Tracking onscreen gender and role bias over time

Will Radford[1] and Matthias Gallé[2]

[1] *Canva, work done while at Xerox Research Centre Europe, wejradford@gmail.com*

[2] *NAVER LABS Europe, work done while at Xerox Research Centre Europe, mgalle@gmail.com*

ABSTRACT

Film and television play an important role in popular culture. Their study, however, often requires watching and annotating video, a time-consuming process too expensive to run at scale. In this paper we study the evolution of different roles over time at a large scale by using media database cast lists. In particular, we focus on the gender distribution of those roles and how this changes over time. We compare real-life employment gender distributions to our web-mediated onscreen gender data and also investigate how gender role biases differ between film and television. We propose that these methodologies are a useful complement to traditional analysis and allow researchers to explore onscreen gender depictions using online evidence.

Keywords: Gender bias, Film, Television

ISSN 2332-4031; DOI 10.1561/100.00000001

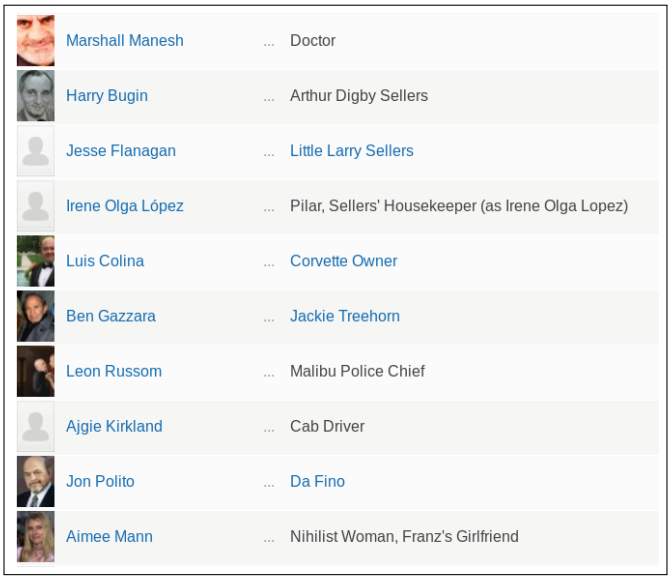


Figure 1: Excerpt from the cast list for "The Big Lebowski".

1 Introduction

Film and television are an integral part of culture and one way that people understand and interact with it. Onscreen scenarios reflect the values from some real or imagined story, but also inform the viewers expectations. However, attempting to directly study film and television presents some difficulties. Watching video for analysis does not scale well to large datasets without significant manual effort. This limits most large-scale study to easily digestible data sources: film popularity, box-office figures, reviews, scripts and other metadata. Although non-video data sources may be easier to study, they limit the types of questions researchers can ask.

Our research question is whether web science can provide viable proxies that allow us to explore interesting social science research questions at scale. Specifically: how onscreen roles are reflected in online data; how these roles relate to gender; how does this relate to reality; how is film different to television; how do these all vary over time? We use data available from a popular media website and examine *cast lists*. Figure 1 is a section of the Internet Movie Database (IMDb)[1] cast list from "The Big Lebowski"[2], showing performer names and images on the left, with their character name on the right. Some character names are proper names (e.g. Arthur Digby Sellers), but some are professional roles (e.g. Doctor) or combinations of role and relation to other characters (e.g. Nihilist Woman, Franz's Girlfriend). We exploit four factors from the data: productions are listed with their release date, they may be marked with a country via their production company, male and female performers are distinguished in the data, and unnamed characters are usually listed by their role or profession. This allows us to count gendered performances of a particular role over time, which can be used to explore social science questions.

This paper is structured as follows: we discuss related work in media gender studies and IMDb in Section 2. Section 3 describes the dataset and the methodology we use to handle noisy user-generated data. We then explore what roles are found onscreen and how they change over time in Section 4 and in Section 5, we examine how roles interact with gender over time. In both cases, we are able to extract plausible trends. We next compare online-mediated gender roles to real-world gender distributions in Section 6, and across media in Section 7, uncovering interesting differences. We believe that web science methodologies can augment traditional manual analysis of onscreeen gender depictions by their online traces.

[1] Alexa ranking 50 (global), 28 (US) as of 25/11/2015.

[2] www.imdb.com/title/tt0118715

2 Background

Gender is a complex sociocultural phenomenon with a vast academic literature and we stress that this work makes limited exploration of gender itself. Instead we focus on some of the issues relating to gender in media as much as our data allows. Under-representation of women is a long-standing gender issue in media, both in terms of the gender of performers and also the subject matter, for example proportions of news stories that focus on females Wood, 1994. In that study Wood notes stereoptypical portrayals of hypermasculine, yet domestically incompetent, male characters and the female characters dependent on them, and complex relationships of power and image. This trend is confirmed in a more recent meta-study of articles in a special issue of the *Sex Roles* journal Collins, 2011, which adds to this observations about the role of race and interesting conjecture about the effect of under-representation and the importance of also finding positive representations of women in media.

Many gender media research questions require manual analysis. In their study of screen portrayals and media employment, Smith et al. consider 26,225 characters[3] from the 600 top-grossing films from 2007–2013 Smith *et al.*, 2014. They find a low percentage of female speaking characters – consistently around 30% over each year of their sample, and only 2% of films features more female than male characters. They also study sexualisation of female characters, finding them more likely to be shown in revealing clothing, nude or referred to as attractive. They note the dearth of female content creators, noting that the number of female writers and directors is at a six year low circa 2014. This extensive and detailed study is only made possible with a team of 71 highly-trained coders and to apply this depth of research at scale would be difficult and costly.

IMDb is an interesting source of data due to its size and popularity on the internet. Boyle notes that "IMDb has been the focus of surprisingly little academic attention" in her study of gender and movie reviews Boyle, 2014. This analyses how gender is expressed (or not) in textual reviews for three different films and the online profiles of the reviewers. Data from IMDb has been used for research in the natural language processing and computational linguistics domain, primarily as the source of a corpus of movie reviews annotated with sentiment Pang *et al.*, 2002. Ramakrishna et al. investigate the extent to which gendered language is used in different genres, finding that Action and Crime films contain more masculine language than films outside that genre, with converse findings for Romantic and Comedy films Ramakrishna *et al.*, 2015. Other resources for gender information have been gathered from the US Census and automatically processed web text Bergsma, 2005; Bergsma and Lin, 2006. A possible application for gender data is in coreference resolution Pradhan *et al.*, 2011, the task of clustering *mentions* that refer to the same entity in a document. For example, lists of male and female names may provide evidence whether the mentions he, Bob and manager should be matched together.

