Professional Networks and Career Opportunities for Directors in Hollywood

Abstract

The broad research question of this study is to characterize which directors get the most career opportunities in Hollywood and why. To that end, three different themes are explored. The first is a descriptive study that looks at the distribution of resources and opportunities in the industry. The others investigate the characteristics of the directors who are the most successful in attracting these opportunities. The second section analyzes how directors and producers interact in the film making market and the third evaluates whether merit or connections play a more important role in making the careers of directors. The study reveals that Hollywood invests most resources on a few directors mostly based on their connections rather than their actual creative potential.

Keywords: Hollywood careers, network analysis, computational social science, machine learning

1. Introduction

In Hollywood, opportunities and resources are highly concentrated in the hands of a few. This 'superstar' phenomenon where disproportionately few individuals rise to the top has been widely studied for decades in the social sciences [26, 12, 1]. This is particularly true in Hollywood because it is an artistic industry [23]. The cause of this has been attributed to multiple reasons in various works.

Faulkner [11] suggests that this could be because producers invest a lot of money into a movie and they want to be as risk-averse as possible. Therefore, they try to hire the best talent in the industry. This is feasible because hiring costs are only a fraction of the overall cost and usually most big-budget movies can afford the best talent. According to Rosen [26] and Adler [1], with the advent of mass media (like the television), art could be produced cheaply and in bulk, so there is no supply constraint. So it is possible to supply the artists' work to a large number of people at almost fixed marginal cost. But in this case, since the number of buyers can be very high, a small difference in talent can lead to a large difference in demand. Therefore to ensure maximum profits, the investors and filmmakers try to get the best artists. Finally, there is another theory that since there is generally no objective way to signal creative talent, most people find it costly to try and evaluate every single candidate in the market. In these cases, once an artist hits a threshold of popularity, 'herd behavior' sets in and most people just demand from already popular faces as it reduces the exploration cost to find good work [3, 14]. All this leads to a market

structure where some directors get disproportionately more opportunities.

Some directors have strong connections with powerful producers, and this connections may possibly present them with more movie-making opportunities. Producers raise funds from various investors and sometimes fund the movie themselves. They make most business decisions on behalf of the permanent investors of the project. While directors are the 'creative' leaders of a movie, producers are the 'business' leaders. In Hollywood, the process of movie-making loosely follows the same stages [9, 10, 27]: laying the fundamental idea for the project, recruiting a preliminary team, hiring writers to improve the script and to develop the project, roping in major directors and stars, seeking investments from potential financiers, and finally hiring the rest of the crew and setting a production timeline. Through the workflow of movie-making, we see that producers are the main 'bosses' of filmmaking - they control the project. They also have a significant say in who gets to work in the project and how much investment will be made. Therefore, associating with good producers is very important for a director and this relationship is another indicator of the director's power and reputation in the filmmaking circles. The formation of director-producer links in this social structure is governed by market economics. On one hand, producers want to hire directors from a large but mainly inexperienced talent pool [8, 18] with their limited funds. Top directors on the other hand receive many offers but cannot accept them all because of limited time. In this marketplace, producers and directors, especially the top ones, both have to makes choices on which

parties to work with.

Aside from connections with producers, it is often beneficial for a director to also associate with other powerful figures in Hollywood. But how important are these connections to a director's career prospects? It is often implied that 'whom you know' is more important than 'what you know' to prevail in Hollywood. One reason for this is because it is often difficult to assess creative talent, and in most cases, reputation is taken as a measure of talent and hence becomes a key driver of success [4, 21]. One way of attracting a reputation is by being involved in critically or commercially successful projects. The other way is by associating with the right artists in the industry and building a strong professional network. While many studies validate the claim that networks are indeed important in determining the career outcomes of artists in Hollywood [25, 7, 11, 21, 27], these studies either do not quantify the effect of network versus the personal traits of the artist on their success or perform standard methods like regression to understand the effects.

This paper is structured into three main sections. The first section (Section 3), quantifies the current state of resource and opportunity allocation in Hollywood. It finds that a few directors get the majority of the resources in the industry. In Section 4 and 5, the role of connections in the making of these top directors is explored. Section 4 performs blockmodeling to understand how various directors and producers interact to make movies. Section 5 mainly looks at a prediction problem - to check whether artists' connections or their merit is a stronger predictor of the

opportunities they get in Hollywood. These sections show that top directors and top producers prefer each others and that connections are have a stronger influence on director opportunities than their actual creative talent.

2. Data

The data for this study has mostly been gathered from four sources: (1) Internet Movie Database (IMDb)[1], (2) The Open Movie Databse (OMDb) API[2], (3) Bechdel Test API[3], and (4) The Numbers website[4]. A total of 3891 movie titles released in at least 500 screens between 2000 and 2018 has been collected. This dataset contains rich information about the movies including its cast and production team, content information, performance ratings, financial information, etc. From the dataset, the key information used for this study are summarized below (more details on the data collection method is explained in the Appendix section 8.1):

- Movie Title, IMDb ID: These are the unique identifiers for each of the 3891 movies. When two movies have the same title, the IMDb ID can be used for differentiation between the two.

- Actor(s), Writer(s), and Director(s): For each movie, the list of writer(s), director(s), and main actor(s) who have contributed to the movie is given along

[1]https://www.imdb.com
[2]https://www.omdbapi.com
[3]https://bechdeltest.com
[4]https://www.the-numbers.com

with their role. Overall, the dataset comprises of 6832 unique actors, 5468 unique writers, and 2361 unique directors.

- Production company: For every movie, the list of companies funding the movie project is also given. There are 712 unique production companies in the dataset. The average number of production companies per movie is 1.1.

- IMDb Rating, Oscars Wins, Number of Awards, Number of Nominations: These variables measure the performance of the movie in terms of its quality. IMDb ratings are ratings given by the public to the movie on a scale from 1 to 10. Oscar wins gives the number of awards the movie received at the Oscars, while the no. of awards and the no. of nominations is for all the major movie festivals including the Sundance and Cannes.

- Box Office Collection, Production Budget: These variables give the financial data for the movie, i.e. how much revenue the movie generated world-wide and how much budget was spent in its making (all in USD).

- Bechdel Test Rating: The Bechdel Test rates a movie from 0 to 3 to measure gender representativeness in movies. This score takes the value 1 if there are two named women in the movie, value 2 if they talk to each other, 3 if they talk about something besides a man, and 0 if none of the above conditions are met.

