related to a love story. Assumed that male and female characters have stereotypical behaviours in the development of relationships within a movie, the author aim at demonstrate her hypothesis through the scientific method of network analysis.

The topic addressed is of particular interest for contemporary gender studies, which are nowadays increasingly questioning on the representation of gender roles in creative works. In particular, the project aims at contribute to the individuation, if present, of gender-based stereotypes influencing the representation of movie characters and thus at shedding light on our society gender prejudices and discriminations they are the manifestation of f.

### Dataset

The aforementioned hypothesis has been proved against real data from movies collected in ‘Moviegalaxies – Social Network in Movies’ open source repository in Harvard Dataverse[[1]](#endnote-1) by Kaminski Jermain, Schober Michael, Albaladejo Raymond, Zastupailo Oleksandr and Hidalgo César. The authors created a large collection of datasets related to 773 films (1915–2012) gathered through the employment of a movie script parser of their own invention that determined same-scene appearance of characters as a proxy of their connectedness. Each co-appearance is measured as one degree unit per scene, which has then been assigned as the weight of edges within the network. Although not free of minor errors, as the authors themselves declare, the data provided are very rich and exhaustive. Provided in JSON format, each movie takes the form of a network composed of a list of edges and a list of nodes. Each edge has its own weight given by the characters co-appearance in the movie; while each node is identified by an id and already provided with information about degree, pagerank, triangles, eccentricity, closeness centrality, betweenness centrality and eigenvector centrality measures. To these, the author of the project, in order to conduct her research, felt the necessity to add primary data related to character gender. By consulting IMDb movie database[[2]](#endnote-2) full cast pages related to the selected movies, each character has been provided with gender information and assigned to the group of male or female.

The dataset used for conducting the research is thus composed of a representative sample of 60 movies from the repository enriched with gender information for each character. The data, processed in a Python-readable form, have then been transformed in undirected weighted networks through networkx, a Python package for the creation, manipulation and study of complex networks.

#### Data sampling criteria

Firstly, it has been necessary to ensure that the sample of 60 movies extracted from the aforementioned population was a statistically significant one, i.e. a sample of the data that would not have invalidated the results of the research. Statistically speaking, a sample is representative of a population when it is both balanced and representative with respect to it: when it is able to reflects the different parts of the population and their sizes in it. In order to refer to objective sampling criteria, the author decided to extract a sample of movies from the database on the basis of their historical periods. In order to assure that movies from different historical periods were properly represented and that their proportions were respected in the sample, the author followed the method here illustrated.

From the same Hardvard Dataverse repository the dataset has been taken from, the author downloaded another file with information related to the metadata of all the 773 parsed movies. This is a tab file storing movies metadata such as the title, the imdb id, the release date and further information. Through the Python csv package, the author processed the file in a Python-readable format in order to extract the needed data.

Initially, all movie release dates were extracted and assigned to 10 date lists representing the decades going from 1915 to 2010. Then, for each movie whose release date was included in one of the decades, the movie title was assigned to the respective group of movies-per-decades. In this way it has been possible to count how many movies in the database were released in a certain decade and also to compute the percentage of them with respect to the total amount of movies in the database. Once that percentage has been calculated, it has been possible to compute the proper number of movies needed to respect the original proportion in a sample of 58 movies. At that point, the Python random sample buil-it function has been used to extract from the movies-per-decades groups the number of movies needed for each decade.

To these, the author manually added 1 movie from the 1915 and 1 movie from the 2012, reaching the total amount of 60 movies in the sample. The Moviegalaxies database indeed reports only one movie for these two decades, so that their percentage was nearly to zero on the overall amount of movies. However, to ensure that even these two periods, extremely relevant to the history of cinema, were represented in the research as well, the author decided to force their addition to the sample, even if this choice unbalances the original proportions.

