Diversity in Hollywood: From Directors to Movies to Actors

CATHERINE DONG
cdong@stanford.edu
and
NICK TROCCOLI
troccoli@stanford.edu
and
JESSICA ZHAO
jesszhao@stanford.edu

1. INTRODUCTION

Underrepresentation of minorities and women in movies has been a persistent issue in Hollywood. Studies of diversity in the film industry have provided high-level statistics aggregated over many movies and actors. We hope to be able to identify more specific patterns of casting practices and their effect on movie success by analyzing the co-working relationships of key players – actors, directors, and movies themselves – in the industry. By examining these relationships, we strive to: 1) Understand the presence and absence of diversity in Hollywood films as related to directors, casts, and revenue, over time; and 2) Uncover gender- and race-based assortativity patterns in the actor-actor and actor-director coworking networks.

2. PRIOR WORK

As motivation for our analysis of this subject, look no further than the current ethnic and gender diversity in film, particularly compared to films viewers. Out of the 11,306 speaking or named characters assessed, only 28.3% were from underrepresented racial/ethnic groups [1]. In contrast, 37.9% of the U.S. population and 45% of movie ticket buyers come from these minority groups. Clearly, the movie industry has failed to reflect the demography of their audiences. Behind the camera, the picture is even more bleak. Only 13% of directors are minorities. This statistic is especially important because Annenberg also found the percentage of minority actors on screen increases from 26.2% to 42.7% when films have a minority director versus a white director. This highlights the importance of our intention to analyze film diversity from a variety of angles, from the director to the cast to the time period. Women have also been historically underrepresented in Hollywood. According to Inequality in 800 Popular Films, only about a third of speaking or named characters in the films studied were women [2]. Moreover, a mere 3.4% of directors were female.

3. DATA

With this clear motivation to examine diversity statistics further, we searched for as comprehensive of a film dataset as we could find. However, due to the depth of data required, including race, gender, film performance, etc., we had trouble finding one dataset to use. Our initial dataset was from Kaggle [4], which contained information about 4919 movies including box office revenue, top actors, director, and IMDB rating. It did not, however, include information such as race or gender of the included actors, or more than the top three actors in each movie. To fix this, we used other datasets to fill in the gaps. We cross-referenced the Kaggle dataset with a scraper of the Notable Names Database website, which provided race/gender information for 43% of actors and directors in the dataset. The unintended consequence of this, however, was a long fetch time for our scraper to gather data for over 10,000 actors. For the actors NNDB could not find information about, we used SexMachine, a Python module that predicts genders based on first name. Relying only on high-certainty predictions, we were able to fill in many of the remaining unknown genders. Finally, we scraped IMDB to pull additional actors for every movie, almost doubling the number that the Kaggle dataset provided. With information pulled from Kaggle, NNDB, IMDB, and SexMachine, we have a sizable dataset with which to run our analyses.

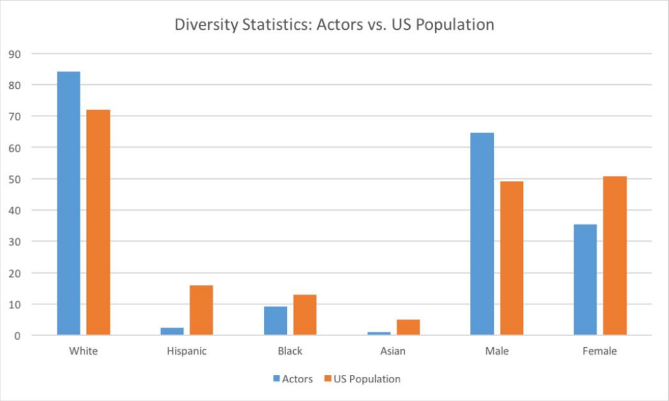

Fig. 1. Percentages of Hollywood actors represented in movies as compared to respective percentages of the US population - breakdown by race and gender.