Detailed gender analyses of media are compelling yet difficult to conduct at scale. We aim to use metadata about screen media as a proxy for the original media to explore, albeit in a limited way, issues about gender and its onscreen representation. User-generated content can contain errors and omissions, so we focus on aggregations of the data, which make our analyses feasible. Web science methodologies that rely on processing large-scale data for exploratory analysis suggest useful starting points, for example using corpora of scanned books to examine culture Michel *et al.*, 2011 or serial numbers extracted from commercial web pages to study global distributions of products and people Talaika *et al.*, 2010. The dataset in this study allows us to study how people report onscreen media using the web, but this kind of data can also influence other media. Specifically, cast information is part of the ecosystem of media reporting, advertising, review and commentary, and this can have real-world impact. A study focussing on the dynamics of online film reviews found that it was their volume, rather than content or rating, that significantly impacts box office sales Duan *et al.*, 2009. The authors attribute this to an indicator of underlying word-of-mouth information flow and that online reviews spread awareness of the film. User data is increasingly being directly used to assist decisions about what media a studio should produce[4] and this is indicative of the complex relationship between onscreen media and the web.

3 Dataset and methods

Our methodology requires two simplifying assumptions. We assume that IMDb is a good proxy for onscreen entertainment, which we believe is a reasonable assumption for recent productions, but less so for older productions as we discuss below. We also assume that popular film and television is more likely to appear in a database like IMDb, and as such its aggregated content is a good estimator of what a random person would watch. Following from this, we ask the question: *"What are viewers likely to learn about roles and gender over time from onscreen entertainment?"*.

We downloaded the plain text data files `actors.list.gz` and `actresses.list.gz`[5] and applied several automated cleaning phases[6]. The files list the performer name, role name, and the titles, types and dates of productions they appear in. We exclude records typed as "credit only" since the performer would not be onscreen, and roles named themselves as we focus on individuals and want to avoid groups of performers. Where a performer is credited by another name (e.g., (as name)) we use this if a role name is missing. Additionally, if a performer is listed as herself, we use her name as the role name. We also remove markers of multiple similar roles: ordinal prefixes (e.g. first or 1st) from 1 to 5 and suffixes (e.g. (1) or (#1)). Any multi-role roles (e.g. model/actress) are split, generating one count for each lower-cased role. Finally, we generate one record per appearance, which may correspond to a film or television episode. Each record is typed into: film (including "straight to

[3] 4,506 of these were speaking roles.

[4] http://www.newyorker.com/business/currency/hollywoods-big-data-big-deal

[5] Accessed on 24/10/14 from http://www.imdb.com/interfaces.

[6] Code at https://github.com/wejradford/castminer.

video"), television and game. Video games are excluded from this analysis since they are relatively infrequent (although they are an increasingly important part of the media environment). We aggregate roles by year and calculate a gender distribution for each role r and year y. Specifically, $p(\mathrm{F}|r, y)$ is the count of records with role r in year y by a performer from the actresses list, normalised by the count of all r and y records.[7]

As with most user-generated content, there are a number of caveats that apply to the data and our analysis. It is possible that records are listed with incorrect years or that performers were misclassified and added to the wrong list file. To evaluate the impact of the latter, we randomly sampled 200 occurrences of actor-role pairs, and manually checked if the labeled gender of the actress or actor correspond to its true gender (using clues from the first name and Internet searches if required). We observed an error rate of 0.5% as we found only one misclassified entry and would expect this to be due to relatively infrequent data entry error. There is also a significant observation bias as, while it may be common for film and television to be listed as it enters production today, older productions are only listed if a user takes the effort to document them. As a result, older counts are susceptible to skew towards television productions with a strong internet-based community dedicated to listing each and every episode.

We do not further process roles and so some may be character names and others professions. We might expect that professions will have higher counts, as it is more likely that generic roles are repeated in many records than character names. This means that we are comparing names and roles, which is somewhat inelegant, but collecting main character roles would require linking to external structured (e.g. Freebase) or unstructured plot synopses (e.g. Wikipedia). One might use a Named Entity Linking system Hoffart *et al.*, 2011 or take advantage of other linked data resources. Another approach might be to filter roles using a list of known names, which may exclude ambiguous surnames such as Butcher, Baker and Pope. More problematic is that central characters typically have more time on screen, but are more likely to be credited with their name than their role, at least in fictional productions. Non-fictional television may be more likely to credit by role and we explore some differences between different media in Section 7.

For the majority of our analysis, we do not distinguish between the production country, which rules out potentially interesting national comparisons[8]. We also do not include any language processing: using stemming for instance host and hostess could be matched, although at the cost of conflating dissimilar concepts within or across languages. Finally, the role descriptions do not follow a fixed schema, so some equivalent role counts may be split by virtue of general synonymy (e.g. director and filmmaker) or different gender forms (e.g. policeman, policewoman, cop, police officer). This problem may be alleviated by mapping IMDb roles onto a semantic ontology such as WordNet Miller, 1995. While our approach is limited in some of these aspects, we believe that it is a pragmatic compromise.

[7] $p(\mathrm{M}|r, y) = 1 - p(\mathrm{F}|r, y)$.

[8] We restrict to productions by US companies for comparing to US employment statistics in Section 6.

4 Roles

After the automated preprocessing described above, we retain 18,224,054 role records from between 1900 and 2020 (Figure 2). The number of entries grows from the early 20th century and increase steadily until the 1990s, when the rate of growth increases. Note that, although the data was collected in 2014, there are records dated later than that, as IMDb lists ongoing and planned productions. We consider all film and television data for counts, but graphs do not show data after 2014 and, unless otherwise specified, are smoothed with a rolling mean with a 5-year window.