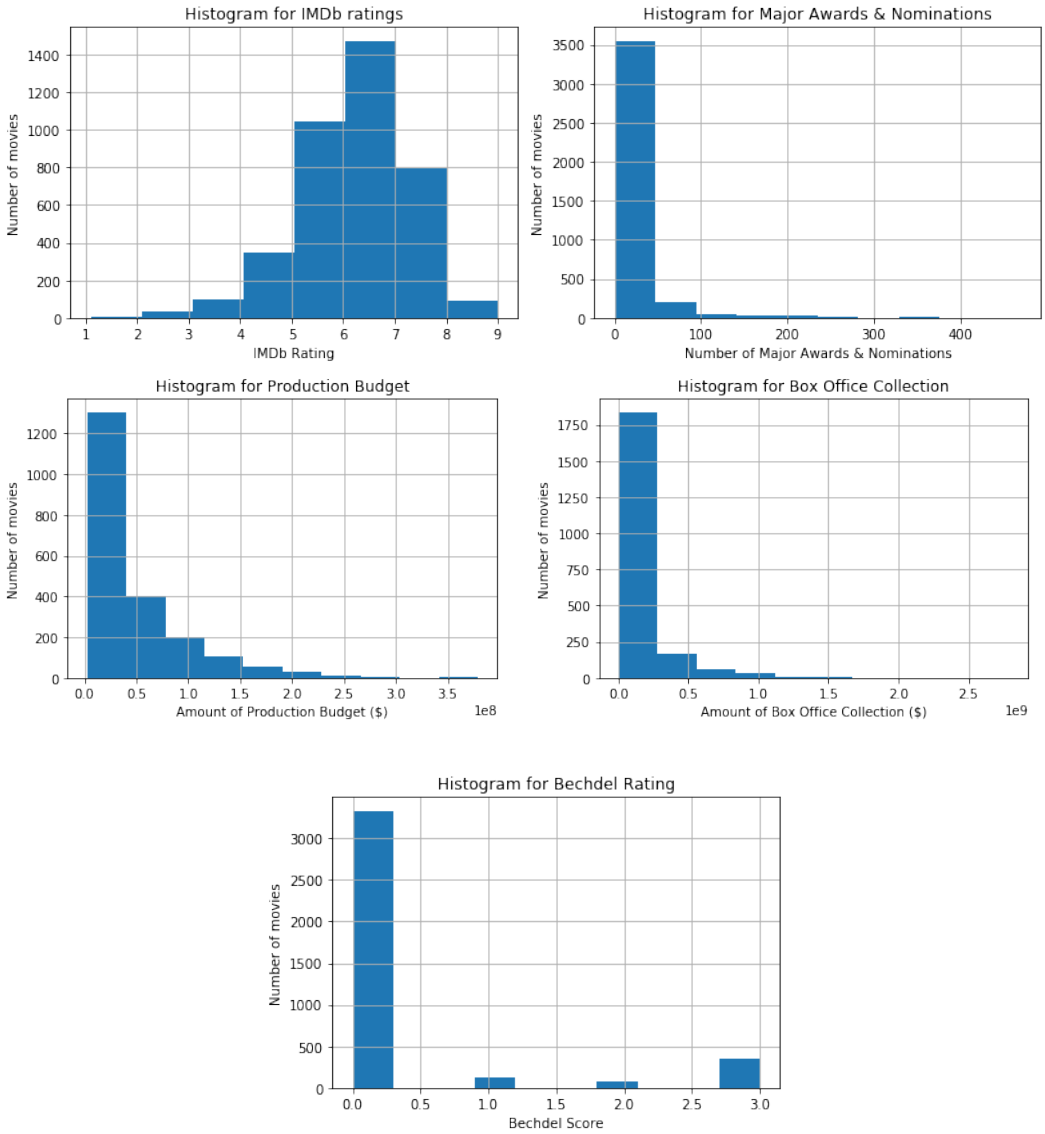


Figure 1: Histograms for some variables: IMDb ratings, awards & nominations, production budget, box office collection, and Bechdel rating.

Rank	Production Company	Number of Movies
1	Warner Bros Pictures	343
2	Sony	319
3	20^{th} Century Fox	242
4	Universal Pictures	240
5	Paramount	222
6	Lions Gate Films	205
7	New Line Cinema	106
8	Walt Disney Pictures	104
9	Metro-Goldwyn-Mayer	96
10	Miramax	92

Table 1: Top producers by volume of movies produced

3. Concentration of resources at the top

To empirically verify this phenomenon, three variables are used to measure the opportunity or resources available to each director: (1) number of movies directed, (2) total amount of production budget received for all movies directed, and (3) career continuity defined as the difference in year of release of the latest movie to the oldest movie directed. Through these three quantifying variables, it can be verified that there is indeed an accumulation of resources at the top [Figure 2]. A vast majority of directors, about 81%, direct one or two movies only. The top 10 percentile of directors has received 51.4% of the total production budget for all the movies in the dataset. The career continuity is not more than 5 years for 75% of the directors. Among the three sets of directors (top 20 percentile) with the highest opportunity as measured by the three variables, there is a significant overlap [Figure 3].

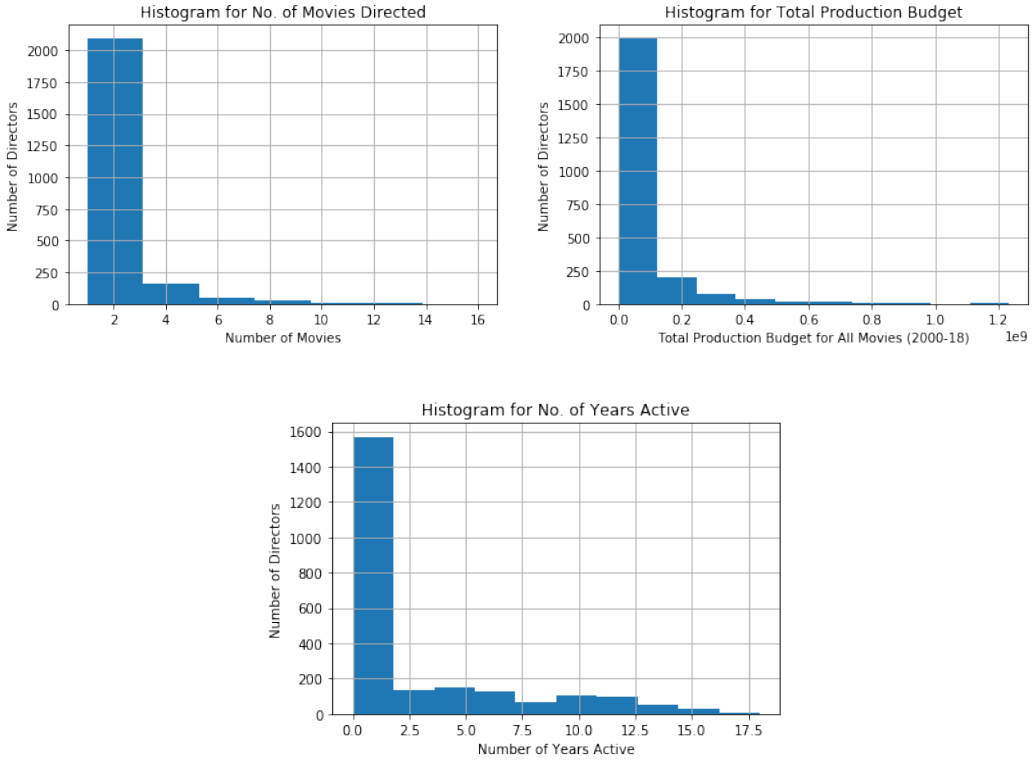


Figure 2: Histograms for some variables: no. of movies directed, total production budget, no. of years active