3.1 Preliminary Data Analysis

3.1.1 *Actors.* We examine by race and gender how actors have been represented in movies as compared against the race and gender distributions of the United States (numbers from the 2010 census) [6]. In Figure 1, we see the percentages of actors against the percentages of the US population by race and gender. We observe that the categories of 'white' and 'male' are disproportionately overrepresented in Hollywood, while every single other category ('hispanic', 'black', 'asian' for races, and 'female' for genders) is disproportionately underrepresnted. This is an unfortunate but expected finding that is consistent with

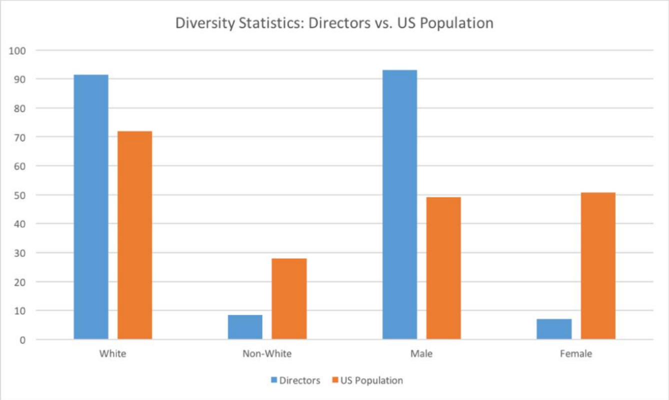

Fig. 2. Percentages of Hollywood directors as compared to respective percentages of the US population - breakdown by race and gender.

the current theme of Hollywood's bias towards casting white males.

3.1.2 *Directors.* Additionally, we examine by race and gender how directors are represented in Hollywood as compared against the race and gender distributions of the United States. In Figure 2, we see the percentages of directors against the percentages of the US population by race and gender. Unsurprisingly, we find a similar trend as for actors in Hollywood - directors who are 'white' and 'male' are disproprotionately overrepresented, while 'non-white' and 'female' directors are overwhelmingly underrepresented. This is also consistent with Hollywood's favorable bias towards white males. Moreover, the homogeneity among the directors likely trickles down towards the casting of actors - a less diverse body making the executive decisions for a film can result in a less diverse cast, as seen in the disparity between bars in Figure 1.

3.2 Network Structure

Once we gathered our data, we represented it as a directed tripartite graph where a node can either be a movie, an actor, or a director. We did this in order to have race/gender data and relations between a variety of different components of the movie industry, which we can break down further as needed in our analyses. Specifically, we are using NetworkX to create a graph where edges go from directors to the movies they directed, and from movies to the actors in those movies. NetworkX also allows us to add metadata to each node such as information about each person and movie. With this approach, we have created a graph with 3,711 movie nodes, 4,204 actor nodes, 1,813 director nodes, and 19,495 edges.

Creating this graph was not without some complications, however. Our first realization was that some people are both actors and directors, sometimes in the same movie. So we remodeled our graph from having separate nodes for actors and directors to having one node for each person connected to all movies theyre related to. Second, we decided midway through our project to change from an undirected to a directed graph. We did this to both make it easier to tell what relationship an edge represents (based on its start and end nodes) but also to give us more nuanced in-and-out degree information for nodes in our graph.

4. NETWORK ANALYSIS METHODS

4.1 Diversity Score

It is difficult to quantify diversity, but for the purposes of our analysis, we create a metric to allow us to analyze the diversity of directors and movies in Hollywood based on each entitys associated actors. We consider racial and gender diversity separately, and create measure for each as follows:

4.1.1 *Diversity of Movies.* We consider a movie to be ”diverse” if the top-billed speaking cast is composed of ”diverse” actors. In our model, actors are racially ”diverse” if they are non-white. Actors are gender ”diverse” if they are non-male. We adopt these measures because Hollywood has historically been dominated by white male figures, and seek to understand how this has changed or not changed over time.

By Race: Each movie is given a racial diversity score, which is calculated as the fraction of its racially diverse top-billed actors vs. overall cast size. For example, if a movie has three white actors and one non-white actor in its top-billed cast, its racial diversity score is 0.25.

