# Hidden Signals of Hollywood: Analyzing Character Interaction Patterns Across 50 Years of Film

**Will Hamilton[1], Yizhi (Elise) Jing[2], Lu Liu[3], Andrew Mellor[4], Moriah Echlin[5], Juste Raimbault[6], Daniel Biro[7], Michael T. Schaub[8], Catriona Sissons[9], Harrison B. Smith[10], and Xiao (Thomas) Zhang[11]**

[1]Affiliation, department, city, postcode, country
[2]Affiliation, department, city, postcode, country
[*]corresponding.author@email.example
[+]these authors contributed equally to this work

## ABSTRACT

Film is a key medium of culture. However, the evolution of film has yet to be analyzed through a quantitative lens. Here, we perform a large-scale analysis of digitized footprints of over 5000 films spanning 50 years. Our data contains information about which characters are on screen at what times, along with full dialogue data for a smaller subset of films. Using this data we extract character interaction networks and construct dynamical models of character interaction patterns, showing that this information reliably predicts genre and other film attributes, such as critical reception. We then use the historical data to characterize how gender roles in film have evolved over the last 50 years. Our analysis provides a quantitative perspective on gender disparity in film and highlights how different statistical attributes (e.g., interaction-network centrality, screen time) are equalizing at vastly different rates. We find that the total share of screen time for male vs. female characters is equalizing at a relatively fast rate, with complete equalization within $\sim$40 years if trends continue; in contrast, when examining various measures of character centrality in the interaction networks, our analysis suggest that full gender equalization will take $>$80 years if current trends continue.

## Introduction

## Datasets

### DigitalSmiths character interaction data

For $\sim$5000 movies we have machine extracted scene data, which specifies who is on screen at what time. This data was extracted by DigitalSmiths Inc. using a proprietary computer vision algorithm along with post-processing hand-curation. The most important aspect of the data is that it lists the appearance times for each character. Some movies contain more data (e.g., "actions" that are being performed on screen) but we have not used this data yet.

### Metadata

For most movies ($\sim$5000), we have a large set of scraped metadata (from IMBD, Rotten Tomatoes etc.), including genre, MPAA rating, Rotten Tomatoes audience score, Rotten Tomatoes critic's score, runtime, cast list, and language. We also have some hand-labeled "descriptors" (from DigitalSmiths) including "themes", "subjects", and "time-periods".

### Gender data

For movie scene data, here we identify gender by the first name of the actor/actress. We use python package SexMachine 0.1.1. There are around 26,000 different actors/actresses in movie scene datasets. After detection, we identify 15,800 actors, 8,000 actresses and 3,000 people who cannot be identified by their first name. Dialogue dataset has a small portion of gender information. To investigate more about genders, we still use the python package to detect the first name of characters. In general, there are 3,900 male characters, 1,900 female characters and 3,000 characters who cannot be identified by their first name.

### Dialogue data

For ∼300 movies we have all dialogues between characters who spoke more than 5 lines to each other. This data is from: http://www.mpi-sws.org/~cristian/Cornell_Movie-Dialogs_Corpus.html.

## Methods: Representing character interactions

### Static interaction networks

We first consider the static networks created from the co-appearances of actors/actress on the screen. Each node in the network represent a character appears in the movie and each edge between actors/actress indicates they have appeared in the same scene. We use the static networks as a paradigm for studying various statistics of the movie networks. First and foremost, we looked into the gender centrality shifts over the decades in movie history. We found that overall the maximal male characters' centralities are higher than those of female characters. However, as time moves forward the differences are becoming much smaller.

### Multilayer interaction networks

To represent the time-dependent evolution of the movie social networks, we create a multi-layer network representation by discretizing the temporal network. For each movie, one-minute bins are created first, in which two characters are considered to have a connection if they appear in the same minute. We then divide the movies into 10-minute time windows, and create each layer of the multi-layer network by aggregating all connections in each time window.