The described method and the respective algorithm can be consulted in the file ‘movie\_sample’ in the author github repository; while the result it produced is here reported as the statistically significant sample used for the research:

* Movies from the 10s: ['Four Feathers']
* Movies from the 30s: ['Lost Horizon']
* Movies from the 40s: ["It's a Wonderful Life", 'Mr. Blandings Builds His Dream House']
* Movies from the 50s: ['Gunsmoke', 'White Christmas']
* Movies from the 60s: ['Star Trek', 'The Manchurian Candidate', 'Planet of the Apes']
* Movies from the 70s: ['Rocky', 'Star Trek: The Motion Picture', 'Barry Lyndon', 'The Getaway']
* Movies from the 80s: ['Gremlins', 'Hard to Kill', 'Top Gun', 'Gremlins 2: The New Batch', 'Raising Arizona', "Babette's Feast", 'Amadeus', 'Escape from New York', 'Batman']
* Movies from the 90s: ['Kundun', 'Lost in Space', 'French Kiss', 'Jackie Brown', 'The Matrix', 'Enemy of the State', 'Tomorrow Never Dies', 'One Eight Seven', 'Buffy the Vampire Slayer', 'The Crow: City of Angels', 'Donnie Brasco', 'Wild Wild West', 'Broken Arrow', 'Copycat', 'Bean', 'So I Married an Axe Murderer', 'Wonder Boys', 'Mumford', 'Nick of Time', 'Life', "My Best Friend's Wedding", "The Devil's Advocate", 'Being John Malkovich', 'Body of Evidence']
* Movies from the 2000s: ['Confessions of a Dangerous Mind', 'Rachel Getting Married', 'RocknRolla', 'Ali', 'Ninja Assassin', 'Mirrors', 'They', 'Punch-Drunk Love', 'Jason X', 'AVP: Alien vs. Predator', 'The Battle of Shaker Heights', 'American Gangster', 'Pirates of the Caribbean: The Curse of the Black Pearl', 'Twelve and Holding', 'Revolutionary Road', 'The Believer', 'Up', 'Orphan', 'Vanilla Sky', 'Jimmy and Judy', 'Black Snake Moan', 'House of 1000 Corpses']
* Movies from the 2010s: ['White Jazz']

#### Sample enrichment with primary data about gender of characters

Once obtained the reliable dataset reported here above, it has been necessary to enrich each movie in the sample with information related to gender of characters. To realize this operation, the author referred to the largest and most authoritative source for movie information on the Web: the Internet Movie Database.

Initially, it has been required to transform each movie's data about characters and their relationships provided by Moviegalaxies in JSON format in networks readable by Python. Through the Python package 'networkx', each movie has been transformed in a networkx graph object representing an undirected weighted network where characters are the nodes of the graph and their relationships are the edges connecting them weighted on the basis of how many time they appear in the same scene. Then, a Python simple function has been created in order to assign the attribute about gender to each node of the networks. By manually providing, movie by movie, a list – realized through the consultation of IMDb full cast page related to each movie – of female characters identified by the same id assigned to each node in the original database, the algorithm has been run assigning the attribute 'gender:female' if the node id was included in the provided list, and the attribute 'gender:male' if it was not. The 60 movie networks enriched with gender information related to each characters obtained through this method can be consulted in the file 'movie\_networks' in the author dedicated github repository.

It is finally to be noticed that only male/female distinction of characters has been considered in the research not to exclude the existence of minorities other than the binary distinction of gender, but because no case was found, within the more than XXX characters investigated, that needed a different gender attribution.

### Validity and Reliability *(not needed for the project proposal)*

The author is aware of the biases in a dataset such sampled. Firstly, its nature and dimensions reflect the minor errors and biases already existent in the whole Moviegalaxies database, that is naturally unbalanced in the representation of movies from different historical periods and genres. By mean of example, no movie from the 20s has been included in the original database; and the great majority of movies in it are representative of a very precise cinematographic reality: the Western cinematographic tradition. Secondly, randomly extracting an amount of movies from each decade through Python's sample function does not exclude the possibility that all the extracted movies are from the same year or however not representative of the whole decade. Furthermore, cinematographical genre distinction has not been taken in consideration in realizing the sample: nonetheless, it has to be noticed that a proper and unambiguous movie genre attribution is a quite difficult task and one that would have made the sampling much more complicated and devious. However, from a first glance to the sample, different genre presence seems to be quite represented in the extracted movies. Finally, it has to be noticed that Python's sample function generates a different result each time the algorithm is run. Different executions generate different movie titles lists to be included in the dataset. However, the author decided to use the ones resulting from the first execution as reported above.