By Gender: Each movie is also given a gender diverse score, which is calculated as the fraction of its gender diverse top-billed actors vs. overall cast size. For example, if a movie has two male actors and two female actors in its top-billed cast, its gender diversity score is 0.5.

4.1.2 *Diversity of Directors.* We consider a director to be ”diverse” if his or her collective movie casts are composed of ”diverse” actors.

By Race: Each director is given a racial diversity score, which is calculated as the average of his or her collective movies racial diversity scores. For example, if a director has directed Movie A, Movie B, and Movie C with respective racial diversity scores of 0.5, 0.25, and 0.0, the directors racial diversity score is 0.25.

By Gender: Each director is also given a gender diversity score, which is calculated as the average of his or her collective movies gender diversity scores. For example, if a director has directed Movie A, Movie B, and Movie C with respective gender diversity scores of 0.1, 0.3, and 0.5, the directors gender diversity score is 0.3.

4.1.3 *Evaluation of the Metric.* We recognize that our metric only provides a naive measure of diversity. It is consistent in that all scores are based on actor diversities – movie scores are calculated based on a movies cast of actors, and director scores are subsequently calculated as an average of his or her collective movie scores. However, there are other potential ways of calculating director diversity. For example, we look at the set of unique actors a director has worked with and compute the proportion of minorities in this set. This would avoid the double-counting of actors with whom the director has worked with multiple times. Our current implementation of director diversity score does double-count actors a director has worked with multiple times, but we reasoned that this is acceptable because directors who work with minority actors on a recurring basis should be awarded a higher diversity score.

4.2 Null Models

In order to interpret our diversity scores and other metrics that we will calculate, we need a baseline of comparison. To obtain this, we developed two different null models.

4.2.1 *Movie-Actor Null Model.* In this null model, we take the original network and shuffle the edges between the movies and the actors to create a bipartite configuration model for this layer of the network. We keep the out-degree of each movie node the same, since each movie must retain its original cast size, but we do not enforce the in-degree of the actor nodes. In other words, all actors are on the same playing field, regardless of their race, gender, or fame. This means that each actor should be cast in about the same number of movies. Note that although we equalize the actors in this way, we are still using our original pool of actors to choose from, so the race and gender makeup of the actors remains the same. The director for each movie remains the same.

The purpose of this null model is to allow us to estimate the expected cast diversities of movies, and by extension the expected movie and director diversity scores, given the existing set of actors in the absence of any sort of casting bias.

4.2.2 *Director-Movie Null Model.* In this null model, the edges between the directors and the movies are shuffled. The out-degree of each director node remains the same, and each movie node still has one in-degree connecting it to some director node. Furthermore, the cast of actors for each movie remains the same. This model works under the assumption that the racial and gender makeup of the existing movies has been pre-defined – perhaps the roles were written for specific races or genders. It then falls to the director to choose which movies to work on, given the diversity of the movie cast.

The purpose of this model is to allow us to estimate the expected director diversity scores, under the assumption above.

4.3 Actor-Actor Assortativity and Modularity

We would like to know whether or not actors in this network tend to co-star with actors of the same race or gender. Since our original network is structured as a tripartite graph in which actors only have edges from movies, we first used the movie-actor bipartite graph to create an actor-actor co-starring network. Then, to assess whether or not actors tend to co-star with similar actors in terms of race and gender, we considered two metrics of this actor-actor graph: assortativity and modularity.

4.3.1 *Assortativity.* We computed the attribute assortativity coefficients for the attributes of race and gender. The attribute assortativity coefficient, r, is the Pearson correlation coefficient of the given attribute between pairs of nodes. In general, $r \in [-1, 1]$, where $r = 0$ means there is no correlation between the attribute and the linking behavior of the nodes. We hypothesize that the assortativity coefficient for race is slightly positive, because rather than having a perfect mixing of races within casts, we have observed that many movies are either largely white or largely non-white. We also hypothesize that the assortativity coefficient for gender is near 0, since most movies feature both men and women, and while there are many all-male movies (25.4% of movies), this may be a product of the majority-male pool of actors, because we also observed very few all-female movies (4.8% of movies).