We then consider the temporal changes of the properties of these networks. It is believed that the story development of movies can be represented through the changes of social networks in the movies; furthermore, since different genres have different ways of story development, this distinction may be found through analysis of network properties. We measure this by the following:

$$\Delta p = \frac{\sum_{t=1}^{n} (p_{t+1} - p_t)}{n} \tag{1}$$

Where p is the network property that we consider on each layer, and n is the number of layers in each temporal network. We currently measure 4 network properties: number of nodes, network density, average degree, and average clustering coefficient. These measurements are then used as inputs for the prediction task.

### Character introduction networks

Character introduction networks, as represented by directed adjacency matrices, contain information about how stories are told by how character nodes are added to the network. These matrices are n x n (n being the number of characters), nonsymmetric, and sparse. Each row represents a character, i, that acts as an introducer and each column represents a character, j, that is introduced. Furthermore, the i,jth element is 1 if character i introduced character j and a 0 if they did not. Character j is introduced by charcter i if character i is the most important character (measured by total screen time) on screen when character j is first seen. If a character first appears alone then they introduce themselves.

### Dynamical models of character co-occurrences

To model character coccurrences we create a discretized temporal network. Each movie is split into bins of length $\Delta t$. If an actor appears onscreen during that interval then they are deemed to have appeared for that entire interval. More formally, we define the matrix $A_{ik}$ as

$$A_{ik} = \begin{cases} 1 & \text{if actor } i \text{ appears in interval k} \\ 0 & \text{otherwise} \end{cases}$$

Sorting character indices by total screen time creates a movie *barcode*, shown in Fig.1. From this matrix it is possible to infer the actor co-occurrence network. Instead, these barcodes are used as the input for the training of a Hidden Markov Model (HMM).

We also use methods of non-linear time-series analysis on these character appearance time-series. In particular, we use transfer entropy to compuate a non-linear measure of correlation or influence between two characters.

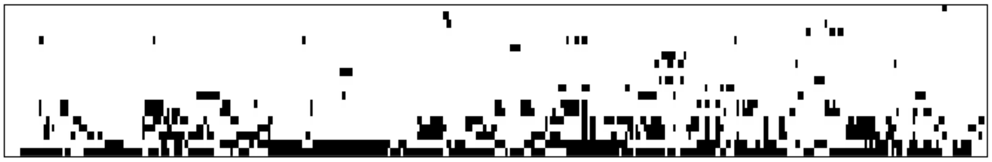


**Figure 1.** A typical discretized actor appearance barcode for the film *Anchorman: The Legend of Ron Burgundy*. Actors with the longest screen time are on the bottom.

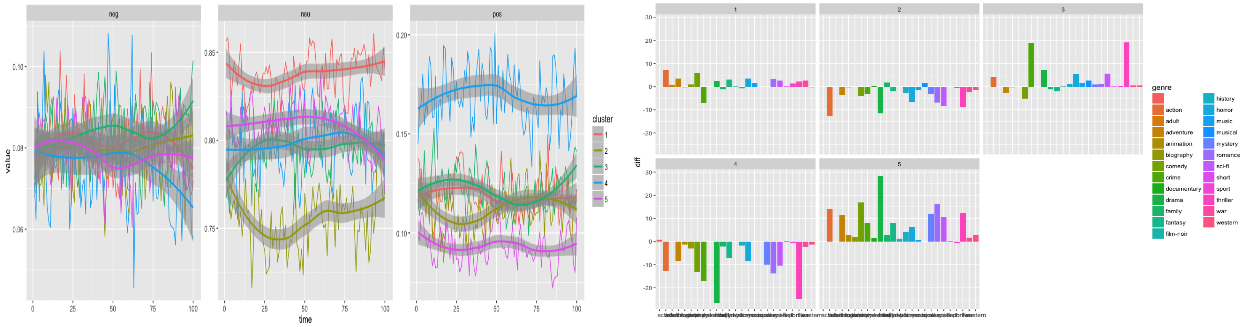


**Figure 2.** Sentiment time-series film classification. *Left :* Trajectories of negative, neutral and positive components of cluster centers. *Right :* Corresponding cluster composition by film genre.