4.1 Role trends

The dataset allows us to track, at a very coarse level, what roles are popular in onscreen media and how has this changed over time. Table 1 shows the top 10 most common roles in 20 year periods from 1900. This shows how roles have changed over time and reflects what roles are reported and seen on screen. Initial roles from 1900 are most often undetermined or stock characters (mary, jack, the girl, the wife, daughter, husband). Roles from 1920-1940 are made up of dramatic roles that appear to be drawn from a crime or noir genre: henchman, policeman, detective. Others are ambiguous, as reporter and dancer could either be in a dramatic or actual role in a news broadcast or variety show. For the two decades from 1940, there seems to be a shift towards news broadcasting (i.e. newsreader, sports newsreader, weather forecaster), narration (i.e. announcer, narrator) and hosted television with host, singer and panelist. The trend of hosted television is maintained for the rest of the dataset, but we see evidence of shifts in trend: model from 1960–1980, additional voices for animated cartoons from 1980–2000, and finally reality television roles from 2000 (i.e. contestant, judge).

While the above analysis shows the enduring popularity of hosted screen entertainment, this can obscure some of the emerging roles through time. Table 2 shows, for the same period, which roles are new and did not appear in the top 50 roles of the previous period. The 1900s list is the same as Table 1 as this is the first period used. The 1920s sees different descriptions of underspecified roles (bit role vs undetermined role). There is a strong focus on hosted and news media from the 1940s and evidence of non-English-speaking entries (corresponsal is Spanish for correspondent).[9] From the 1960s, there is evidence of popular roles in children's television (member of the short circus from "The Electric Company"), television soap operas (paul williams, victor newman[10] from "The Young and the Restless"). Newly popular roles in the 1980s and 1990s included game and quiz shows (contestant, lexicographer from "Countdown Masters"), different television soap operas (ridge forrester from "The Bold and the Beautiful") and new terms (anchor and the gendered form co-hostess). Roles thusfar from the two decades from 2000 reflect the recent trend for

[9] In the absence of detailed language data and reliable translations, we consider these distinct roles.

[10] This character seems to first appear in 1980, so may be listed under an incorrect year. In lieu of canonical sources for "The Young and the Restless": http://en.wikipedia.org/wiki/Victor_Newman

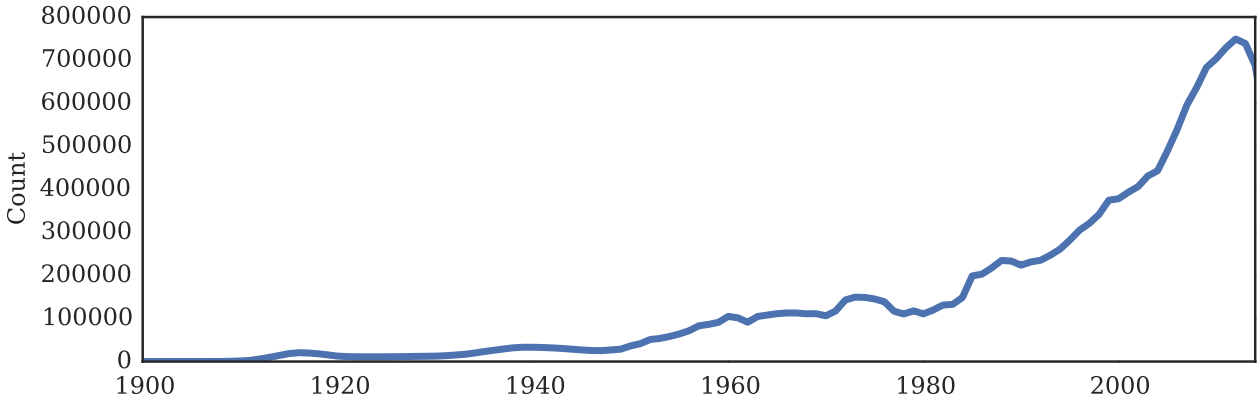


Figure 2: Count of roles over time.

1900-1920	1920-1940	1940-1960	1960-1980	1980-2000	2000-2020
undetermined role	minor role	newsreader	host	host	host
the wife	henchman	host	model	hostess	contestant
the husband	reporter	reporter	announcer	newsreader	narrator
mary	dancer	narrator	presenter	presenter	guest
the father	policeman	panelist	various	announcer	presenter
the girl	townsman	townsman	narrator	narrator	judge
jack	undetermined role	announcer	singer	guest	panelist
the sheriff	detective	sports newsreader	guest	various	various characters
the maid	party guest	singer	reporter	additional voices	co-host
the mother	waiter	weather forecaster	various characters	reporter	various

Table 1: Top 10 roles for 20 year periods from 1920.

1900-1920	1920-1940	1940-1960	1960-1980	1980-2000	2000-2020
undetermined role	henchman	newsreader	model	additional voices	zombie
the wife	reporter	host	various	contestant	housemate
the husband	dancer	panelist	various characters	musical director	police officer
mary	townsman	announcer	member of the short circus	lexicographer	alex
the father	waiter	sports newsreader	paul williams	anchor	interviewee
the girl	narrator	weather forecaster	victor newman	interviewer	laura
jack	barfly	corresponsal	brady black	ridge forrester	audience member
the sheriff	doctor	correspondent	jack abbott	emcee	david
the maid	bit role	presenter	roman brady	phil	sam
the mother	singer	sports reporter	george	co-hostess	bar patron

Table 2: Top 10 **newly popular** roles for 20 year periods from 1920.

zombies, which typically feature many unnammed zombie characters and thus has a large impact on the count data. We see a continued trend of more first-name roles (laura, david and the gender-ambiguous alex and sam[11]), and roles that reflect current naming conventions (police officer rather than policeman and bar patron rather than the earlier bar fly).

We propose that the dataset is an interesting way to explore how onscreen roles, and how they are referred to, change over time. We see evidence for a main hosted model of onscreen entertainment, with secondary trends, such as reality television. In older performances there seems also to be evidence of a skew towards television programmes that have been comprehensively documented, presumably by a dedicated internet-based community.

[11] These may be more frequent as both male and female performers can have those names as a role.

