Diversity in Hollywood: From Directors to Movies to Actors

Catherine Dong, Nick Troccoli, Jessica Zhao
{cdong, troccoli, jesszhao}@stanford.edu, Stanford University, Department of Computer Science

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.

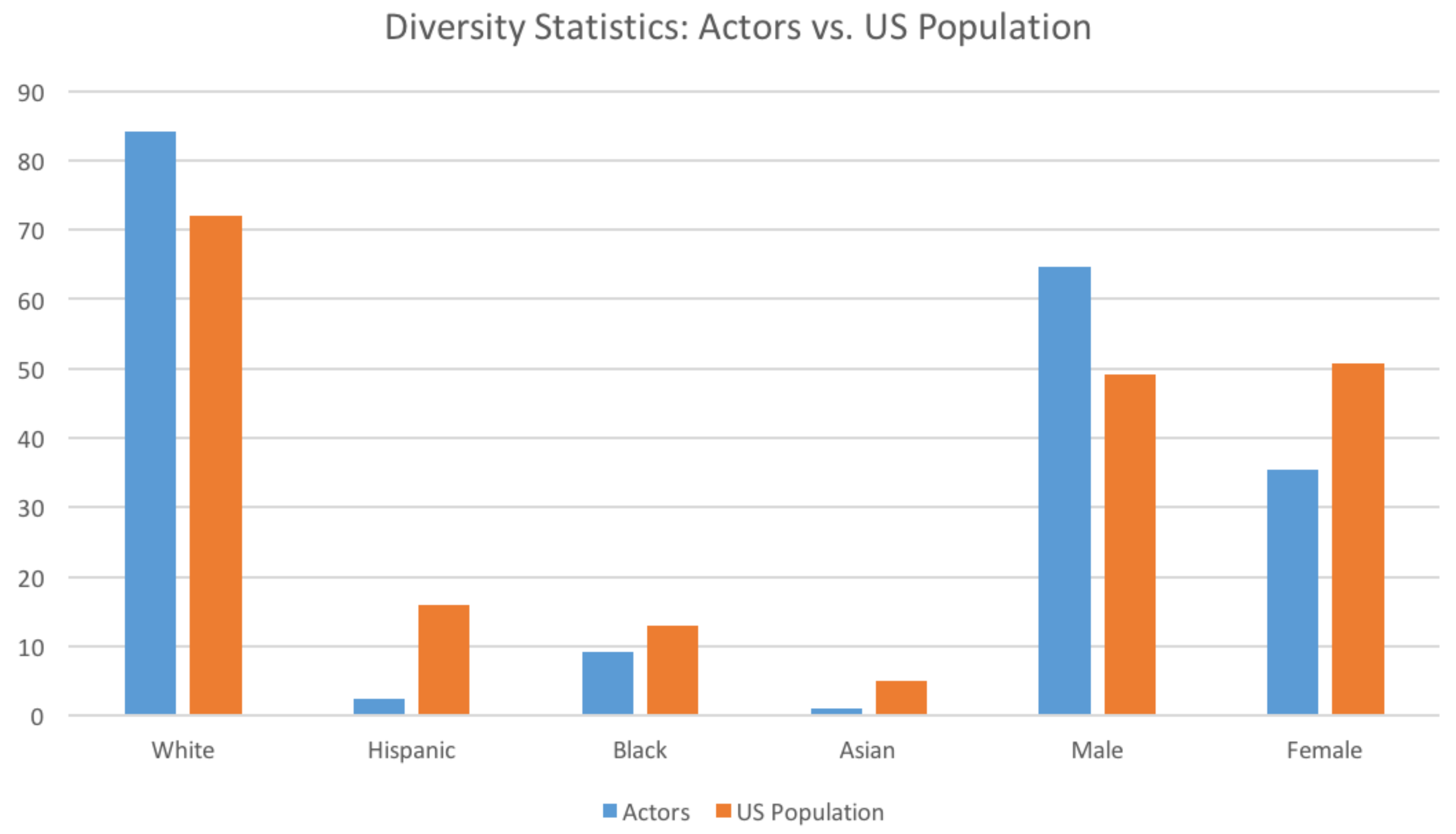
Data Collection

Custom-curated dataset combining:

- Kaggle
- Notable Names Database (NNDB)
- Internet Movie Database (IMDB)
- SexMachine

19495 edges connecting
7311 movies
4204 actors
1813 directors

Preliminary numbers indicate a disparity in diversity between representation in Hollywood and representation in the US population.