## Time-series analysis of character dialogue

### *Film Classification based on sentiment profiles*

A first approach into dialogue data is the quantification of sentiment level in each line, in order to establish profiles of movies. Sentiment analysis is a widely used method in Natural Language Processing, and tools available are numerous[1]. Interesting examples of application include fields such as literature[2] or political science[3]. We expect strong differences across genres, as for example a drama and a comedy will not have the same narrative structure ("happy end" e.g.). We construct time-series using a sentiment quantifier $\mathscr{S} : s \mapsto (pos, neu, neg) \in [0;1]^3$ that associates a fragment of text with levels of positive, neutral and negative sentiments. It yields time-series with endogenous observation times, for film $i$, at times $t_j$, for sentiment type $s$, $S_i^{(s)}(t_j)$. Time-series are normalized into $p$ standard bins for temporal sampling, yielding a representation of films as vectors of $\mathbb{R}^{3p}$ (films with less lines than $p$, are discarded ; with in our case $p = 100$ a negligible proportion was discarded). We can then cluster the time-series to identify typical "sentiment profiles". We use a rough k-means clustering (with 10 repetitions of initial centers to minimize randomness) on features. The number of cluster is taken at an inflexion point in clustering coefficient as a function of cluster numbers, giving $k = 5$. Trajectories of cluster centers and composition of clusters by film genre are shown in Fig. 2. A relatively good correspondence between sentiment profile and genre is observed (e.g. drama finishing with a high sentiment level).

### *Jenson-Shannon Distance over time*

The imbalance of gender in movies is a reflection of our culture. Another aspect of the dialogue data is to study the dialogue difference between male and female characters. Here we use single-word Jensen-Shannon distance to measure the divergence of dialogue. The higher the distance, the more different word usage.

### *Homophily of gender based on network structure*

Next we would like to investigate why there are gender differences in movies. One assumption is homophily. which means both females and males tend to communicate more to the same gender. To verify this assumption, we Here we measure homophily by counting the ratio of network neighbors from certain gender to the ratio of the gender in the whole network. For example, if a female character in a movie has 50% female neighbors, and the network has 50% female characters, then homophily ratio of this character is 1, which means no homophily. If the ratio is smaller than 1,

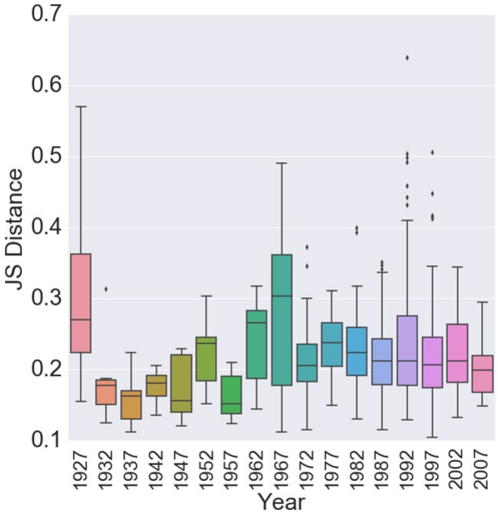


**Figure 3.** The Jenson-Shannon distance from 1927 to 2007. Here Each bar contains movies within five years. In general the distance increases over time before year 1967 and than remains stable.