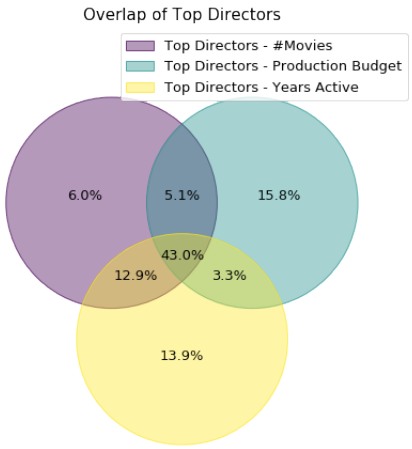


Figure 3: Venn diagram for top directors

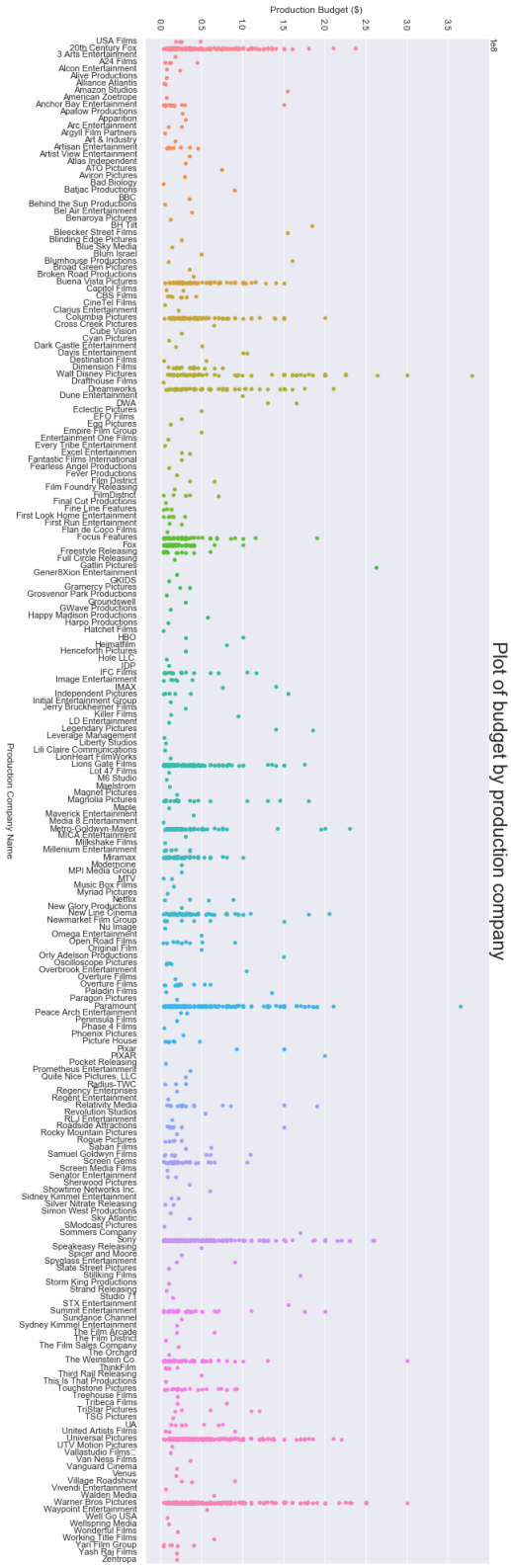


Figure 4: Different production budgets for each production company

4. Collective action of directors and producers

To search for an underlying pattern in the hundreds of links formed through complex interactions between directors and producers, blockmodel analysis is used. Blockmodel analysis has been carried out through the CONCOR (CONvergence of iterated CORrelations) algorithm implemented using the UCINET software [5]. This method groups directors and producers in the network based on structural equivalence [2, 6]. Two directors are considered to be similar or equivalent in their relationships with producers if they have worked for a similar set of producers. Essentially, blockmodeling permutes the rows and columns in a matrix representing the links between different agents in the network. The resulting matrix may be viewed as a collection of submatrices each with structurally equivalent links. The method, based on hierarchical clustering, finds correlation values between column vectors to construct a new matrix. These values quantify the similarities in links between different producers. Same procedure is applied to the newly found matrix iteratively till the matrix is entirely composed of either +1 or -1 values in all cells. Based on this, a bipartition is created in the matrix and more partitioning is done till the desired level of consistency within blocks is achieved. To summarize, the CONCOR algorithm takes a full matrix representing links as the input and returns information about the sub-matrices or 'blocks' each of which is structurally different from each other.

Since most of the initial matrix (one row per director, one column per producer) was sparse, performing some form of aggregation for rows/columns seemed reasonable

(large sparse matrices take more computation time). Also, in general, bigger and more prominent production companies take on larger quantities of movies [Table 1] and are also capable of producing multiple big-budget projects [Figure 4]. Since this paper wants to understand how directors with opportunites and powerful producers interact, aggregation was done based on the number of movies the director or the production company had undertaken.

Each row denoted by DX (where X is the number of movies directed) represents a single group of directors and each column is denoted by PY (where Y is the number of movies produced) represents a single group of producers. For example, the column P343 represents all the links of the group of producers who have produced 343 movies. In this particular dataset, Warner Bros is the only production company belonging to that group 1. Now consider the row D1. All the directors who have directed exactly one movie belong to this group amounting to a total of 1504 directors in all. The value in a cell gives the number of movies where the directors and producers belong to the given row and column groups respectively.

Figure 5 gives the permuted matrix for movie volume post blockmodeling. The blocks are separated by thick black lines. The directors are separated into two groups and the producers into four groups, so there are a total of $2 \times 4 = 8$ blocks. By simple observation, it is found that the first group of directors (i.e. D - I) comprises small directors with less than 7 movies and the second group (i.e. D - II) comprises prominent directors with movies between 8 and 16. For the producers, the

first two groups (i.e. P - I and P - II) contain small and medium-sized production companies, for example, Yari Film Group (P17), Relativity Media (P24) are both medium-sized, and National Geographic (in P3) (which although is a big brand name is a small player in Hollywood productions), Lucasfilm (in P3). The only exception in the group is Fox Searchlight (in P91) which is a large producer's group. The third producer's group (P - III), mostly contains mid and large companies. Finally, the fourth group (P - IV) exclusively contains large producers, like Dreamworks (in P77), Paramount (in P222), and Lions Gate Films (in P205).

Several patterns emerge from this analysis:

- *Presence of structural holes*: several structural holes can be observed from the permuted matrix. In Table 4, blocks that exhibit a relatively high amount of interaction (measured by the number of movies in this case) stand out and are assigned 1s (blocks 1,2,3,4,8). But some blocks are empty or are very sparse indicating low or no transactions between the associated groups, these are assigned 0s (blocks 5,6,7). This shows that top directors mostly don't form many links with medium and small-sized production companies - they work almost exclusively with large producers.