4.3.2 *Modularity.* We also clustered the actors by the attribute of interest (race or gender) and measured the modularity of these clusters. Modularity is the difference between the fraction of edges connecting nodes within the clusters and the expected fraction of edges within these clusters.

4.4 Actor-Director Assortativity

In our literature review, we found that the percentage of minority actors on screen increases when films have a minority director. We would like to verify this statistic by estimating the race and gender assortativity between actors and directors in our network as compared to the baseline assortativity given by the director-movie null model. We estimate assortativity of the bipartite actor-director graph by calculating the fraction of edges for which the actor and the director have the same race or gender.

4.5 Correlation Between Movie Diversity and Box Office Performance

A practical issue of having such un-representative movie casts is that the demographics of these casts do not reflect the demographics of the audience, which may lead to reduced box office sales. In order to evaluate the impact of on-screen diversity on box office performance, we calculate the Pearson correlation coefficient for movie diversity score and movie revenue (normalized by movie budget).

4.6 Times Series Analysis

With all of these aforementioned analysis methods, much of our interest comes down to not only what the current state of past and present films tell us about diversity, but how these statistics have changed over time. Therefore, we created a way to build up our actor-director-movie graph over time according to film release dates, and calculate aforementioned statistics like diversity scores and assortativity, for the graph at each time step. Specifically, we decided to calculate the statistics for just the movies/directors at each time step, so as to be able to clearly graph how the diversity statistics of new movies/entrants in Hollywood over time. For diversity scores, we predict a mostly flat trend over time with a possible slight increase, as, per our previously cited statistics, diversity in Hollywood has not improved much over time. For assortativity, we estimate a relatively flat trend as well, indicating a not-insignificant but not changing interaction among racial minorities and whites.

5. RESULTS

5.1 Aggregated Results

5.1.1 *Movies.* We created the bipartite-configuration model for the movie-actor bipartite graph and used this as a null model for our analyses of the movie diversity.

We computed various statistics for the movies in our dataset, including the fraction of movies with all-male casts, all-female casts, all-white casts, and all-non-white casts. We also calculated the fraction of movies in which the cast is at least 50 percent female. Figure 3 shows this statistics in relation to the expected statistics derived from the null model. From this, it is evident that in the our real movie network has an overabundance of movies with all-male casts and movies with all-white casts. In addition, there is a lack of movies with all non-white casts and all-female casts.

5.1.2 *Directors.* One of our goals was to determine whether or not minority directors tend to work with more diverse casts. If this is the case, then one potential method for increasing on-screen diversity would be to increase diversity behind the camera by promoting more minority people to director positions. We used the actor-director assortativity heuristic described in section 4.4 to asses this on the real actor-director network as well as the director-movie null model. We found that the fraction of edges for which

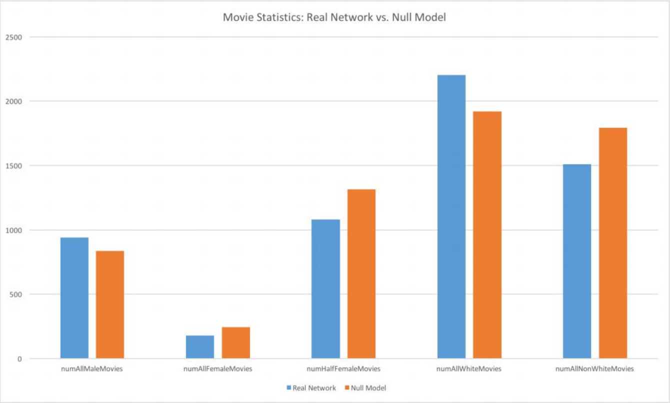

Fig. 3. A comparison of movie diversity statistics in the real network and the movie-actor null model.

the actor and the director were of the same race was 0.587. This is slightly higher than we would expect. The same calculation on the null model yielded 0.502 for the fraction of same-race edges. The difference was less significant in the case of gender. We found that the fraction of edges in the real network for which the director and actor were of the same gender was 0.630 for the real network, which was only slightly higher that what we expected based on the null model – 0.605.