4.2 Role volatility

While this analysis shows when roles became popular, it does not answer questions about decreasing popularity. A related problem we studied was the identification of *volatile* roles: those roles that changed from popular to unpopular the most often. For this, we modified a popular tool to measure *bursty* features over time Kleinberg, 2002. The basic Markov model assumes that for a given year, one role is in one of two states: normal or bursty (over-represented). The model is parameterised with probabilities for emission of observed role frequencies and

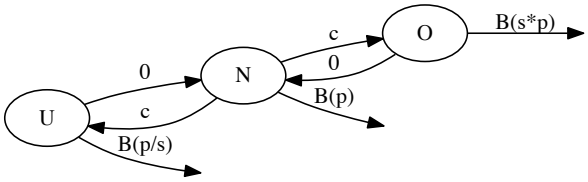


Figure 3: Our model to capture over and under-represented years.

transition between states. Dynamic programming is used to infer the most probable sequence of state assignments given observed frequencies. State sequences inferred in this way can capture modest changes observed for a long period, but ignore small fluctuations within a bursty period where it might be too costly to transition to another state.

The modification we propose also permits to model under-representation of roles, as well as over-representation. Figure 3 shows the model with states for generating roles in under-represented years (state U), in normal years (state N) and over-represented years (state O). Entering one of these abnormal years incurs a cost c, defined as in the original model as $\gamma \log(n)$ (we used $\gamma = 0.1$), while returning to the normal distribution is free.

The input is the relative frequency of role r over years ($[r_1, \ldots, r_n]$), and the assumption is that the distribution of the given role r for one year follows a binomial distribution. That is, it assumes that role r is generated with probability p, and therefore the probability of having k occurrences of role r if the total number of occurrences of all roles is d is shown in Equation 1.

$$\binom{d}{k} p^k (1-p)^{d-k} \tag{1}$$

The normal rate of emission of a role (p) is set to the proportion of that role over all role occurrences: by using probabilities and not absolute counts the model becomes immune to the increasing number of overall roles over years (Figure 2). This probability gets scaled by a parameter s for over-represented years, and scaled-down by s for under-represented years (we used $s = 2$).

Table 3 shows the roles that changed the most[12], which included roles such as kidnapper, pirate, headmaster and the German krankenschwester (i.e., nurse), which are hard to attribute to one period. On the other spectrum there are roles whose frequency changed radically, although only once. In our datasets these were zombie, boyfriend[13] and hipster which all had a sudden spike in recent years.

[12]Calculated over roles occurring more than 500 times in the period 1950–2014, excluding proper names.

[13]This may indicate more stories from the female point of view, so include a less-central boyfriend role.

Role	Changes
performer	22
krankenschwester	22
kidnapper	22
headmaster	21
mechanic	20
heckler	20
pirate	20
granny	20
resident	20
correspondent	20
guest host	20

Table 3: The most volatile roles from 1950–2014.

Role	F	Role	M
host	123688	host	369794
hostess	74766	narrator	75237
presenter	39538	announcer	58139
newsreader	34113	presenter	51686
model	30243	guest	45996
guest	29264	various	33488
contestant	28566	newsreader	32265
reporter	25837	various characters	31705
nurse	20765	contestant	31347
dancer	19004	reporter	31142
panelist	17801	panelist	25953
various	14372	judge	25000
judge	14110	co-host	22072
narrator	13609	doctor	18222
co-host	12226	additional voices	17981
various characters	12016	policeman	16546
girl	11548	performer	14872
singer	11488	man	13627
woman	11158	bartender	13259
waitress	11079	various roles	12520
correspondent	10685	singer	12430
mother	9959	correspondent	12343
laura	9917	dancer	12161
maria	9845	waiter	11839
performer	8481	police officer	11122
sarah	8190	cop	10715
lisa	8145	david	10071
anna	7962	student	10042
additional voices	7922	soldier	9999
co-hostess	7845	guard	9791
student	7587	detective	9685
mary	6949	paul	9302
rita	6888	tom	9149
rosa	6708	sports newsreader	9070
alice	6707	john	9016
jane	5990	jack	8943
various roles	5919	commentator	8780
julie	5782	townsman	8514
secretary	5668	mike	8491
sara	5539	max	8448
linda	5427	extra	8313
receptionist	5398	boy	8253
extra	5190	frank	8245
eva	5112	mark	8037
marta	5007	tony	7916
jenny	4963	george	7889
sandra	4963	sam	7800
lucy	4912	interviewee	7787
ana	4849	musician	7785
teresa	4803	joe	7759

Table 4: The 50 most frequent roles by gender.

5 Gender

One of the most valuable characteristics of the dataset is that each performer has gender information. Aggregating by role allows us to consider biases of the gender of onscreen roles. Figure 4 shows how roles over time are split between two genders, with counts for each gender and also the proportion of female roles ($p(F)$). From 1940, we see a gradual increase in the proportion of roles played by female actors from 0.25 to 0.4. Before this period, total counts are somewhat lower, so it is difficult to draw conclusions.