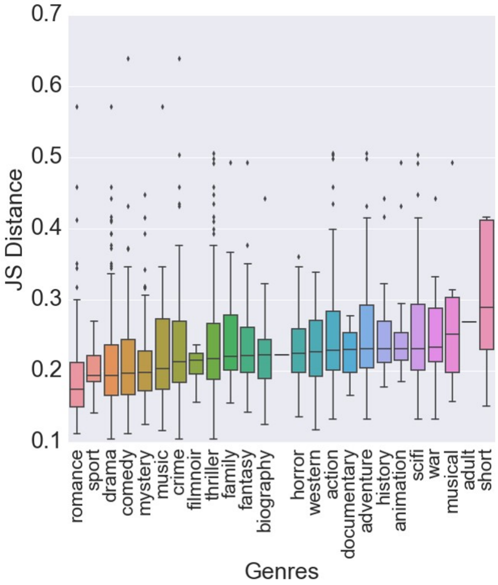


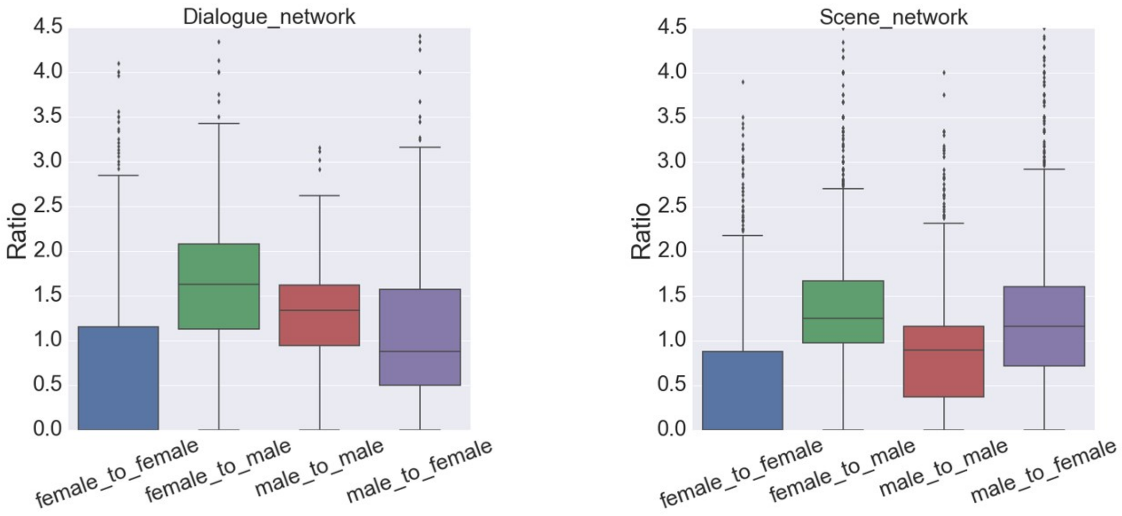
**Figure 4.** The Jenson-Shannon distance for different genres, ranging from romance to short movies. Here romance has the lowest distance, which means male and female characters speak similarly in romance movies.

it means that they link less than network average. If the ratio is larger than 1, it means that there is homophily.

## Predicting metadata from interaction structures

We used random forest regression to predict genre and Rotten Tomatoes scores. We used a massive set of features, including some metadata, network stats from the various interaction networks, aggregate transfer entropy stats, and statistics from the HMMs (number of state changes, entropy). Random forests are a useful machine learning approach for preventing over-fitting in this sort of situation.

For all experiments, we used 5-fold cross-validation to optimize hyperparameters. For genre, we currently have a



**(a)** Dialogue network

**(b)** Scene network

**Figure 5.** (a) The barplot for homophile ration of dialogue network. Here each spot is each movie (only consider the giant component). Both female and male tend to link more to male. (b) The barplot for homophile ration of scene network. Here each spot is each movie (only consider the giant component). For scene network, we only consider characters shown in at least two scenes and total time larger than 5% of the whole scene time. Both female link more to male, while male link more to female (average ratio ¿ 1). The difference may reveal that male and male in the same scene do not often communicate with each other.

six-class accuracy of $\sim$50%, which is a 3× improvement over random; for binary classification between genres, the accuracy is in the range $[60\% - 80\%]$. For predicting Rotten Tomatoes scores, we have an out-of-sample $R^2$ of 0.28 for audience ratings and 0.25 critic's scores.

Random forests also give a measure of "importance" for different features. Interestingly, almost all features where important for genre prediction, but runtime was by far the most important for Rotten Tomatoes. For predicting the critic Rotten Tomatoes scores, the total amount of "transfer entropy", which measures the "predictability" of a movie, was also important.