Methods

A Diversity Metric

To perform our analysis, we design a naive measure of diversity. Because Hollywood has traditionally favored white males, we consider racial diversity as non-white and gender diversity as non-male. We calculate these two diversity scores for movies as the fraction of the cast's diverse top-billed actors. We calculate the two scores for directors as the average of his or her collective movies diversity scores.

Null Models for a Baseline Comparison

Movie-Actor Null Model

Purpose: To estimate the expected cast diversities of movies, and by extension the expected movie and director diversity scores.

Construction: 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.

Director-Movie Null Model

Purpose: To estimate the expected director diversity scores.

Construction: We take the original network and shuffle 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.

Assortativity & Modularity

Within our actor-actor network, we would like to know whether or not actors tend to co-star with actors of the same race or gender. Additionally, 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 examining our actor-director network and 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.

Assortativity
Pearson correlation coefficient of
the given attribute between pairs of nodes

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

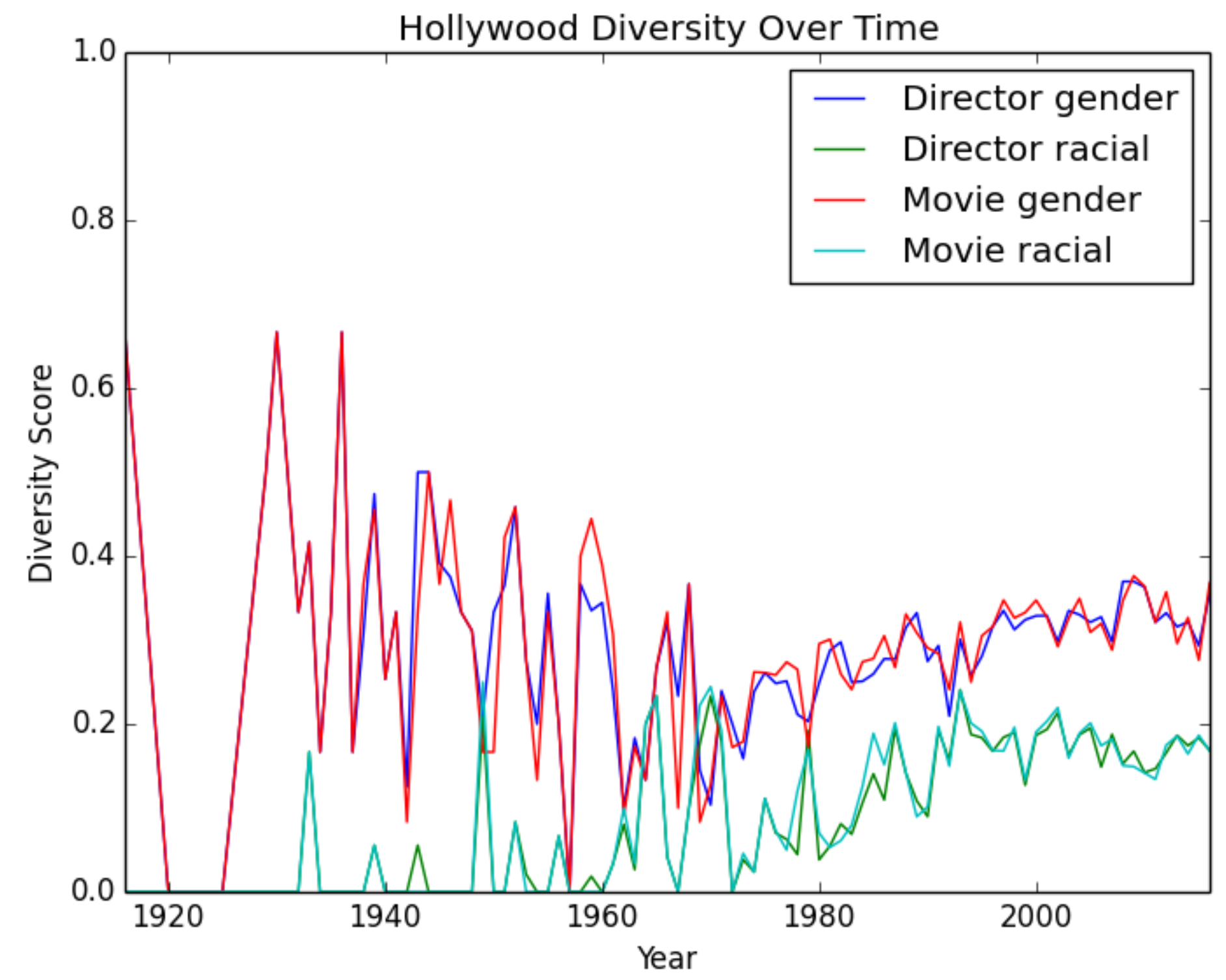
Modularity
The difference between the fraction of edges connecting nodes within
the clusters and the expected fraction of edges within these clusters