Table 4 shows the 50 most frequent roles per gender. Of

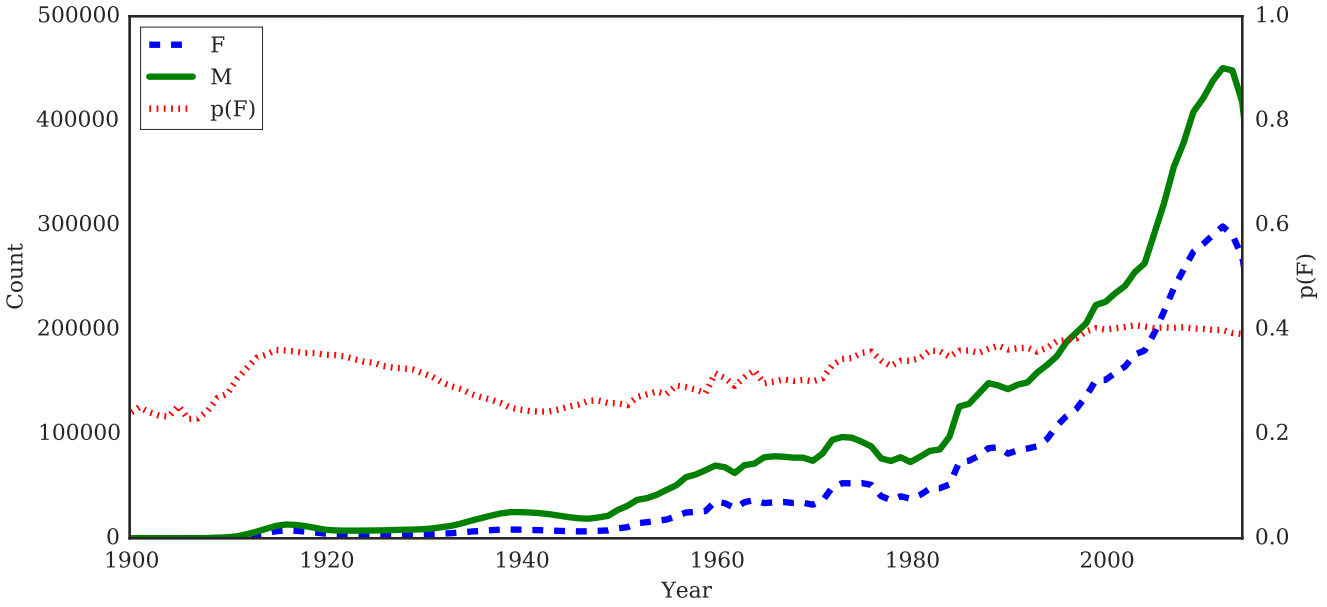


Figure 4: Count of roles from each gender over time, as well as the gender distribution $p(\mathrm{F})$.

Strongly male		Moderately male		Gender neutral		Moderately female		Strongly female	
Role	$p(F)$	Role	$p(F)$	Role	$p(F)$	Role	$p(F)$	Role	$p(F)$
general	0.01	athlete	0.20	obstetrician	0.42	dancer	0.61	international reporter	0.81
priest	0.01	comedian	0.20	orphan	0.42	shopper	0.61	mannequin	0.81
thug	0.01	school student	0.21	student	0.43	office assistant	0.61	stenographer	0.84
truck driver	0.01	servant	0.23	violin	0.43	computer voice	0.63	lexicographer	0.85
rapist	0.02	factory worker	0.23	art student	0.44	nutritionist	0.63	switchboard operator	0.85
referee	0.03	rebel	0.23	cafe patron	0.44	recepcionista	0.64	gossip	0.86
u.s. soldier	0.03	psychiatrist	0.24	swimmer	0.45	personal finance expert	0.65	doll	0.87
attorney general	0.04	lecturer	0.24	margaret thatcher	0.45	autograph seeker	0.65	receptionist	0.88
cop	0.05	scout	0.25	reporter	0.45	computer	0.65	legal analyst	0.88
pirate	0.05	teenager	0.29	victim	0.47	democratic strategist	0.66	flight attendant	0.89
terrorist	0.06	paranormal investigator	0.29	mourner	0.47	interior designer	0.67	witch	0.89
thief	0.06	translator	0.31	singer	0.48	psychic	0.70	stripper	0.89
detective	0.06	casino patron	0.32	schoolchild	0.48	ballet dancer	0.71	dr. quinn	0.91
gambler	0.07	hospital patient	0.33	church member	0.48	librarian	0.72	telephone operator	0.93
director	0.07	hitchhiker	0.34	production manager	0.49	schoolteacher	0.73	cheerleader	0.93
stranger	0.10	zombie	0.35	hostage	0.50	fortune teller	0.75	nurse	0.94
doctor	0.13	geophysics	0.35	sports anchor	0.50	the secretary	0.75	prostitute	0.95
ninja	0.14	winner	0.35	escort	0.54	regional newsreader	0.77	blonde	0.95
lawyer	0.15	vampire	0.36	nudist	0.58	angela merkel	0.77	belly dancer	0.96
paramedic	0.15	baseball fan	0.36	hotel receptionist	0.58	social worker	0.78	courtesan	0.97
alien	0.17	researcher	0.38	therapist	0.59	politics reporter	0.79	pageant contestant	0.97
editor-in-chief	0.18	sports reporter	0.39	cashier	0.59	psychotherapist	0.79	maid	0.98

Table 5: Examples of common roles with different gender distributions.

course, some of the roles of Table 1 appear again here, but it is already possible to see biases towards one of the genders. model and receptionist are frequent roles which are mostly female, as are hostess, girl, woman, waitress and mother, together with a series of frequent female first names. On the male side side, there seems to be strong bias for narrator, announcer, doctor, detective, bartender together with a series of security or military roles (police officer, cop, soldier, guard), and again some gender-specific roles like policeman, man and waiter.

We can also analyse the gender distribution of common roles to characterise how gender relates to roles at a high level. As an example, we filtered the most common mentions with an overall count above 100[14], and partitioned them into five bins according to their gender distribution (from $p(F)$ between 0 and 0.2, between 0.2 and 0.4 and so on). In Table 5 we show some of these roles. maid and receptionist are frequent roles which are mostly female, as are belly dancer, stripper and cheerleader. On the male side side, there seems to be strong bias for referee, doctor and lawyer; together with some criminal or negative roles (rapist, terrorist, thief, thug and a series of security or military roles (u.s. soldier, cop, general). Note also how psychiatrist is moderateley male, therapist is gender neutral and psychotherapist is moderately female. While psychic is moderately female, paranormal investigator is moderately male. As gender neutral, we can find swimmer, student, church member and obstetrician. We see some surprising entries for female politicians: margaret thatcher as gender neutral and angela merkel as moderately female, both due to male performers in satirical productions[15]. Note how computer and computer voice are moderately female, which we discuss below.

[14]This threshold is chosen empirically.

[15]Margaret Thatcher was frequently played by the male actor Steve Nallon in the puppet-show Spitting Image.

Profession	Keywords	$p(F)$
IT	software, computer, hacker	0.40
Doctor	medical, dr, doctor surgeon, psychiatrist	0.28
Corporate	corporate, ceo, coo	0.34
Law	prosecutor, lawyer	0.15
Politics	minister, dictator, parliament senator, president	0.11
Science	science, professor	0.13
Religion	priest, priestess, reverend pastor, prior, allamah imam, rabbi, guru, lama	0.15
Engineering	bishop, ayatollah, swami engineer	0.05

Table 6: Gender distribution grouped by profession.