We also compared the average director gender diversity score and average director racial diversity score for the real network and the director-movie null model and found that there was no significant change. This implies that individual directors are not especially biased in terms of the movies they choose to work on.

5.1.3 *Actors.* From our tripartite director-movie-actor network, we derived the actor-actor co-starring network in order to asses the race assortativity and the gender assortativity of the actors. We computed the race assortativity coefficient to be 0.257. This was significantly higher than the baseline race assortativity computed from the movie-actor null model, -0.002. As expected, both the race assortativity coefficient and the gender assortativity coefficient was near zero for this random graph. On the real network, the gender assortativity coefficient was a statstically insignificant 0.029.

These results imply that that actors of the same race tend to appear in the same movies, but there is no correlation between gender and actor-costarring relationships.

5.1.4 *Diversity Scores and Box Office Performance.* We computed the Pearson correlation coefficient for movie racial diversity scores and box office performance (movie revenue divided by movie budget) as well as movie gender diversity scores and box office performance. We found no significant correlation for either of these metrics. This seems to imply that the diversity (or lack thereof) of movie casts is does not have a significant impact on ticket sales.

5.2 Evolution of the Hollywood Network Over Time

Our time series graphs clearly show a disappointing trend of little change in Hollywood diversity over time. However, there are some interesting and encouraging trends.

5.2.1 *Diversity Scores.* The most prominent feature to note in Figure 5 is the slight increase in racial and gender diversity scores over time. While the beginning of the graph is rather inconsistent (due to the small data set size we have for movies before the 1950s-1960s) you can clearly see a slight upward trend in diversity scores beginning around 1970. This falls in line with our expectations of a slight (but minimal) increase in cast and director diversity over time as social norms change to accept a wider diversity of actors/directors in Hollywood.

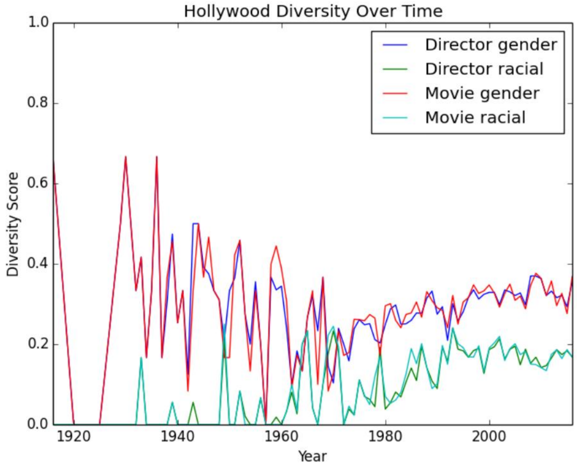

Fig. 4.

5.2.2 *Actor-Actor Assortativity.* The most prominent feature to note in Figure 4 is the mostly flat gender assortativity score, and the slightly increasing racial assortativity over time. Again, we disregard early data pre-1950 due to the small number of data points/movies. Interestingly, there is a significant bump in racial assortativity at 1970, which, upon further investigation, happens because of an influx in new movies with very homoegenous casts. ”Patton”, for instance, released in 1970, has a cast made up mostly of white males. Similarly, ”Cotton Comes to Harlem”, also released in 1970, has a cast made up almost entirely of minorities. This concentration and homogeneity results in a spike in assortativity that declines as later movies do not rise to similar levels of concentration. As for the gender assortativity, the results imply that there is no significant tendency for actors of the same gender to work together.

5.2.3 *Actor-Director Assortativity.* We saw a slight decrease in actor-director assortativity over time for both race and gender. This may imply that directors and actors have become more open to working with people of different races or genders over time.

5.2.4 *Diversity Scores and Box Office Performance.* We found absolutely no correlation between diversity scores and box office performance.