Considered the previously cited sampling criteria, the model used for the research can said to be valid with respect to the Western cinematographic tradition from 1915 to 2000 and reliable with respect to its capacity of reproducing data similar to the observed population. The dataset is representative of movies significantly extracted from different historical periods and enriched with gender information related to their character.

### Measures

The main aim of the research is to analyse whether there are stereotypes in the representation of movie character relationships due to gender of characters. Relationships, as previously outlined, are represented through scene co-appearances of characters in a movie and assume a specific strength dependently on how many times they are retrieved by the movie script parser. The analysis has been conducted on three different major points:

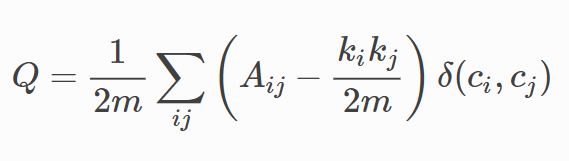
1. whether movie characters tend to develop relationships mainly with characters of the same gender;
2. whether movie characters tend to develop stronger relationships with character of the opposite gender;
3. whether it could be possible, given the first two assumptions, to detect network communities in movies able to represent the gender division of characters.

Finally, descriptive statistics has been used on the results provided by network analysis in order to investigate whether there are different character relationship representations in movies of different historical periods and genres.

#### 1) Do movie characters tend to appear in the same scenes more than characters of different gender? Modularity.

Network analysis has been conducting different studies, gathered under the definition of the so-called ‘homophily’ or ‘assortative mixing ’studies, on the tendency of actors to develop relationship with other they perceive similar to them. More specifically, homophily, or assortativity, is a preference for a network's nodes to attach to others that are similar to them in some way. The study of whether movie female characters tend to develop relationships with other female, just as male with other male, can definitely be included under the same label of homophily studies. Furthermore, the best solution network analysis provides to calculate the extent to which this tendency of nodes to connect with nodes of the same kind in a network is the so-called measure of modularity.

Modularity is mathematically defined by the following formula:



where mm is the number of edges, AA is the adjacency matrix of the graph, kiki is the degree of ii and δ(ci,cj)δ(ci,cj) assumes the value of 1 if ii and jj are in the same community or module and 0 otherwise. In this sense, modularity is able to quantify the level of non-randomness in the placement of edges in a network’s partitions.

Thanks to the respective built-in function in the networkx package, it is quite easy to calculate modularity using Python. Indeed, by applying the function defined by the package as .modularity(), which takes as parameters the networkx graph 'G', the node sets ‘communities’ representing a partition within G’s nodes, and optionally the edge attribute that holds the numerical value used as a weight for G's edges, the algorithm returns as result the modularity Q of the partition. This latter is a floating-point number that is always strictly less than 1 and takes positive values if there are more edges between nodes of the same type than we would expect by random chance or negative values if there are fewer such edges than we would expect by chance.

Within the context of the project, modularity of graph partition representing on the one hand the group of male characters and on the other hand the group of female ones has been calculated for each movie in the sample on the basis of the following method. Firstly, a Python list has been created with all the 60 movie networks enriched with node attributes referring to the gender of each character as previously described. Then, a simple Python function able to iteratively access to each of the movie network in the list has recreated the two groups of male and female characters for each movie, which have then be used as the communities parameters of the modularity built-in function representing the partition of the interested graph. The weight of edges has been considered by iteratively accessing the respective attribute value for each network. The algorithm, running for each movie in the list, also allowed to store each network’s modularity value in a separated list. On this Python’s built-in functions from the statistics package have been employed in order to compute the average modularity for all the 60 movies and the standard deviation of the same value.

#### 2) Do movie characters tend to develop stronger relationships with character of the opposite gender? Connection strength computation model.

Of course, in order to establish how much relationships between movie characters are influenced by gender-based stereotypes, the calculation of the probability of their connection on the basis on gender is not a sufficient measure. Another relevant factor to consider is indeed how much strong character relationships are and whether the strength of their connectedness vary on the basis of character gender. As previously outlined, the networks coming from the representation of character relationships in movies are weighted ones, where weight is measured as 1 unit for each time the characters co-appear in the same scenes and assigned as attribute to the respective edge connecting them. The weight such calculated is thus a representation of the strength of the connectedness of characters, and it is indeed the variable used to calculate the strength of their relationships. Given the gender variables considered in this project, the meaningful connections that can be valuated are of three different kinds: male characters connected with male characters, female characters connected with female character and finally male characters connected with female characters. They are exactly those whose strength has been calculated in the second task of the project on the basis of the following method.