Country	%	Country	%
None	52.7	Mexico	1.1
United States	23.8	Italy	0.8
Great Britain	4.1	Sweden	0.6
Germany	1.8	India	0.5
Australia	1.7	Finland	0.5
France	1.7	Netherlands	0.4
Spain	1.6	Poland	0.3
Canada	1.5	Chile	0.3
Philippines	1.3	Argentina	0.3
Japan	1.3	Denmark	0.3

Table 7: The distribution of countries.

In Smith *et al.*, 2014, the authors analyze 120 movies and show strong biases in the representation of executive roles. Inspired by that report, we looked for key roles in areas such as law, IT and religion and looked at the aggregated count of male and female actors in these roles. For each keyword listed in Table 6, we looked for all roles that contained that word. We made exceptions for president where we looked only for exact matches, and bishop where we ignored those mentions that end with it to avoid including surnames.

Legal professions had around 15% female representation, which coincides with the values reported in Smith *et al.*, 2014, while the medical domain (doctors) had a female probability of 0.28. In contrast to the results in Smith *et al.*, 2014, Religion does not score at the bottom with regards to female presentation (although very low with 0.15). From the professions we selected, Engineering was the lowest (0.05). The highest scoring profession was IT (0.40), which is partly due to the fact that many computer voices were female (the probability that a female plays a computer is 0.65; and enterprise computer from "Star Trek" was almost exclusively female). Strong conclusions are hard to draw from this analysis since we manually selected roles, for example omitting judge, which could apply to law or entertainment.

We can also examine role gender over time, searching for qualitative evidence that the gender associated with a specific role changes. Figure 5 shows the distribution of nurses, where we matched any role containing the query term. Onscreen nurses have been traditionally almost uniformly female until the 1990s and now one in five nurses are played by male performers.

6 Reality

Our analyses to this point have only referenced IMDb data, but it is also interesting to examine how onscreen gender distributions compare with their real-world counterparts. The US Bureau of Labor Statistics publishes yearly estimates of its Occupational Employment Statistics (OES), and we accessed, parsed and unified that data from 2002 until 2014 United States Department of Labor, 2015, requiring some manual effort to aggregate role descriptions that had changed over the years[16]. To fairly compare against IMDb data, we assign a country to each record using the country of its production company.[17] Table 7 shows the 20 most frequent values across all records. Roughly half of all records lack country information and the next most popular country is the United States (`us`), and we use only this subset for the analysis in this section. The two main issues with this mapping are that the incompleteness means we reject a large part of the data, and that multi-country productions are unaccounted for.

Figure 6 shows how onscreen gender distributions map to those listed in the OES. In both cases, the data was restricted to 2014. Intuitively, points on the diagonal line have an onscreen portrayal consistent with the OES distributions. If a point is above the line (e.g. reporter), then those roles are over-represented onscreen by female performers. Conversely, points below the line suggest an under-representation onscreen by female performers. For example, surgeons, teachers and nurses are played more frequently by male performers than their OES counterparts.

We also looked at evolution of gender representation of specific roles over time. Figure 7 shows how the distributions in IMDb and OES data vary over time. For some roles, the female representation onscreen consistently underperforms reality (surgeon, teacher). We see a divergence in representation for others where female reporters are increasingly over-represented onscreen, and vice-versa for nurses.

There are several limitations of this analysis that should be taken into account before drawing strong conclusions. Comparing user-generated roles with strict OES roles introduces bias since we selected the mapping and selected roles. Linking roles from the different sources to a common ontology would present a useful way to reduce manual effort in this step. We were not able to retrieve OES data for 2001 and 2002, and the ontology used changed significantly after 2003, reason for which we only considered census data from that year onward. Overall, this analysis allows us to draw an interesting exploratory counterpoint between onscreen gender representation and real-world figures.

[16] These mappings are included in the code release.

[17] Found in `production-companies.list.gz`.

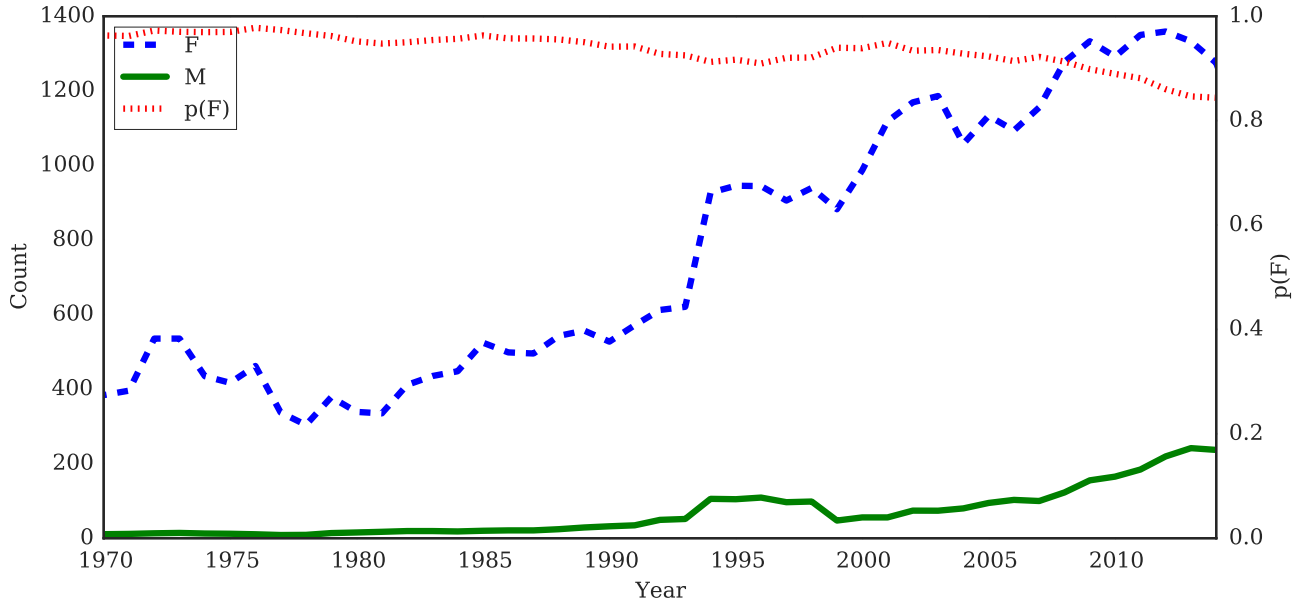


Figure 5: Gender counts and proportions over time for nurse.