	P - I	P - II	P - III	P - IV
D - I	639	1114	773	1774
D - II	33	47	55	233

	P - I	P - II	P - III	P - IV
D - I	624	1099	741	1548
D - II	29	39	45	133

	P - I	P - II	P - III	P - IV
D - I	0.03	0.03	0.03	0.09
D - II	0.09	0.09	0.12	0.70

	P - I	P - II	P - III	P - IV
# Movies	0.05	0.04	0.07	0.13
# Links	0.05	0.04	0.06	0.09

Table 2: CONCOR results: (i) aggregate number of movies, (ii) aggregate number of unique links, (iii) Density matrix for the number of movies, (iv) ratio of values from row D - II by row D - I from (i) and (ii)

I	II	III	IV
V	VI	VII	VIII

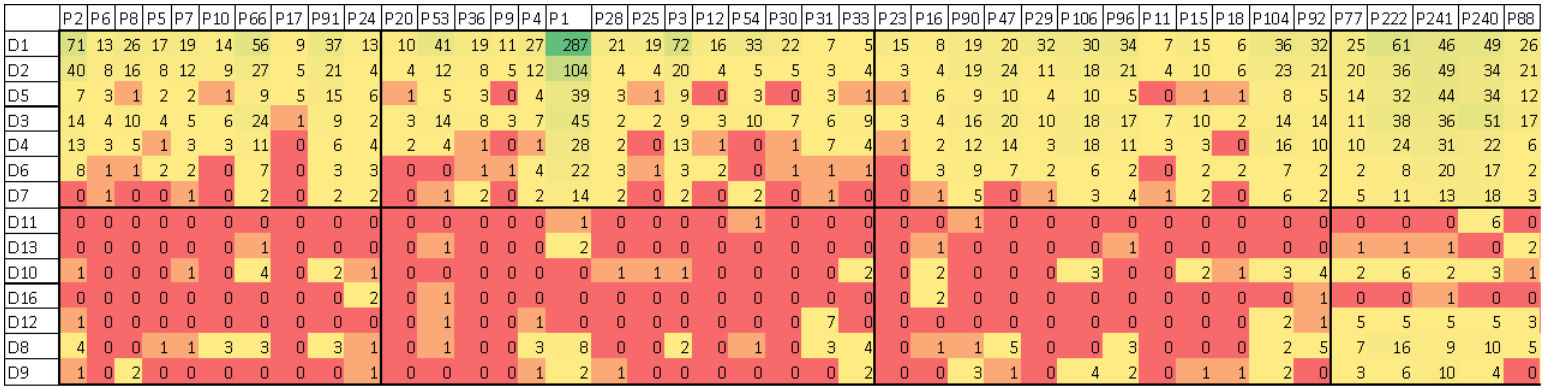
Table 3: Block numbering (for reference only)

1	1	1	1
0	0	0	1

Table 4: 0/1 binary representation of link presence

Group	Members
D - I	Directors - 1,2,3,4,5,6,7
D - II	Directors - 8,9,10,11,12,13,16
P - I	Producers - 2,5,6,7,8,10,17,24,66,91
P - II	Producers - 1,3,4,9,12,20,25,28,30,31,33,36,53,54
P - III	Producers - 11,15,16,18,23,29,47,90,92,96,104,106
P - IV	Producers - 77,86,88,205,222,240,241,319,343

Table 5: Blockmodel clusters for directors and producers



	P2	P6	P8	P5	P7	P10	P66	P17	P91	P24	P20	P53	P36	P9	P4	P1	P28	P25	P3	P12	P54	P30	P31	P33	P23	P16	P90	P47	P29	P106	P96	P11	P15	P18	P104	P92	P77	P222	P241	P240	P88
D1	71	13	26	17	19	14	56	9	37	13	10	41	19	11	27	287	21	19	72	16	33	22	7	5	15	8	19	20	32	30	34	7	15	6	36	32	25	61	46	49	26
D2	40	8	16	8	12	9	27	5	21	4	4	12	8	5	12	104	4	4	20	4	5	5	3	4	3	4	19	24	11	18	21	4	10	6	23	21	20	36	49	34	21
D5	7	3	1	2	2	1	9	5	15	6	1	5	3	0	4	39	3	1	9	0	3	0	3	1	1	6	9	10	4	10	5	0	1	1	8	5	14	32	44	34	12
D3	14	4	10	4	5	6	24	1	9	2	3	14	8	3	7	45	2	2	9	3	10	7	6	9	3	4	16	20	10	18	17	7	10	2	14	14	11	38	36	51	17
D4	13	3	5	1	3	3	11	0	6	4	2	4	1	0	1	28	2	0	13	1	0	1	7	4	1	2	12	14	3	18	11	3	3	0	16	10	10	24	31	22	6
D6	8	1	1	2	2	0	7	0	3	3	0	0	1	1	4	22	3	1	3	2	0	1	1	1	0	3	9	7	2	6	2	0	2	2	7	2	2	8	20	17	2
D7	0	1	0	0	1	0	2	0	2	2	0	1	2	0	2	14	2	0	2	0	2	0	1	0	0	1	5	0	1	3	4	1	2	0	6	2	5	11	13	18	3
D11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	6	0
D13	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0	2
D10	1	0	0	0	1	0	4	0	2	1	0	0	0	0	0	0	1	1	1	0	0	0	0	2	0	2	0	0	0	3	0	0	2	1	3	4	2	6	2	3	1
D16	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
D12	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	2	1	5	5	5	5	3
D8	4	0	0	1	1	3	3	0	3	1	0	1	0	0	3	8	0	0	2	0	1	0	3	4	0	1	1	5	0	0	3	0	0	0	2	5	7	16	9	10	5
D9	1	0	2	0	0	0	0	0	0	1	0	0	0	0	1	2	1	0	0	0	0	0	0	2	0	0	3	1	0	4	2	0	1	1	2	0	3	6	10	4	0

Figure 5: Permuted matrix for the number of movies developed together

- *Large number of novice directors and producers*: Since D1 and P1 represent producers and directors who have worked on exactly 1 movie each, the number of movies in their respective cells also represent the number of directors and producers in the group. From the table, the row and column corresponding to D1 and P1 are densely populated. Therefore, this is a market that is populated at the bottom with many inexperienced directors and producers. This number dwindles near the top signaling that only a few make a successful climb to the top, and supports the argument from section 3 that Hollywood is dominated by groups of elites.

- *Elite producers and directors prefer each other*: since the number of big directors is very small, the density between the two director groups is compared instead of their volumes (i.e. no. of movies and links). Density is measured by the no. of actual links in the group by the no. of possible links, the no. of possible links is the no. of links if all directors and producers were connected to each other (results in Table 2). All producer groups have denser connections with top directors are compared to others. Also ratio of movies with large directors w.r.t small directors is increasing with the size of the producers (Table 2-(iv)). So top directors and producers prefer each other.

- *Exploitation vs. exploration tradeoff for producers*: large directors are exclusive to large producers, but producers while preferring top directors, also nurture fresh talent (2 (i) and (ii)) - blocks 1,2,3). This is akin to the exploration

vs. exploitation paradigm discussed in numerous social science literature. Exploration is a high-risk high-reward case involving actively seeking out new information, while exploitation is a low-risk low-reward case taking advantage of existing information [15, 19]. Because it is difficult to signal artistic talents, producers tend to trust directors with good track-records. But they are also on the lookout for fresh talent because novice directors cost less, and they keep up the leadership pipeline. So producers are generally involved with both groups (albeit with different intensities) to balance out the costs and risks for exploration and exploitation.

- *Affinity towards diversification*: most of the ties between directors and producers are one-shot, and directors and producers believe in the strength of diversity. Only some 7% of the links are recurrent. 5.3%, 1.1%, 0.5% of the director-producer pairs have worked on 2, 3, and more than 3 projects respectively.