The process of accessing the data is the same as the previous task: an iteration on the list of movie networks enabling the author to access to the data of each movie. Then, the task has been articulated in three different moments regarding the previously outlined combination of variables: male-male, female-female and male-female. Firstly, for the first two combinations, gender attributes have been used in order to recreate the two respective groups of male characters and female ones. In each of them, an iteration along the list of male and female characters has been performed in order to check whether they had relationship with other characters of the same gender and, if an edge was retrieved between the two nodes, the weight associated to it was stored in a list dedicated to the computation of the overall relationship strength respectively for male with male and female with female. The same has been done for the third combination, regarding relationships between male and female, except for the recreation of lists of male or female character, as it was no more a significant information. Once having retrieved the weight of each relationship connecting firstly male with male, secondly female with female and finally male with female, built-in Python statistical algorithm have been used to calculate the mean and the standard deviation of character relationship strength.

The final result of this process is a set of three value representing respectively the average strenght/weight of the relationships between male with male characters, female with female characters and male with female characters. An analysis of their meaning will be conducted in the section dedicated to the discussion of the results.

* T-test for dependent variable??? A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.
* test of differences between means. Now we are interested in goodness of fit: if an observed measure of central tendency (thus, in case of interval/ratio variables, a mean) is significantly different from another already known measure.

Null Hypothesis: Strength of relationships between male and female characters is the same as the average

Alternate Hypothesis: Strength of relationships between male and female characters is stronger than the average

Perform the t-test to understand if male-female strength distribution is different from strength distribution regardless the gender: t = 5.316577084154557

p = 1.0657581217543566e-07, t2 = 2.9190366119786075

p2 = 0.0035161524891003364: reject the null hypothesis [https://towardsdatascience.com/inferential-statistics-series-t-test-using-numpy-2718f8f9bf2f]

#### 3) Is it possible to predict gender-based communities of characters by analysing network's topology? Community detection algorithms.

Assumed that stereotypes exist which influence the representation of male and female characters within a movie so that all nodes representing male or female characters in a network modelling the same movie tend to have similar behaviour, it becomes natural to question whether it could be possible to apply some measures enabling us to distinguish these two groups of nodes within the network. In network analysis, the problem of finding groups of nodes in network is called community detection. The aim of the third task of the research is thus exactly the one of performing some community detection approaches in order to analyse whether it would be possible to find character communities based on their gender.

Generally speaking, the primary goal of community detection is to find the natural divisions of a network into groups of nodes such that there are many edges within groups and few edges between them, with the main aim to provide a better understanding of its structure and organization in communities which are similar in some way.

Although it is quite immediate to realize what network communities are network and what are the parameters which may help to identify them within it, community detection is actually a still quite challenging task in network analysis. It is particularly difficult, indeed, to approach the problem from a mathematical and abstract point of view, reason why many different approach and methodologies exist in order to perform community detection within networks. Within the context of this research, the author applied two different approaches in order to understand whether it would have been possible to detect communities in movie networks corresponding to male and female character groups.

The first approach applied to the movie networks in the dataset is also one of the most widely used and it is known as the method of modularity maximization. It deals with the problem of community detection by approaching the task as an optimization problem and particularly it considers the optimal network partitions the ones able to maximize its modularity. As previously outlined, modularity is an assortative mixing measure which assumes high value when connections are primarily between nodes of the same type and a low value when they are not. Starting from this assumption, the modularity maximization method, by assimilating communities to groups of nodes of the same kind, considers good divisions of the network precisely those that have high values of modularity.