6. CONCLUSION

As our data analysis shows above, both from static snapshot analysis and time series analysis, we see a slight but insignificant increase in Hollywood diversity over time. From the time series analysis, which showed slight upward trends in diversity scores of films and directors, to our investigation of null models, which showed that movie casts are less diverse in real life than in our null models, we see a disturbing trend that has not changed significantly over

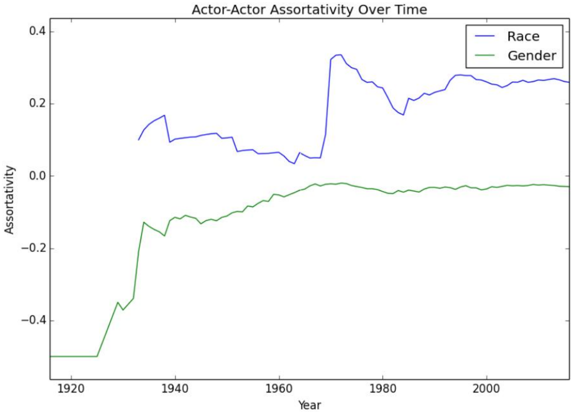

Fig. 5.

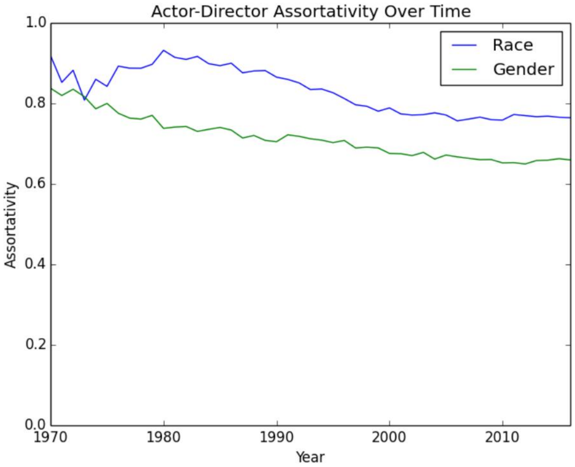

Fig. 6.

time. We hope that our investigation into and calculation of these statistics serves to encourage more pushes towards further diversity in Hollywood.

7. REFERENCES

1 Smith, Stacy L., Marc Choueiti, and Katherine Pieper. ”Inequality in 800 Popular Films: Examining Portrayals of Gender, Race/Ethnicity, LGBT, and Disability from 2007-2015.” Media, Diversity, Social Change Initiative (n.d.): n. pag. Web. 17 Nov. 2016.

2 Smith, Stacy L., Marc Choueiti, and Katherine Pieper. ”INCLUSION or INVISIBILITY? Comprehensive Annenberg Report on Diversity in Entertainment.” Institute for Diversity and Empowerment at Annenberg (IDEA) (n.d.): n. pag. Web. 17 Nov. 2016.

3 Herr II, Bruce W., Ke, Weimao, Hardy, Elisha, and Borner, Katy. (2007) Movies and Actors: Mapping the Internet Movie Database. In Conference Proceedings of 11th Annual Information Visualization International Conference (IV 2007), Zurich, Switzerland, July 4-6, pp. 465-469, IEEE Computer Society Conference Publishing Services.

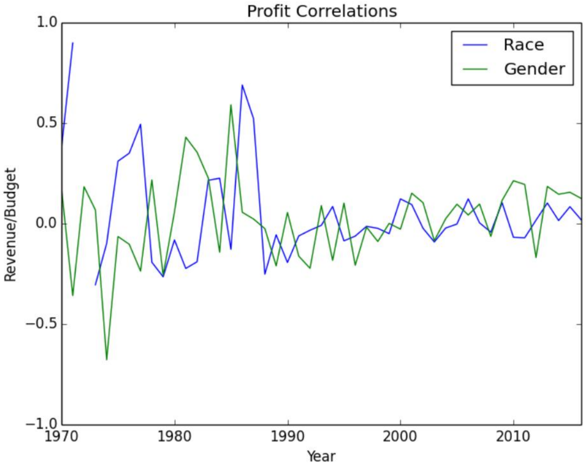

Fig. 7.

4 https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

5 Newman, Mark EJ. ”Mixing patterns in networks.” Physical Review E 67.2 (2003): 026126.

6 www.census.gov/newsroom/releases/archives/2010census/