CS224W Project Proposal

Project Title: Diversity in Hollywood: What are the Numbers?
Team Members: Catherine Dong, Nick Troccoli, Jessica Zhao

Part I. Abstract (200 words)

Underrepresentation of minorities and women in movies has been a persistent issue in Hollywood. One reason for poor casting of actors from these groups is the perception that they will negatively affect movie marketability. 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 patterns connecting movie cast diversity to box office performance. Pinpointing the largest culprits of diversity evasion as well as laying out diversity's history in Hollywood urges actionable steps for industry change.

Part II. Related Work (2-3 pages)

Summary of Literature

To deepen our understanding of diversity in Hollywood, we studied the three following papers: "Inclusion or Invisibility? Comprehensive Annenberg Report on Diversity in Entertainment," "Inequality in 800 Popular Films," and "Movies and Actors: Mapping the Internet Movie Database."

"Movies and Actors: Mapping the Internet Movie Database" gave an overview of studies around movie and actor networks in general and provided a great starting point. The paper's overall goal was to visualize the connections between all movies and actors, similar to our project goal. They used a variety of metrics to organize and display their data, including genres, sorting by year, actors that acted in movies together (to study co-actorship), and even Academy Award nominations. They modeled their visualization using multiple layers, including a movies layer, actor layer, popular actors layer, Academy Award best actor/actress layer, and an Academy Award best picture layer. At the heart of it all was a bipartite graph, much like we've studied in class, of actors and movies, with edges going from movies to actors in those movies.

While the end product of the paper were mostly visual and thus hard to summarize in text, a few interesting results stood out. First, when visualizing the winners and nominees of the best actor/actress Academy Award, they found those nominated and who won are tightly clustered together which, as the paper suggests "may mean that in order to increase one's chances of an academy award for best actor/actress, one should work closely with actors in this cluster." A second interesting result was the density of the visualization they generated; in fact, they had difficulty rasterizing it and displaying it at a manageable size. Filtering down data was an issue at the graph level as well; when modeling co-acting as weighted edges between actor nodes (based on how frequently they co-acted), they discarded all links with a weight less than three. Finally, they visualized their data over time, plotting both movie growth as well as number of starring actors in movies. All of these results tie into our project, as will be discussed in the next section.

While "Movies and Actors: Mapping the Internet Movie Database" gave an overview of visualizing movie and actor networks over time, two other articles, "Inclusion or Invisibility? Comprehensive Annenberg Report on Diversity in Entertainment" and "Inequality in 800 Popular Films," provided a more specific focus into film diversity, the topic we want to explore. There's long been an issue of proportional representation of minority races and women. There has been increased awareness and interest in this subject in recent years, especially after the #OscarsSoWhite movement around the 2016 Academy Awards.

Data shows that minorities and women are indeed being underrepresented in Hollywood, both onscreen and behind the camera. USC's Annenberg School for Communication and Journalism has conducted yearly studies on the gender and racial makeup of top films.

According to "Inclusion or Invisibility? Comprehensive Annenberg Report on Diversity in Entertainment," of the 11,306 speaking or named characters assessed, only 28.3% were from underrepresented racial/ethnic groups. Furthermore, only 14% of films depicted an underrepresented lead or co-lead. None were played by an Asian actor. 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.

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. In film, a mere 3.4% of directors were female.

So why is the representation of minorities and women so poor in Hollywood? One reason is the perceived impact that casting actors from these groups has on the success of a production.

For example, directory Ridley Scott, who faced scrutiny for casting white actors as Egyptian characters in the big-budget film "Exodus: Gods and Kings," said that if his "lead actor is Mohammad so-and-so from such-and-such ... [he's] just not going to get it financed. So the question doesn't even come up."

While many directors and production studios view casting minorities as a risky business decision, some studies have shown that movies with better minority representation actually perform better at the box office. Kaden Lee of Brown University found that movies featuring black actors outperform those without black actors by 40% in ticket sales.

Critique of Literature

The USC Annenberg reports on diversity in film and TV present many important statistics on the state of the industry as a whole. Most of these statistics were based on data aggregated across all the movies that were analyzed. However, the report lacks finer-grained analysis on specific movies, actors, and directors.

The Brown University paper provided a comparison of movie revenue for films that featured white actors versus those that included black actors. What it lacks is information about the impact of casting actors of other ethnicities as well as women.

