Diversity in Hollywood: From Studios to Directors to Actors
CS224W Project Milestone
Team Members: Catherine Dong, Nick Troccoli, Jessica Zhao

Part I. 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 studios -- 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 and studios over time; and 2) Uncover gender- and race-based assortativity patterns in the actor-actor and actor-director coworking networks.

Part II. Prior Work

Surveying the Current Landscape

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.

On Clustering

We are particularly interested in unearthing trends among groups of actors, like minority actors. We were drawn to examining clustering after reading "Movies and Actors: Mapping the Internet Movie Database," specifically when the authors investigate the winners and nominees of the Best Actor / Actress Academy Award. They found that those nominated and those who won are tightly clustered together. This "may mean that in order to increase one's chances of an Academy

[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.

Award for Best Actor / Actress, one should work closely with actors in this cluster" [3]. We are curious if similar clustering exists among minority actors or non-minority actors, which results in lopsided cast diversity.

Part III. Data Collection

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 8600 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 3804 of the remaining 4903 unknown genders. With information pulled from Kaggle, NNDB, and SexMachine, we believe we have a good dataset with which to run our analyses.

The race and gender distribution in our data is as follows, for labeled actors and directors:

White: 3052
Black: 282
Hispanic: 83
Multiracial: 60
Asian: 50
Asian/Indian: 39
Middle Eastern: 11
American Aborigine: 7

Male: 2508
Female: 1078

Part IV. Network Structure

Once we gathered our data, we represented it as a 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 an undirected graph where edges exist between directors and the movies they directed, and movies and the actors in those movies. NetworkX also allows us to add metadata to each node such as

[3] Herr II, Bruce W., Ke, Weimao, Hardy, Elisha, and Börner, 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.
[4] https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

information about each person and movie. With this approach, we have created a graph with 4,919 movie nodes, 6,255 actor nodes, 2,399 director nodes, and 19,495 edges.

Creating this graph was not without some complications, however. First and foremost was the realization 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 they're related to. We are still experimenting with other ways to model our graph, including using a directed graph where edges go from directors to movies and from movies to actors. This way, it would be easier to tell the relation between a person and a movie without resorting to two nodes per person, but having separate nodes for the same person may give us bad degree information.

Part V. Preliminary Analyses and Results

Assortativity

With this tripartite graph mostly complete, the first metric we explored was assortativity. Assortativity refers to the the likelihood that nodes form edges with other "similar" nodes, much like in the aforementioned example of Academy Award clustering. In this case, however, our measures of similarity are the attributes "gender," "race," and "isWhite" (a generalization of race in which all non-white actors are grouped together). We compute the assortativity coefficient for the actors in our network for each of these attributes to determine whether actors in the same gender/race group tend to co-star together.

Technical Details

The assortativity coefficient, r, is given by the following equation[5]

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} = \frac{\mathrm{Tr}\,\mathbf{e} - \|\,\mathbf{e}^2\,\|}{1 - \|\,\mathbf{e}^2\,\|}$$

where e_{ij} is the fraction of edges in a network that connect a vertex of type i to one of type j, and a_i and b_i are the fraction of each type of end of an edge that is attached to vertices of type i. The "gender" attribute has 2 types, the "race" attribute has 9 types, and the "isWhite" attribute has 2 types.

The range of r is $[-1,\ 1]$. Positive values indicate that actors with the same gender/race tend to co-star together, while negative values indicate that actors with different genders/races tend to co-star together.

As described above, our full network is a tripartite graph of directors to movies to actors. We aim to calculate the attribute assortativity coefficients for only the actor nodes in the network. Thus, we first create an actor-actor association multi graph in which actors are connected to other actors with whom they co-star. There are parallel edges between pairs of actors who costar in multiple movies together.

[5] Newman, Mark EJ. "Mixing patterns in networks." *Physical Review E* 67.2 (2003): 026126.

Results
Using this actor-actor network (and discarding actor nodes for which we do not have race or gender information), we computed the following attribute assortativity coefficients:

"race": 0.122
"isWhite": 0.152
"gender": 0.0254

From these results, we see that there is a slight positive association between actors of the same race and little positive association between actors of the same gender. We plan to run statistical analyses on these values to determine their significance. Note that our current data only has the top three actors from each movie. These numbers may change significantly when we analyze more members of the casts.

Null Models
These results do not mean much, however, without a baseline comparison. In order to identify baselines for our analyses, we will generate three different null models. These null models tell us what the expected distribution of races and genders would be across our nodes.

Technical Details

Null Model 1: In this null model, we look at the bipartite movie-actor graph and rewire the edges such that the degree for each movie and each actor remains the same. Intuitively, this means that each movie still has the same number of characters and each actor still gets cast for the same number of roles, but the distribution of actors among movies is randomized. We can run the assortativity analysis on this null model to determine how assortative our real network is compared to what we would expect if each actor were cast to a random set of movies.