5. Connections vs. Merit in Predicting Careers

This part of this study employs various machine learning models, which can pick up more complex signals, to understand which of the two - professional connections or individual talent plays a more prominent role in a director's Hollywood career. First, measures are defined for connections, merits, and opportunities they are likely to influence (Section 5.2). Then, by employing some machine learning models, we can find out which of the two - connections or merit - variables hold more power in

accurately predicting the opportunities a director gets.

5.1 Network Representation

Hollywood's professional relationship network between co-artists can be represented as a simple bipartite graph [Figure 6]. This graph has 2 sets of nodes - one set of nodes are the movies, the other set of nodes are people i.e. the actors, directors, and writers. Links are formed between a movie and a person when that person works for the movie as a writer/actor/director. Therefore all the links in the graphs are between movies and people and no links exist between 2 people or 2 two movies. A person may be linked to one movie as an actor but to another as a director. Sometimes, a person takes on multiple roles, for example as both a writer and a director, for a movie. In such cases, multiple links exist between the person and the movie. There are a total of 11,673 edges and 18,566 nodes in this graph.

In this vast network, artists form professional relationships with each other when they have worked on a movie together (i.e. they are 'co-workers'). Consequently, different artists are also indirectly connected to each other through their 'co-workers' (and their 'co-workers' and so on). This leads to the formation of an interconnected network of artists. This second network of artist-artist connections comprises entirely of writers, directors, and actors. Artists derive professional reputation by associating with other important or 'central' artists in the second network and by associating with critically or commercially successful movies in the first.

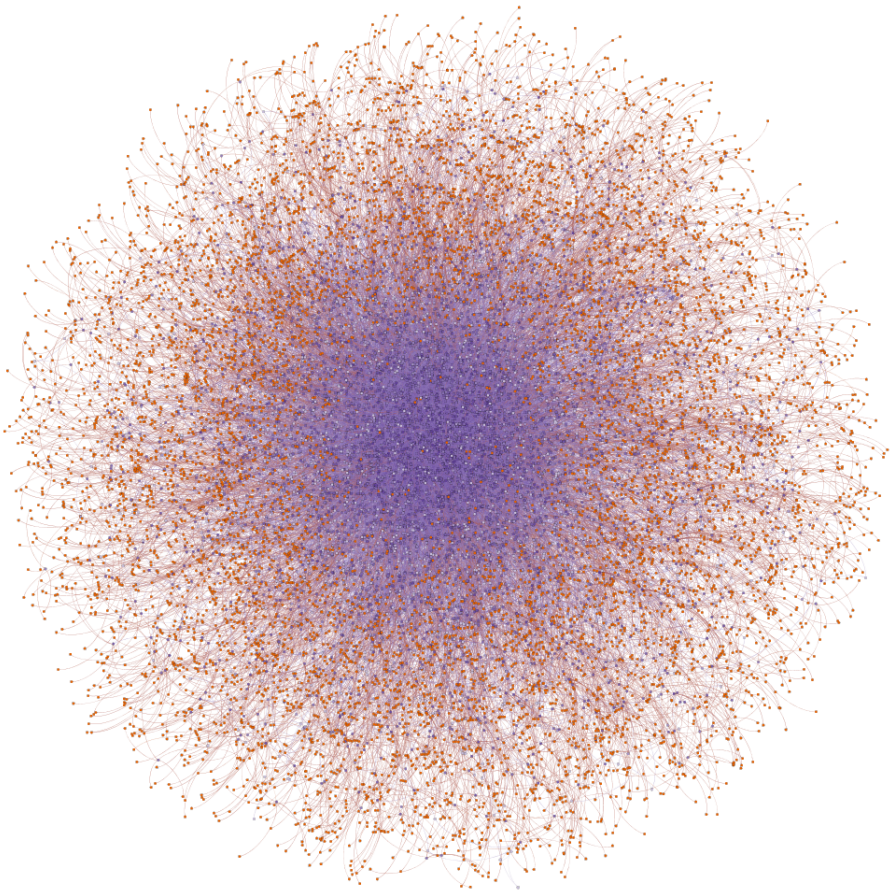


Figure 6: Network of movies (blue nodes) and artists (red nodes)

5.2 Feature Engineering

The first set of variables is the 'connection' variables which indicate the reputation of the artist based on her position in the network with other artists. This is mainly derived from the artist-artist network. Several network coefficients calculated using various graph algorithms (from AWS Spark GraphX [29] and NetworkX [16]) are used for this, these are:

- Degree centrality: this measure gives the importance of a node in the network based on the number of links it has (sum is taken). This measure treats all links to be of equal importance.

- Eigenvector centrality: for this measure, all nodes are given a relative score but unlike the degree centrality, connections to high scoring nodes are valued more than those to low scoring nodes. So not all links are treated equally.

- Closeness centrality: this measure values nodes based on its closeness to other nodes, the closer a node is to other nodes, the more its score. It is calculated as the inverse of the average distance to all the other nodes in the graph. Intuitively, if a node is closer to all the other nodes, the more efficient it is in the transmission of information.

- Betweenness centrality: this is a measure for how much the node is in the path between other nodes. It is computed as the ratio of paths that break if the node is deleted by the total possible path between all other pairs of nodes. This is based on the concept that if a node is important in connecting many other nodes, the individual in the node becomes a kind of 'gatekeeper' and can control the communication channels of other nodes.

- Eigenvector centrality of the production company: this variable measure the prominence of the production companies a director has worked with.

The second set of predictor variables are the 'merit' variables. These variables are

proxies for the individual talent of the director and are mostly derived from her repertoire of movie projects, i.e. the artist-movie network: (1) average IMDb rating, (2) average box office collection, (3) average no. of awards and nominations, (4) sum of no. of oscar wins, and (5) average Bechdel score.

Finally, the third set is the 'opportunity' variables that are the target of the prediction problem. These are calculated from the artist-movie network. These quantify the professional opportunities presented to a director: (1) no. of movies, (2) years active (measured by as the year of release of the latest minus the oldest movie), (3) total production budget raised. (1) and (2) measures the work volume and (3) measures the extent of financial support for the projects.