The most effective way to apply modularity maximization method for community detection in Python is through the best\_partition() algorithm implemented by the community package. The algorithm refers to the Louvain's 'best partition' method to detect communities within a network, an heuristic algorithm for approximately maximizing modularity over divisions

of a network into any number of communities. Louvain agglomerative algorithm is a well example of modularity maximization methods, thanks to its approach to community detection which works by taking single nodes and joining them into groups, then repeatedly joining groups with other groups in an effort to find the configuration with highest modularity. The community best\_partition() algorithm computes the partition of the graph nodes which maximises the modularity using the same Louvain heuristices. The communities partition resulting from it is the partition of the (approximate) highest modularity, i.e. the highest partition of the dendrogram generated by the Louvain algorithm.

### Results

What is the connection among: the gathered data, the applied measures, and the properties found?

#### 1) Characters in movies tend to develop relationships with opposite gender characters. Low modularity values

The aim of the first task of the project was to compute modularity of movie networks partitioned in male and female groups of nodes in order to understand whether there were more relationships between same gender characters than we would expect by chance. The values resulting from the application of the networkx package's modularity() algortihm to calculate the modularity of each movie in the dataset have then been used to compute the arithmetic mean for the average modularity of all the movies, resulting:

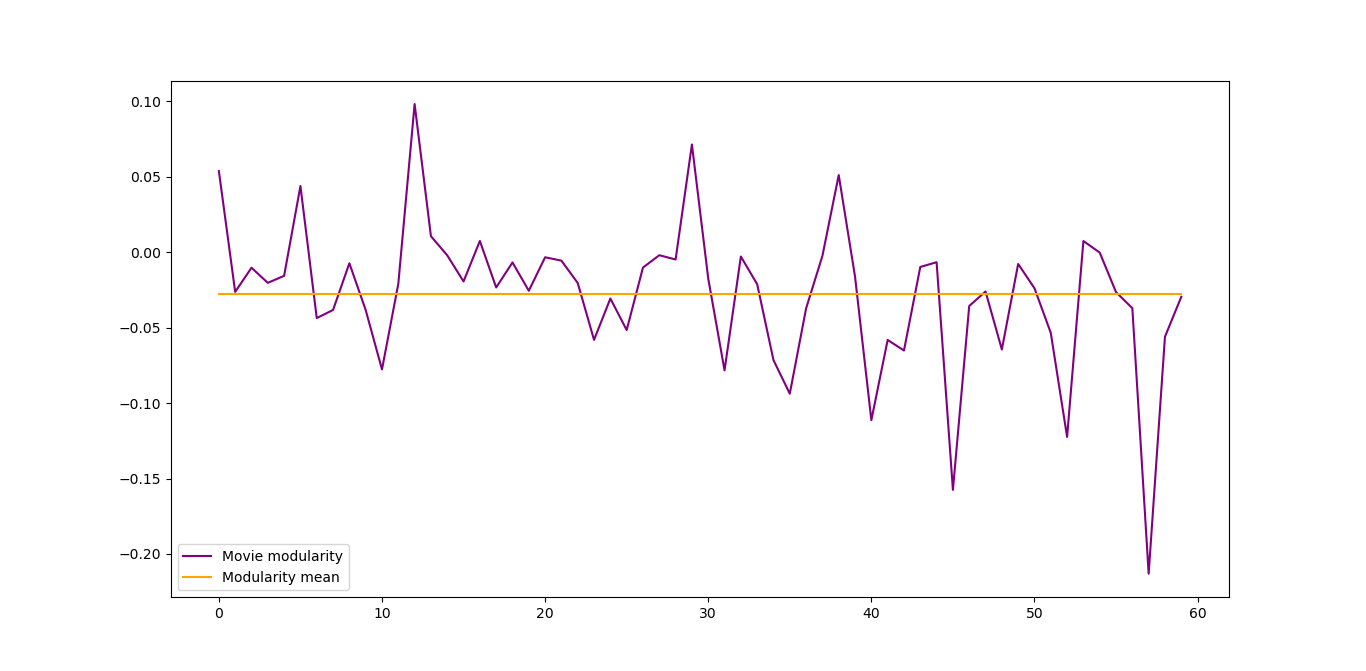
Modularity mean: -0.027670926735455754

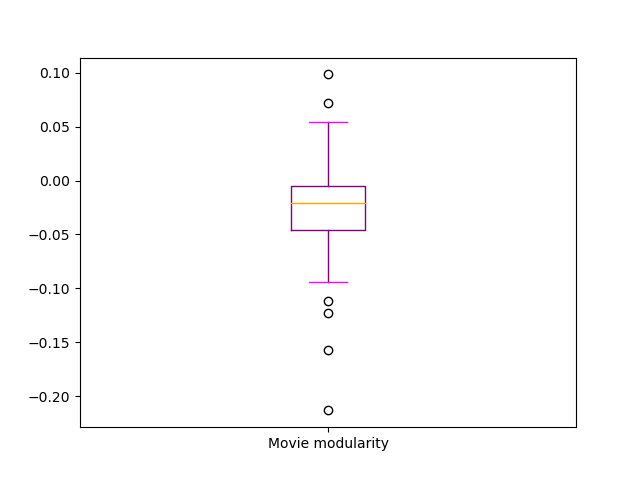
The mean value of modularity of all 60 movies is slightly below zero, indicating a disassortative trend of male and female nodes in the networks representing the movies. Indeed, although rather close to the neutral value of modularity, i.e. 0, a negative average value allows to reject the initial hypothesis according to which gender-based stereotypes led to a representation of the characters in movie biased towards a major co-presence in scenes of characters of the same genre. Contrarily, movie characters seem to have a slight tendency to appear in scenes with characters of the opposite gender.

However, although the arithmetic mean is the measure of central tendency in this case able to best return the average value of modularity, it is not enough to describe the behaviour of the data without their respective measure of dispersion. Hence, the calculation of the standard deviation, the measure of the amount of variation or dispersion of the set of values used for the mean, has been performed in order to return a better understanding of the data, resulting.

Modularity standard deviation: 0.048620735964211764