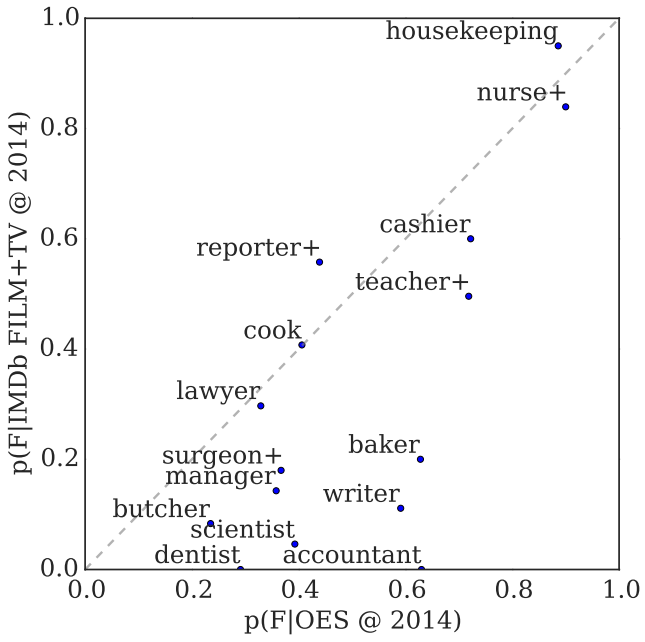


Figure 6: Proportion of female in IMDb and OES. + indicates significant at $p < 0.05$ in a two-tailed, two-proportion Z-test.

7 Media

The analysis above does not distinguish between the different types of media that are covered by IMDb. In this section, we investigate how role and gender varies on film and television. Figure 8 shows the counts over time of datapoints from a film or a television screening. Film has a longer history, whereas television is a more recent phenomenon with a faster growth, presumably due to its relatively cheaper production costs. The proportion of female roles is also different: during the 1960s and 1970s, female performers were under-represented, but independently of the medium on which they appeared. However, since the mid-1980s, the trends have diverged and, while both have increased, a higher proportion of roles are played by females on television than on film.

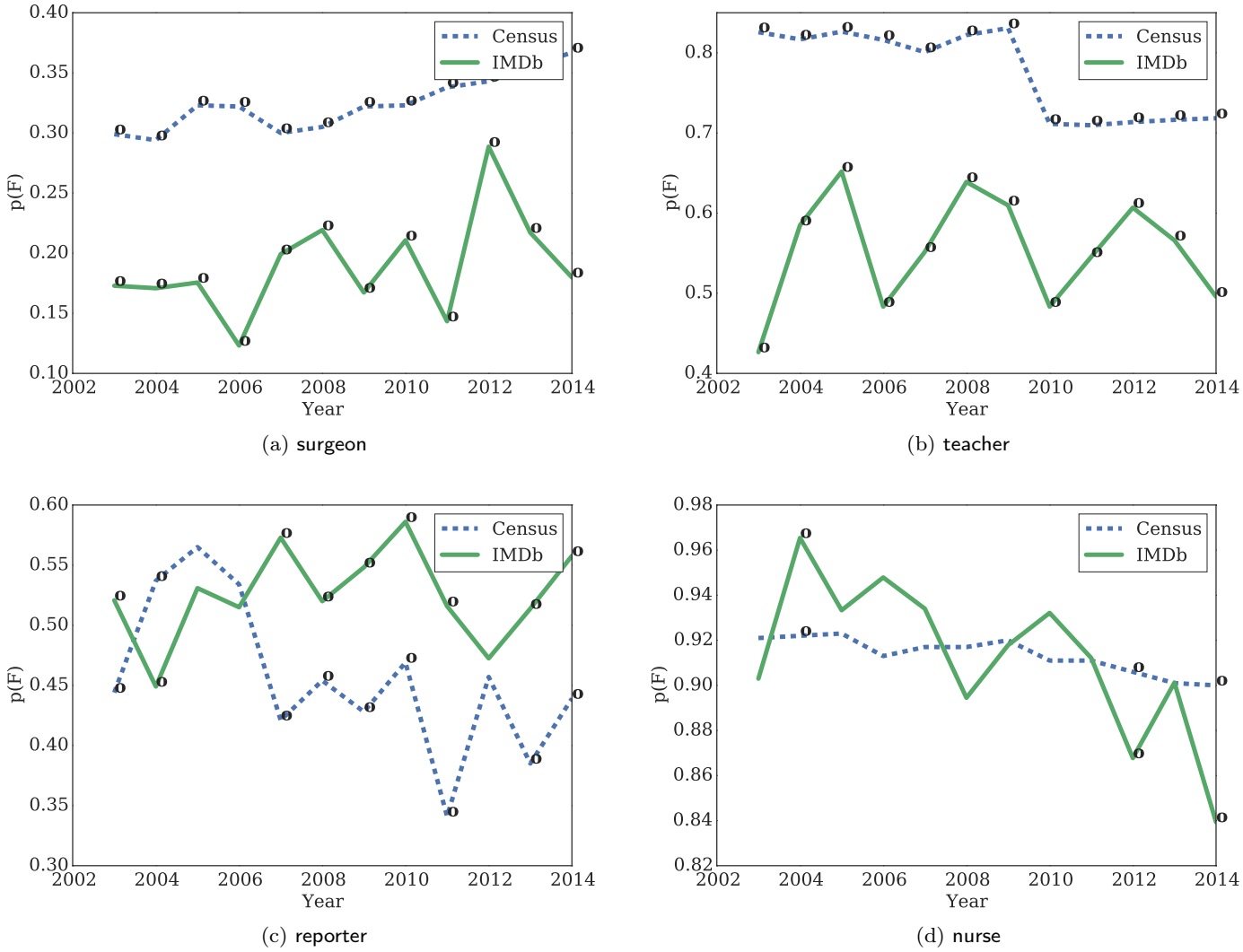
We are also interested in how roles evolve over time, and how this relates to the different media. In general, for a given time-step, we calculate a distribution over individual roles (P_t). This can then be compared to the distribution at the next time-step (P_{t+1}). We calculate the Bhattacharyya distance[18] Bhattacharyya, 1943, as specified in Equation 2, between each year.