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \frac{s_v s_w + 1}{2}$$

Time Series Analysis

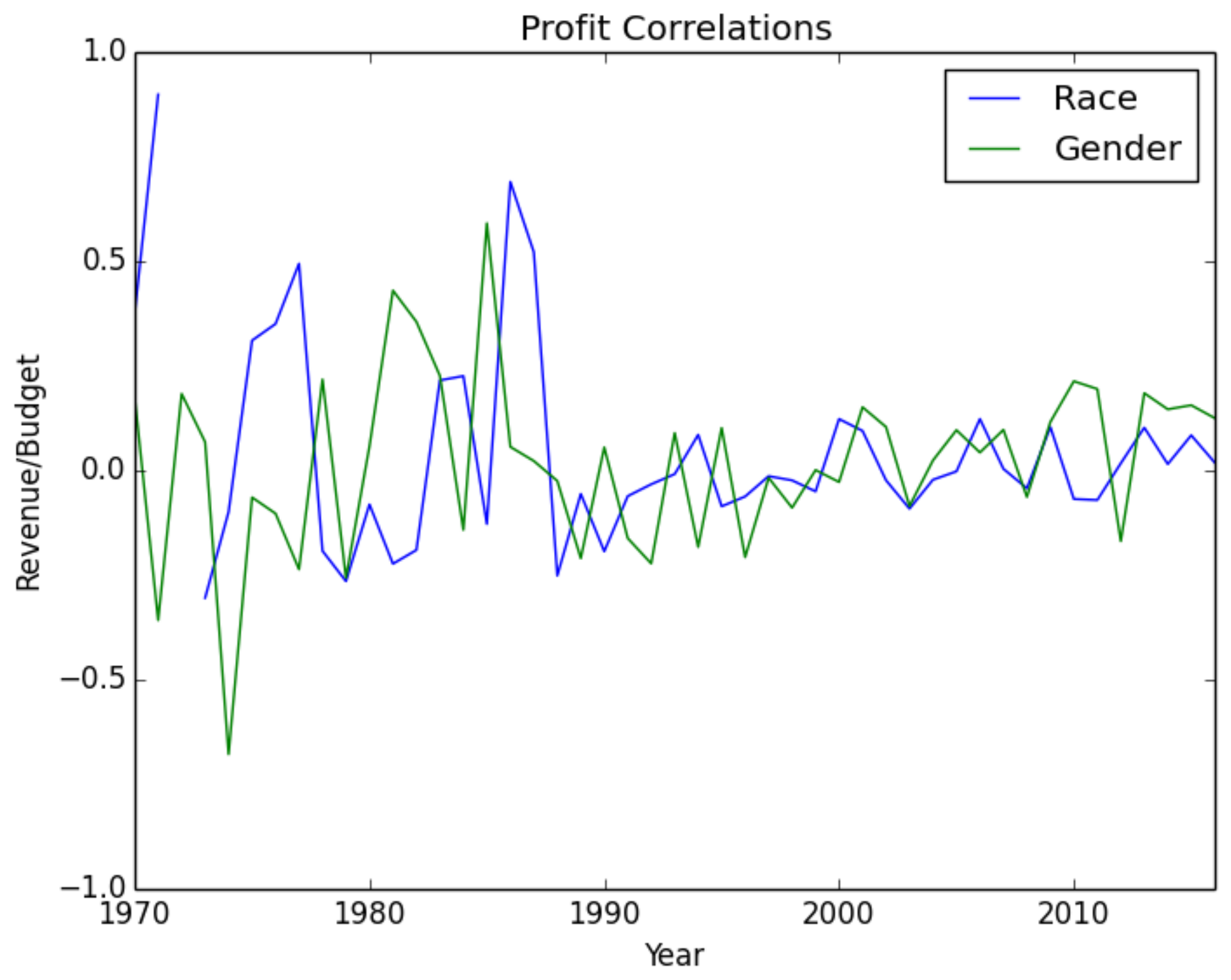
Over time, we observe a slight upward trend in racial and gender diversity scores

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.