Null Model 2: This null model is similar to the previous one but we do not enforce the degrees of the actors. In other words, the number of characters in each movie remains the same, but actors get cast to these roles randomly, and each actor will receive approximately the same number of roles. A comparison of our real network to this null model can show us how diverse movie casts would be if parts were cast randomly from the existing pool of actors.

Null Model 3: In this null model, we will randomly assign genders and races to the existing set of actor nodes according to the demographics of the U.S. population. A comparison of our real network to this null model can reveal which movies have the greatest discrepancy between the demographics of their cast and that of the U.S. population.

Results
We generated Null Model 1 described above and ran attribute assortativity analysis on this network. The attribute assortativity coefficients are below:

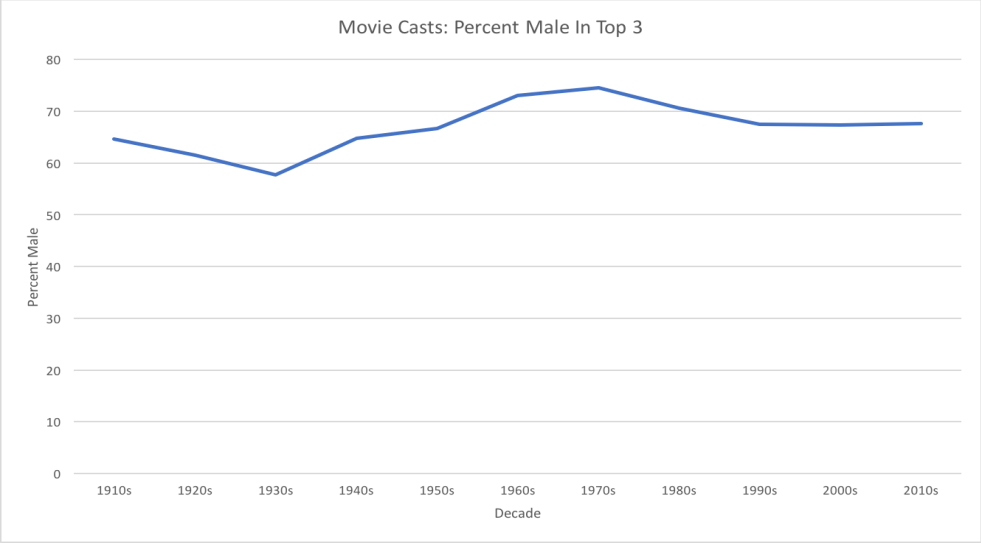
"race": 0.00380
"isWhite": 0.00211
"gender": 0.00132

As expected, the attribute assortativity coefficients for the null model are nearly zero.

Part VI. Future Work

Time Series Analysis

For the next half of our project, we plan to run all the analyses described above across different segments of time. The reason for this is movies and cast diversity are a fluid statistic over time, and we are curious to see what time trends underlie the data in our graph. As an example, in the graph below, you can see a clear stagnation in the high percentage of male actors in our dataset over time.



While we have movies spanning from the 1910s - 2016, for some future analyses we plan to trim off data from before 1980 due to the low number of movies from that era in our dataset. We will divide our data by movie release year into 5 buckets: 1980's, 1990's, 2000-2005, 2005-2010, and 2010-2015. We plan to break down 2000-2015 into 5-year spans because of the high number of movies in that time span, and also to be able to see more fine-grained patterns over the course of the last 15 years. We will then generate a separate network for the movies in each of these buckets and run the analyses above to identify trends in gender/race distribution and assortativity over time in the film industry.

Missing Genders and Ethnicities For Actors

We also hope to fill in more of the gaps in our race and gender data. As mentioned previously, we are currently using NNDB and SexMachine to match actors to their races and genders, which has given us a solid base of results. However, there is still a significant amount of information missing. Out of our 8492 actors and directors, 1099 are missing their race and 4906 are missing their gender. To deal with this gap, we have a variety of potential strategies we will explore in the next weeks.

First, we can hand-classify by looking up the actors' biographies. The strengths of this approach include accuracy; the weaknesses of this approach include the tedious labor and possible inability to manually fill in all remaining holes.

Second, we can use a photo-to-race classifier which, given an actor's headshot, will assign the most likely race. This will most likely require an existing classifier and a databank of images. The strengths of this approach include efficiency; the weaknesses of this approach include the overhead of downloading many images and the probability of finding a robust photo-to-race classifier.

If we are unable to fill in missing race information specifically, we may instead opt to more heavily weight the gender analyses than the race analyses, since we are missing much less data on gender than race. However, we are still looking for other ways to fill in gaps in our race data.