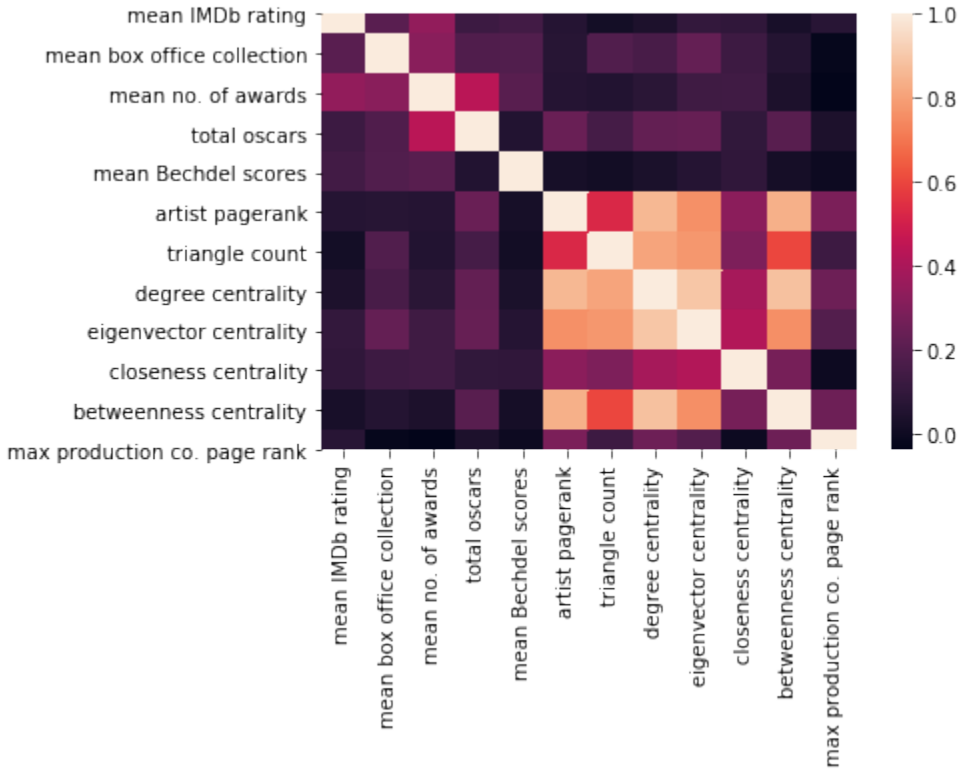


Figure 7: Correlation for merit and connection variables

	mean	std	min	25%	50%	75%	max
years active	3.2	4.1	1	1	1	4	19
total no of movies	2.3	2.9	1	1	1	2	45
total production budget (mil$)	95.9	192.3	1.9	15	38.2	85	4291.1
mean IMDb rating	6.1	1.1	1.1	5.6	6.2	6.8	8.9
mean box office collection (mil$)	95.7	136.3	0	16	68.2	103.8	2048.4
mean no of awards	13.8	32	0	1	4	12	470
total no of oscars	0.1	0.9	0	0	0	0	40
mean bechdel score	0.3	0.8	0	0	0	0	3
artist pagerank	1.05	0.27	0.85	0.92	0.95	1.05	5.66
triangle count	43.7	74.4	0	10	21	41	1109
degree centrality ($\times 10^{-4}$)	9.72	11.4	0.78	3.88	5.43	0.1	140
eigenvector centrality ($\times 10^{-3}$)	4.17	7.76	0	0.39	1.58	4.22	116
closeness centrality ($\times 10^{-1}$)	2.18	0.62	0	2.11	2.31	2.47	3.24
betweenness centrality ($\times 10^{-4}$)	2.11	7.20	0	0	0	0.69	159
max production co. page rank ($\times 10^{-1}$)	6.98	2.10	4.35	5.82	5.93	8.01	20.8

Table 6: Descriptive statistics for all variables

5.3 Methods

Both the feature variables and the target variables are continuous numbers. Based on the available data and its type, three different models are shortlisted for fitting the data: (1) linear regression, (2) tree-based methods, and (3) neural network. The three models are implemented using the scipy package in python [28]. First, a multiple linear regression model is fit to ensure that the features chosen indeed are significant in predicting the target variable. Let Y_i and X_is be the data points, βs are the coefficients, and ϵ is the model fitting error. OLS (Ordinary Least Squares) regression, minimizes the Residual Sum Square, $RSS = \sum_i \epsilon_i^2$ from the following equation to estimate the βs:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_n X_{ni} + \epsilon_i$$

From the correlation heatmap (Figure 7), it can be seen that many 'connections' variables are correlated with each other. Primarily, this is because they measure the same kind of reputation. Therefore, some of these variables have to be dropped to make the model more 'interpretable' [17]. Also, linear regression uses every feature that is fed into it. This can cause potential over-fitting which may lead the model to perform poorly in an out-of-sample dataset. Therefore, a variation of regression called the LASSO regression is used. LASSO regression filters out the most influential variables by imposing an L_1 norm regularization penalty based on the coefficients to the RSS error:

$$\sum_i \epsilon_i + \lambda \sum_{j=1}^{n} \beta_j \tag{1}$$

From Table 6, it is clear that different variables are of entirely different magnitudes. Therefore, for a variable of small order (say one of the centrality measures), the coefficient will be very large. Here, LASSO will unfairly penalize the larger coefficients more than the smaller once. Therefore before running the regression, all data is normalized with a mean of 0 and a variance of 1.

Once the significance of the variables is established, the focus is shifted to prediction. One method that is suitable for this data is the tree-based method, random forest. Random forest is an ensemble method, it constructs several random, independent,

and shallow (i.e. splitting only on small no. of variables) decision trees. A decision tree is a set of decisions, where a single decision based on a single variable is tested at every level. Once all the tests are administered to a given data point, the model classifies the data into a 'bucket' and each bucket is assigned a particular prediction value. For a random forest, the final prediction is taken as a weighted average of the predictions of the individual trees [20]. The second prediction method used is artificial neural networks (ANN). The ANN model is inspired by the working of neurons in animal brains. It is an ensemble of a non-linear transformation of a linear function of input features [22, 17]. ANNs are notorious for being 'black box' models meaning that while it can be used to approximate any functional form but studying a trained neural network cannot give any information about the underlying true functional form. This makes a model's interpretation very hard and the relative significance of each given variable cannot be obtained. But this drawback is inconsequential in this study because the focus is entirely on prediction.

For each of the three types of methods, i.e. regression, random forest, and ANN, two models each are tuned one using only the 'connection' and the other using only the 'merit' features. Whichever of the two sets of variables has the highest power in terms of predicting the 'opportunities' for a director, is concluded to be the more important aspect in making a director's Hollywood career.

5.4 Results

Table 7 summarizes the results from the LASSO regression. Since the variables were

first standardized before running the regression, the magnitude of the coefficients does not hold any direct value. However, one main observation is that all coefficients have non-negative values, this means that none of the features calculated adversely affects opportunities. Another observation is that only a few variables have non-zero values after running LASSO regression. This is because all the relatively insignificant coefficients are set to 0 by the LASSO model. The value of lambda that determines this cutoff (as given in equation (1)) is determined by doing a grid search to find the lambda value that gives the least RMSE error for 5-fold cross-validation [13]. (Check Figures 8 to 10 in Appendix to see how the coefficient magnitudes change with lambda).

	no. of years active	no. of movies	total production budget
mean IMDb rating	0	0	0
mean box office collection	0	1	671
mean no of awards	0	0	0
total no of oscars	0	0	0
mean bechdel score	0	0	0
artist pagerank	2	39	565
triangle count	0	1	0
degree centrality	27	8	412
eigenvector centrality	0	0	1068
closeness centrality	2	0	0
betweenness centrality	0	0	0
max production co. page rank	1	0	0

Table 7: Descriptive statistics for all variables

For no. of years active, the significant variables are (1) artist pagerank, (2) degree centrality, (3) closeness centrality, and (4) max production co. page rank. All four of these are 'connection' variables. For no. of movies, important variables

are (1) mean box office collection, (2) artist pagerank, (3) triangle count, and (4) degree centrality. Three of these four variables are 'connection' variables. For the total production budget, variables used are (1) mean box office collection, (2) artist pagerank, (3) degree centrality, (4) eigenvector centrality. Here again, three of these four variables are 'connection' variables. Therefore in all three cases, the majority of the top-4 significant variables are the ones representing the strength of the director's connections.