The low value of standard deviation indicates that the calculated modularity values generally tend to be close to the mean; nonetheless from the graphical visualization of the of the modularity trend for all the movies (Fig. 1) and the boxplot graph visualizing the standard deviation (Fig. 2) allow to spot relevant outliers among the whole amount of investigated networks.





Although many values are in fact around the calculated mean, there are peaks of positive modularity as well as peaks of negative modularity. The movie with the greatest positive modularity, and therefore a greater tendency to represent scenes with characters of the same genre, is .... while the one with the least modularity is ..... In any case, in the second half of the sampled movies, the most recent ones, shown in Fig. 1 there seems to be a trend towards negative modularity values. An investigation into the differences between various historical periods and movie genres will in fact be conducted in the second part of the discussion of the results.

#### 2) Character relationship strength within a movie is not particularly affected by characters' gender. Similar average values for male-male/female-female/male-female connectedness strength

The second task of the project regarded the differences in the strength of developed relationships between characters. Particularly, it has been investigated whether there were differences in average strength of relationships between male characters with other male, female characters with other female, and male characters with female ones. The arithmetic mean and the standard deviation of all the relationships strenght of these three different kind have been calculating, resulting:

Average female-female relationship strenght: 1.7263182835851236

Female-female relationshp strenght standard deviation: 0.18407555429307337

Average male-male relationship strenght: 1.607859263662848

Male-male relationshp strenght standard deviation: 0.05983709412361456

Average male-female relationship strenght: 1.8569831208925593

Male-female relationshp strenght standard deviation: 0.30968522560156625

Average relationship strength regardless the gender: 1.6781771795384097

Relationshp strenght regardless the gender standard deviation: 0.06549409502833627

#### 3) There are rare cases in which community detection methods are able to detect groups of nodes formed only by characters of the same sex.

#### 4) Are there meaningful tendencies in the representation of movie characters along historical periods and movie genres? Descriptive statistics.

### Critique *(not needed for the project proposal)*

Do you think your work solves the problem presented above? To which extent (completely, what parts)? Why? What could you have done differently to answer your research problems (e.g., gather data with additional information, build your model differently, apply alternative measures)?

1. Moviegalaxies – Social Network in Movies, Hardvar Dataverse repository by Kaminski Jermain, Schober, Michael, Albaladejo Raymond, Zastupailo Oleksandr and Hidalgo César, <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/T4HBA3> [↑](#endnote-ref-1)
2. IMDb Movie Database, <https://www.imdb.com/> [↑](#endnote-ref-2)