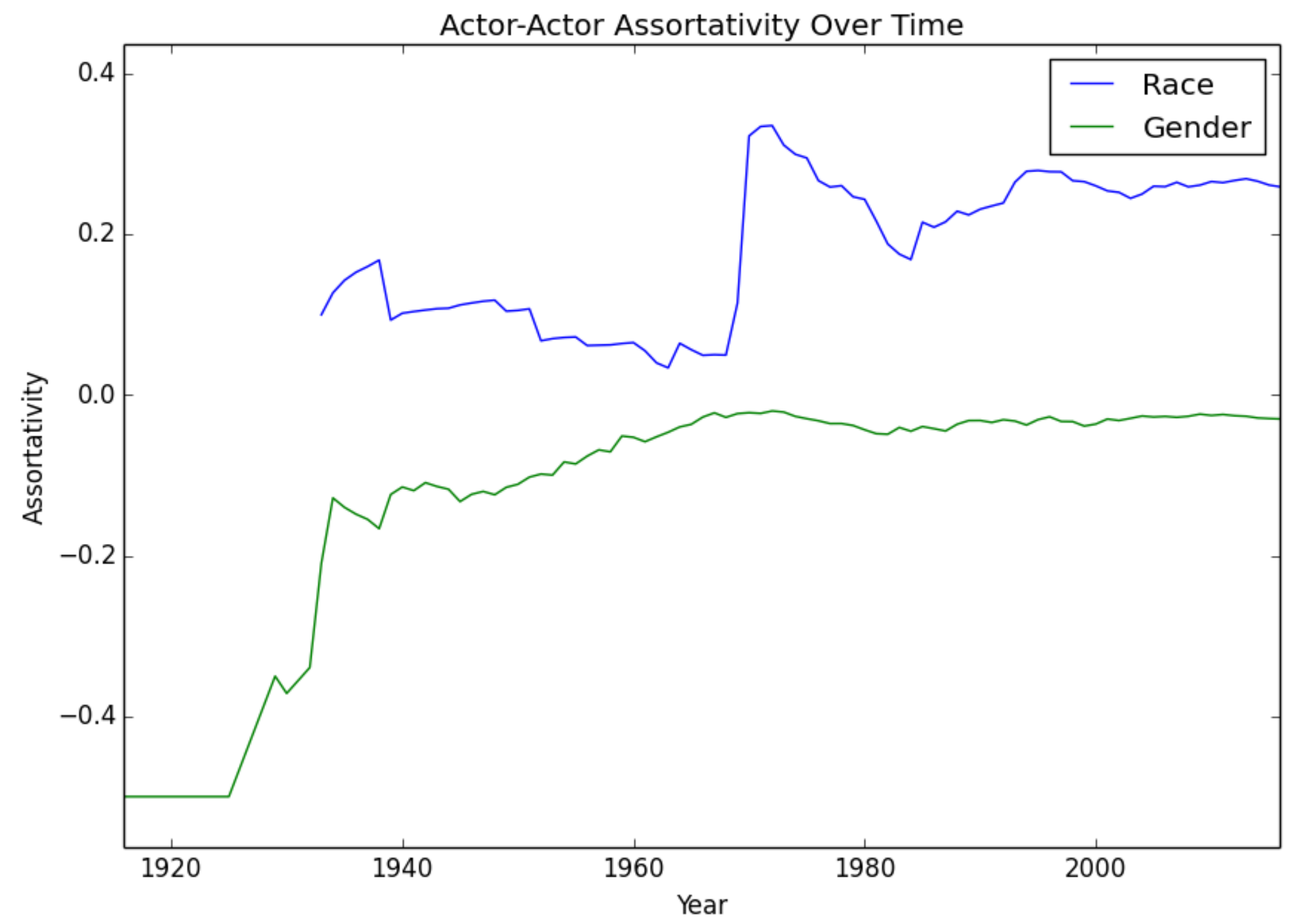
No correlation found between diversity measures and box office performance

This may imply that directors do not have to be overly concerned about casting particular races or genders when considering the financial performance of a film.