S.No.	Model	RMSE for no. of years active		
		Merit	Connections	Connections & Merit
1	LASSO	16.83	6.97	6.90
2	Random Forest	7.3	3.08	3.03
3	Neural Network Classifier	9.8	0.71	3.23
S.No.	Model	RMSE for no. of movies		
		Merit	Connections	Connections & Merit
1	LASSO	7.76	0.26	0.26
2	Random Forest	3.54	0.30	0.43
3	Neural Network Regressor	4.77	0.18	0.21
S.No.	Model	RMSE for total production budget		
		Merit	Connections	Connections & Merit
1	LASSO	30606	14603	11971
2	Random Forest	15787	12229	7351
3	Neural Network Regressor	20829	11844	5293

Table 8: Cross-validation RMSE for various models for outcome variables (i) no. of years active, (ii) no. of movies, (iii) total production budget

Each of the three models (LASSO, random forest, neural network) has their hyper-parameters tuned by performing a randomized grid search over a range of possible values. The model with the lowest cross-validation (5-fold) RMSE is chosen in each case (Table 8 gives the optimal lowest RSME values). For every opportunity outcome variable and every machine learning model, connections give a better prediction

accuracy (i.e. lesser RMSE score) as compared to the merit variables. This implies that connections predict a directors success more accurately than her merit would.

6. Discussion and Limitations

The study sheds light on both the existing market conditions for directors in terms of opportunities and investigates the influence of connections on a director's success. Only a few directors enjoy most resources and opportunities in Hollywood. These elite directors exhibit similarities in their pattern of interactions with other artists and producers. Elite directors and prominent producers show strong affinity towards each other. Also, a director's position in a network of other artists and directors is a stronger predictor of her rise in the industry as compared to her actual creative talent and potential.

There are a few limitations to this study. First, a link is assumed to exist between two agents only if they have worked on any movies together. While this helps chart the professional connections of an artist, many informal connections formed through daily interactions and such are omitted from the data. These informal connections could also play a role in the director's career outcomes. Second, only the subset of major movies made in the last 20 years have been used in the study. A more exhaustive compilation of not just all the movies but also similar contents like TV shows and documentaries could give more robust results. Finally, there is a chicken-or-egg causality problem. On one hand, connections give directors more chances to

flourish, but at the same time, a successful or popular director naturally will have a stronger network of colleagues. This is similar to the Matthew effect [24] where established individuals attract disproportionate acclaim and visibility for their future projects. This study assumes that one (i.e. merit and connections) predictS the other (opportunities) but more complex models could be looked at to understand this 'feedback' effect.

References

[1] Moshe Adler. Stardom and talent. *The American economic review*, 75(1):208–212, 1985.

[2] Phipps Arabie, Scott A Boorman, and Paul R Levitt. Constructing blockmodels: How and why. *Journal of mathematical psychology*, 17(1):21–63, 1978.

[3] Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.

[4] Howard S Becker. *Art worlds: updated and expanded*. University of California Press, 2008.

[5] Stephen P Borgatti, Martin G Everett, and Linton C Freeman. Ucinet for windows: Software for social network analysis. *Harvard, MA: analytic technologies*, 6, 2002.

[6] Ronald L Breiger, Scott A Boorman, and Phipps Arabie. An algorithm for blocking relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of mathematical psychology*, 12:328–383, 1975.

[7] Gino Cattani and Simone Ferriani. A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the hollywood film industry. *Organization science*, 19(6):824–844, 2008.

[8] Richard E Caves. *Creative industries: Contracts between art and commerce.* Number 20. Harvard University Press, 2000.

[9] B Daniels, D Leedy, and SD Sills. Movie money: Understanding hollywood's (creative) accounting practices. *CPA Journal*, 69(2):74–74, 1999.

[10] Kimberly D Elsbach and Roderick M Kramer. Assessing creativity in hollywood pitch meetings: Evidence for a dual-process model of creativity judgments. *Academy of Management journal*, 46(3):283–301, 2003.

[11] Robert R Faulkner. *Music on demand.* Transaction Publishers, 1983.

[12] Robert H Frank and Philip J Cook. *The winner-take-all society: how more and more Americans compete...* Free Press, 1995.

[13] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, 2019.

[14] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[15] Anil K Gupta, Ken G Smith, and Christina E Shalley. The interplay between exploration and exploitation. *Academy of management journal*, 49(4):693–706, 2006.

[16] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[18] Candace Jones, William S Hesterly, and Stephen P Borgatti. A general theory of network governance: Exchange conditions and social mechanisms. *Academy of management review*, 22(4):911–945, 1997.

[19] David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative science quarterly*, 52(4):667–694, 2007.

[20] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[21] Mark Lutter. Creative success and network embeddedness: Explaining critical recognition of film directors in hollywood, 1900–2010. 2014.

[22] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[23] Pierre-Michel Menger. Artistic labor markets and careers. *Annual review of sociology*, 25(1):541–574, 1999.

[24] Robert K Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

[25] Patricia F Phalen, Thomas B Ksiazek, and Jacob B Garber. Who you know in hollywood: A network analysis of television writers. *Journal of Broadcasting & Electronic Media*, 60(1):160–170, 2016.

[26] Sherwin Rosen. The economics of superstars. *The American economic review*, 71(5):845–858, 1981.

[27] Paul F Skilton. Similarity, familiarity and access to elite work in hollywood: Employer and employee characteristics in breakthrough employment. *Human Relations*, 61(12):1743–1773, 2008.

[28] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[29] Reynold S Xin, Joseph E Gonzalez, Michael J Franklin, and Ion Stoica. Graphx: A resilient distributed graph system on spark. In *First international workshop on graph data management experiences and systems*, pages 1–6, 2013.

8. Appendix

8.1 More details on data collection and pre-processing

The first step in data collection was to compile the list of movies to be scraped. This list was procured by performing web scraping on the list of prominent movies from 1972-2018 on IMDb (hyperlink: https://www.imdb.com/list/ls057823854/). Along with the title and the year, a unique ID assigned by IMDb to the movie was also scrapped. This ID has a prefix 'tt' followed by seven numerical digits. Next, this list was filtered to include only the data from 2000 onwards. Now, these IDs were used to call the OMDb API to retrieve all information about the specific movies. The call to this API is embarrassingly parallel. Therefore, this section of the code was parallelized on 5 cores using mpi4py package on python. Subsequently, there was a 38.7x speedup in running the code.

Now some cleaning and pre-processing were performed on the obtained dataset to make it suitable for model-building. All the movies not released in the US were removed from the list (~5%). Any item in the list that was a television series or a documentary was also removed. Further obscure movies with very little information about them (i.e. empty values for rating, actor/director/writer list) were also removed (~1%). Several columns that were very sparsely populated were removed (example: information from rotten tomatoes like rating, tomato meter score, reviews as well as others like DVD and website information for the movie). Further, any

missing or incomplete information was replaced with a float value 'NaN' (python equivalent for a none value) and all numerical values were converted to numbers (years, runtime, rating, etc.).