We strive to fill these gaps in specifically-available analysis, the range of studied ethnicities, and the relationship between diversity and box office performance with our proposal below.

Part III. Proposal (1-2 pages)

Problem Definition

From our literature review, we have seen that underrepresentation of minorities and women is a real issue in the movie industry. We have also identified that one reason for poor casting of actors from these groups is the perception that they will negatively affect movie marketability. The data on diversity in the industry we have seen have all been aggregate data across many movies, actors, directors, and studios. 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.

Our goal is to identify the Hollywood directors and studios that work with the least and most diverse movie casts and examine the change in diversity over time. As with many industries, the Hollywood film industry is largely profit-driven and therefore, we hypothesize that Hollywood directors and studios favor less diverse casts because those movies tend to perform better at the box office and therefore bring greater financial income. By performing our analysis on Hollywood films beginning from the industry's inception in the early twentieth century to modern day, our goal is to clarify that hypothesis in two ways: 1) Understand the presence and absence of diversity in Hollywood films as related to directors and studios over time; and 2) Uncover patterns connecting movie cast diversity to box office performance.

Dataset

For data on directors, studios, and actors, will be using data from the Internet Movie Database (IMDB) at www.imdb.com. IMDB provides thorough data on actors, directors, producers, and more from movies and TV shows. For data on the gender and ethnicity of actors, we are using the Notable Name Database (NNDB) at www.nndb.com. We will scrape NNDB and match people to their genders and ethnicities by name. For data on box office profits, we will be using Box Office Mojo at www.boxofficemojo.com. Box Office Mojo provides a detailed breakdown on revenue generated by a film.

Algorithm

We will use a SNAP undirected graph with four categories of colored nodes: Directors, Studios, Movies, and Actors. An edge exists between a Director and a Studio if they have collaborated; between a Director and a Movie if the director directed the movie; between a Studio and Movie if the studio funded the movie; between a Movie and an Actor if the actor starred in the movie.

We will use Gephi software to visualize our data and build a searchable web interface, in which one can enter any director or studio's name to find the entity's overall diversity score and a breakdown of the entity's associated movies and actors by gender and ethnic diversity.

Our methodology for each of the objectives described in our *Problem Definition*:

Objective 1: Understanding Movie Cast Diversity

To measure the diversity of a movie cast, we will define a metric that looks at the ethical and gender composition of the movie's top-billed, speaking roles. We measure gender diversity and ethnic diversity separately and calculate the score for each respective category by taking the percentage of diverse cast members from the top-billed cast. For gender diversity, non-male figures have been historically underrepresented in Hollywood and thus we take the percentage of non-male cast members from the top-billed list. For ethnic diversity, non-white figures have been historically underrepresented in Hollywood and thus we take the percentage of non-white cast members from the top-billed list.

With our movie diversity scores, we look at the relationship between Directors → Movies and assign a director's diversity based on a weighted average of his or her collective movies' diversity. Similarly, we look at the relationship between Studios → Movies and assign a studio's diversity based on a weighted average of the studio's collective movies' diversity.

We apply these calculations to determine the directors and studios who work with the most and least diverse casts not only in the present day, but in films throughout the last century. By doing so, we can model the growth (or lack thereof) in gender and ethnic diversity from the birth of Hollywood film in the early twentieth century to its modern day incarnation.

Objective 2: Patterns between Diversity and Box Office Performance

To examine the potential financial motivations of directors and studios in selecting cast diversity, we connect movie diversity scores to the movie's box office performance and see if there is a correlation between the two entities. Of course, the presence of correlation does not imply that less diverse casts bring higher box office profits or vice versa. However, we are interested in the data we will see that can better shape our understanding of diversity in Hollywood.

Methods for Evaluation

To evaluate the soundness of our algorithm in answering our posed questions, we must determine the efficacy of our diversity metric. We do so partially through our second objective of comparing diversity scores to box office performance and measuring that correlation. As directors and studios are historically known to be biased towards less diverse top-billed cast due to supposedly higher revenue, we expect our diversity metric to be inversely correlated with box office performance.

We additionally plan on improving our diversity metric as we begin working with the data by periodically sampling random sets of directors, studios, movies. We will examine each entity's diversity scores individually - looking at the entity's neighboring edges and nodes - to see if there are any overlooked factors we must incorporate into our diversity metric to make our measurements more accurate.