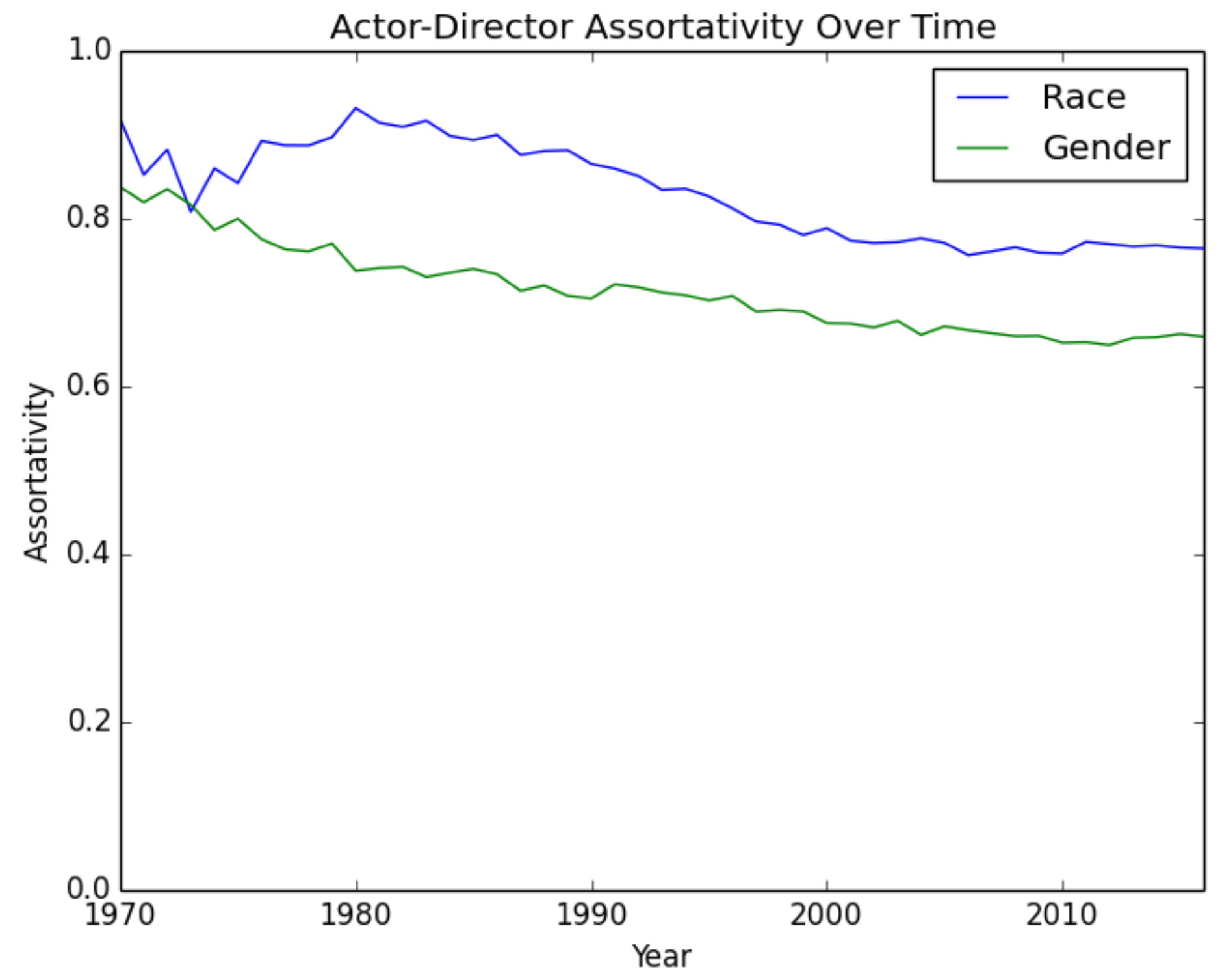
Actor-Actor Assortativity: Primarily flat gender assortativity score; slightly increasing racial score over time.

This may imply that there is no significant tendency for actors of the same gender to work together.



Actor-Director Assortativity: Slight decrease in actor-director assortativity over time for both race and gender.