Some additional variables were added to this table. Using the same ID, the movies' Bechdel scores were procured from the Bechdel test API. A separate table was created with the names of all directors, writers, and actors with their gender. This gender was first guessed using the genderguesser package in python. Then for ambiguous gender names, the list was filled out manually by extensive searching on Google. Unfortunately, only two genders - female and male were supported by the package, and only those are used in this study. Some box office information was downloaded from numbers.com. This data was merged with the movie database by matching both the titles and year of release. Unfortunately, in this data, some 30% of the financial information (production budget and box office collection) and 45% of the production company data was missing. Some of the financial information could be scraped from google.com's search results. The remaining were filled out mostly manually by searching on google. In the end, about 4% of the movies still lacked financial information. This was filled by extrapolation (i.e. taking median and mean) of the existing data with the same genre, production company, and year of release.

8.2 Regularization path plots for LASSO regression coefficients

Coefficient magnitudes vs lambda for the various target variables

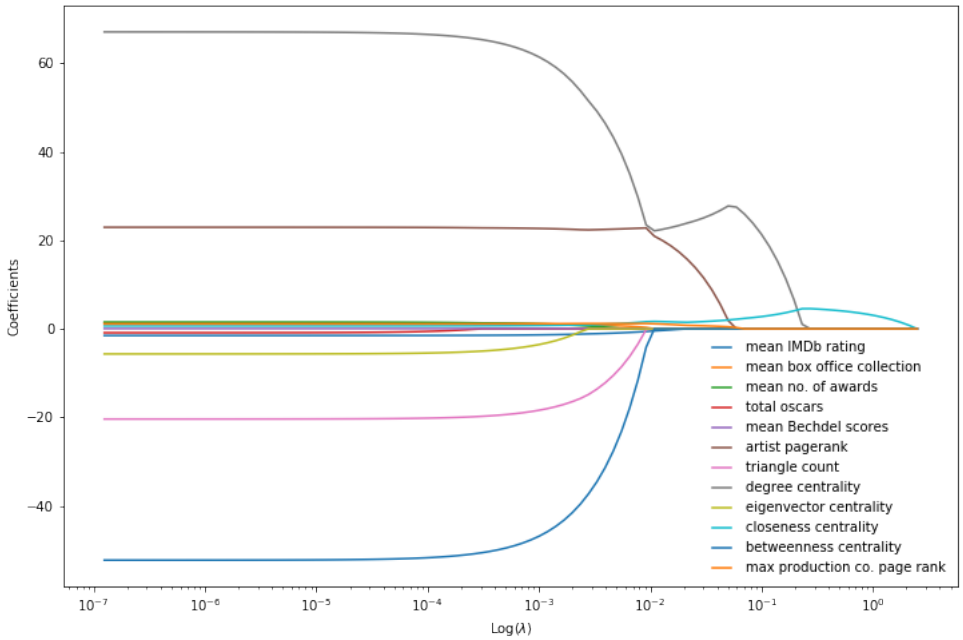


Figure 8: Target variable = no. of years active

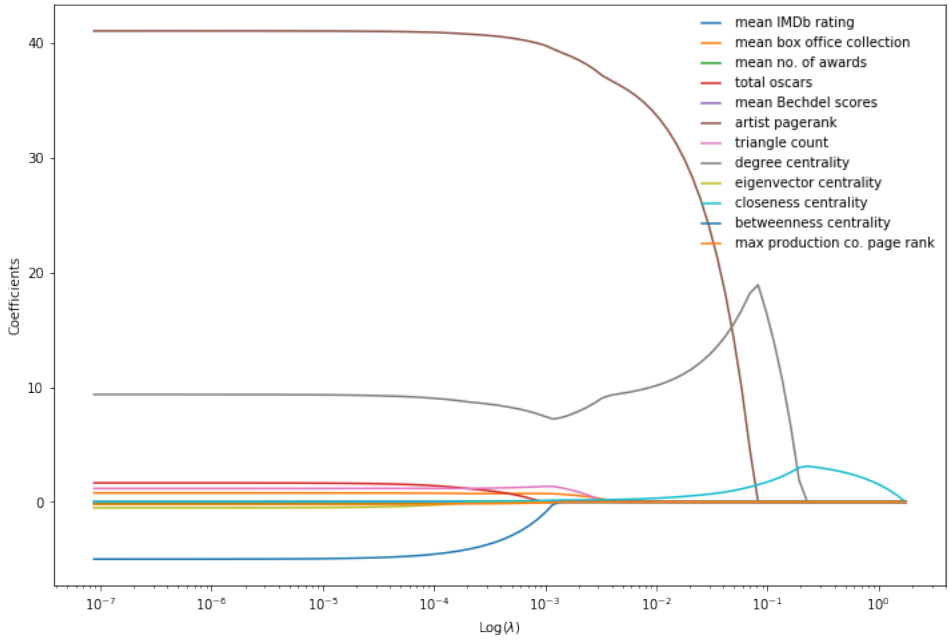


Figure 9: Target variable = no. of movies

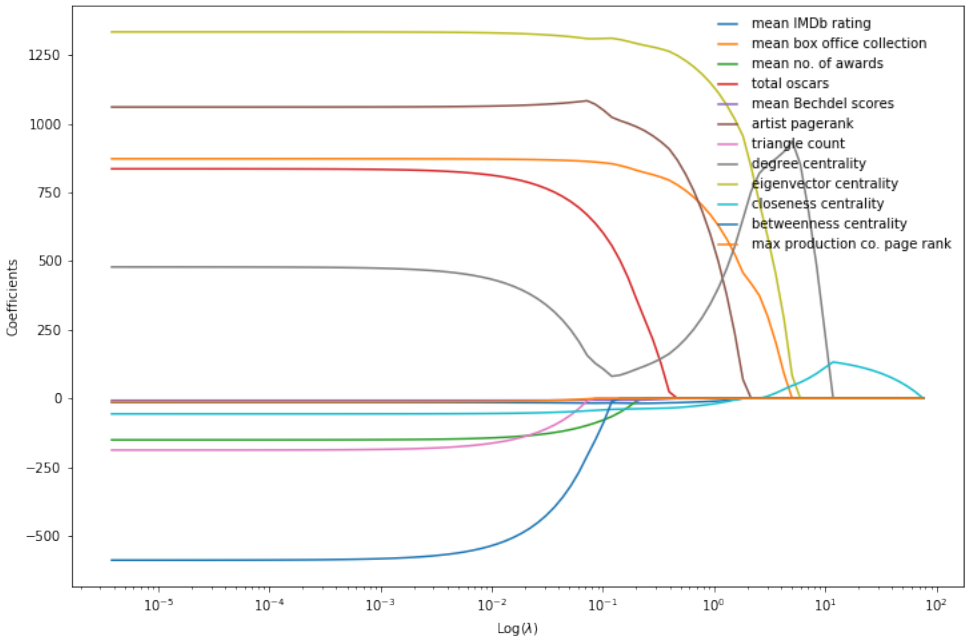


Figure 10: Target variable = total production budget