## The evolution of gender roles

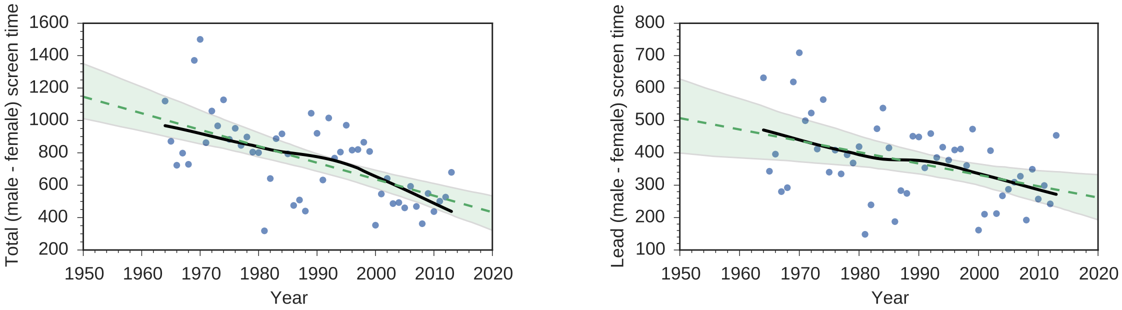
Since we have data for films spanning 50 years, we can use this information to analyze if and how the depictions of different gender roles have evolved.

### Screen time

The share of screen time is the most straightforward way to analyze changes in the portrayal of male vs. female characters. Figure **??** shows that the total share of screentime for male vs. female screentime is equalizing faster than the ratio for lead male vs. female characters. The figures show the average difference within years along with robust linear regression fits and LOWESS kernel regression estimates.

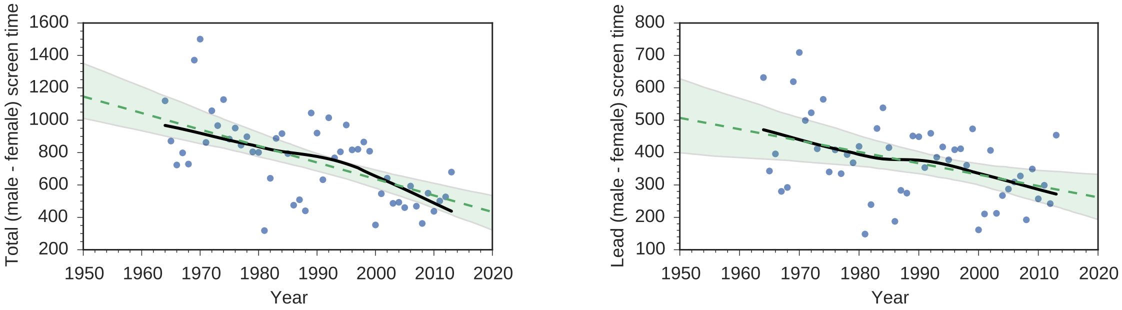
### Centrality

We also analyzed how network centralities of characters of different genres have changed over time using the static co-appearance networks. Figure **??** shows how betweenness and eigenvector network centralities of lead characters have shifted. We see that these measures appear to be equalizing even slower than screen time. Analogous results held with the introduction networks as well.



**(a)** Total screen time difference for male vs. female characters over the last 50 years.

**(b)** Dcreen time for lead male vs. female characters over the last 50 years.



**(a)** Difference in betweeness centrality for lead male vs. female characters over the last 50 years.

**(b)** Difference in eigenvector centrality for lead male vs. female characters over the last 50 years.

## References

1. Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M. & Benevenuto, F. Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* **5**, 23 (2016).
2. Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M. & Dodds, P. S. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* **5**, 31 (2016).
3. Rheault, L., Beelen, K., Cochrane, C. & Hirst, G. Measuring emotion in parliamentary debates with automated textual analysis. *PLOS ONE* **11**, 1–18 (2016). URL `http://dx.doi.org/10.1371%2Fjournal.pone.0168843`.