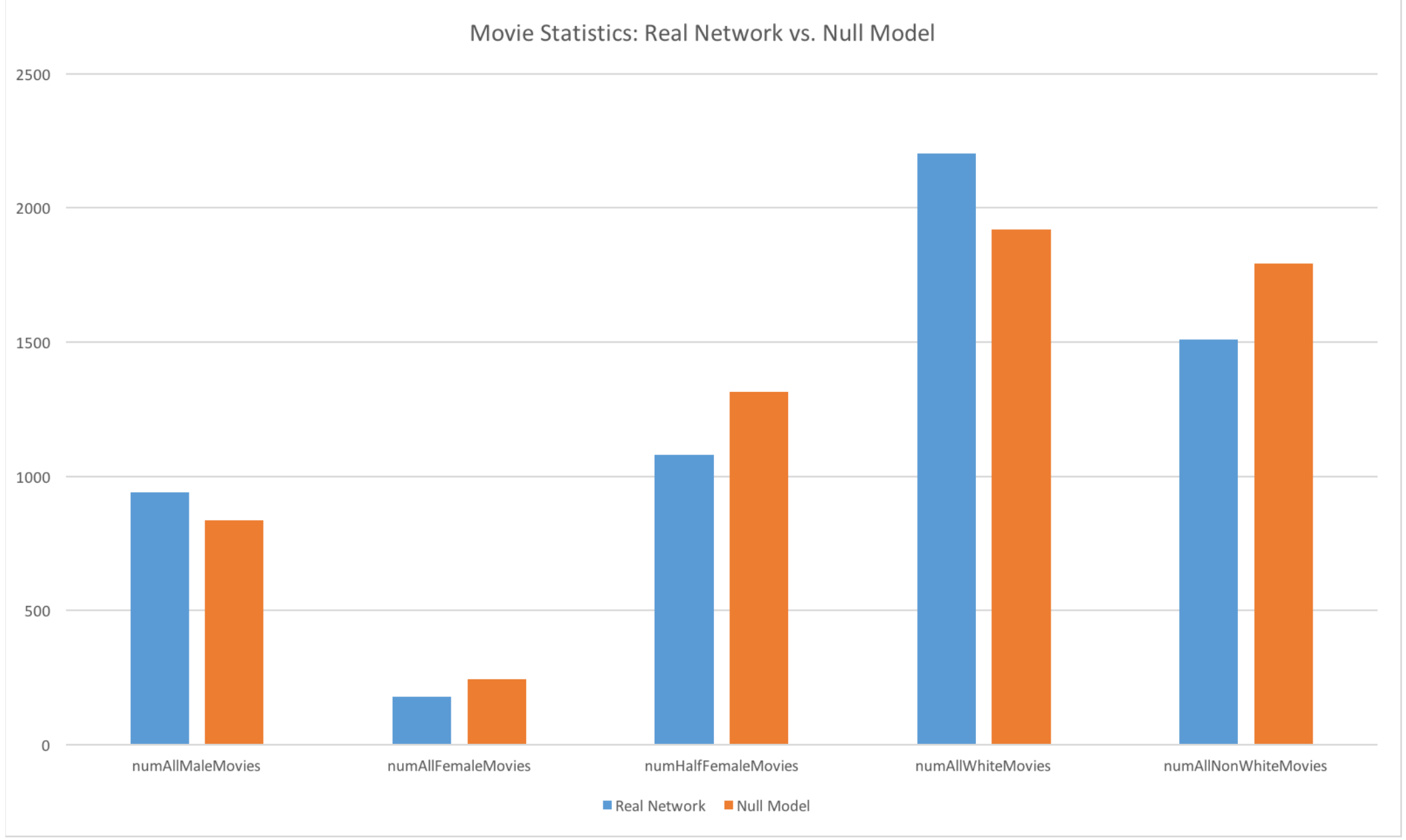
This may imply that directors and actors have become more open to working with people of different races or genders over time.



Conclusions

Real network has overabundance of movies with all-male casts and movies with all-white casts

There is a distinct lack of movies with all non-white casts and all-female casts. However, comparing director diversity scores with the null model scores indicates no significant disparity, implying individual directors are not particularly biased in terms of the movies they choose to work on - an encouraging observation.



As our data analysis illustrates, 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 an alarming trend that has not changed significantly over time. We hope that our investigation into and calculation of these statistics serves to encourage a push towards further diversity in Hollywood.

Acknowledgments

We would thank our TA Will Hamilton for his feedback and guidance as we worked on framing our research questions and performing our analysis. Many thanks to the CS224W teaching team for their time and to Professor Leskovec for a great course.