$$H(P_t, P_{t+1}) = \frac{1}{\sqrt{2}} \| \sqrt{P_t} - \sqrt{P_{t+1}} \| \quad (2)$$

Figure 9 shows the trend in inter-year distance for film and television role distributions. The first thing to note is that there is usually a large inter-role-distribution distance between years. This diversity is declining over time, such that a year's role distribution is more similar to the previous year in 2013 than it was in 1960. We also observe that diversity is decreasing faster for film than television. One possible reason for this is that larger film production costs mean that producers are more conservative, preferring roles that are more established.

Finally, we examine how gendered roles distribute across film and television. Table 8 shows the popular roles for male and female performers in film and television. Separating by medium reveals that differences in film seem to be more pro-

[18] or Hellinger distance.



(a) surgeon (b) teacher

(c) reporter (d) nurse

Figure 7: Gender distribution in IMDb and OES over time. **o** indicates significant at $p < 0.05$ using the two-tailed, two-proportion Z-test. Note that these do not use a rolling mean.

Film				Television			
Role	F	Role	M	Role	F	Role	M
dancer	10774	narrator	20776	host	115039	host	353585
nurse	9066	host	16210	hostess	72736	announcer	55780
host	8647	policeman	9975	presenter	37633	narrator	54461
mother	7090	doctor	9613	newsreader	33796	presenter	48812
girl	7022	reporter	8750	model	27845	guest	42886
waitress	6120	bartender	7777	contestant	27839	newsreader	31928
woman	5766	man	7517	guest	27322	various	31009
student	4850	extra	7216	reporter	22150	contestant	30639
extra	4697	dancer	6884	panelist	17374	various characters	30189
maria	4515	zombie	6810	various	13322	panelist	25272
anna	4360	soldier	6750	judge	13149	reporter	22392
sarah	4200	waiter	6507	nurse	11699	co-host	21312
narrator	4090	cop	6336	co-host	11549	judge	21144
mary	4005	police officer	6312	various characters	11288	performer	13794
reporter	3686	student	6302	correspondent	10396	additional voices	13450
zombie	3665	henchman	6094	narrator	9518	correspondent	12008
party guest	3367	john	5621	singer	8574	various roles	11182
laura	3334	detective	5557	dancer	8229	singer	9475
lisa	3238	boy	5395	performer	7868	sports newsreader	9064
singer	2913	father	5332	co-hostess	7808	doctor	8608

Table 8: The 20 most frequent female and male roles across film and television.

nounced than television. There are fewer roles common to both genders in film than in television. The former is composed of stereotypically (e.g., nurse, soldier) or explicitly gendered roles (e.g., policeman, waitress), while the latter is more balanced with both males and females in common television roles.

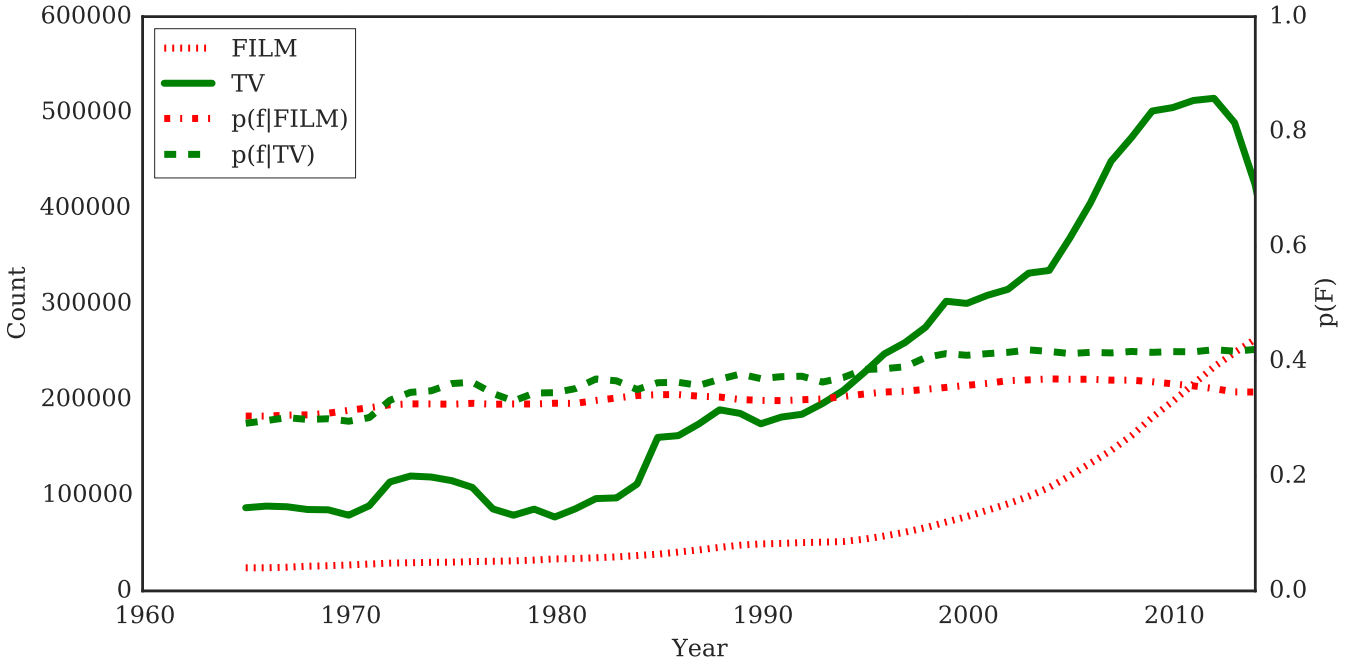


Figure 8: Count and proportion of female roles in film and television.

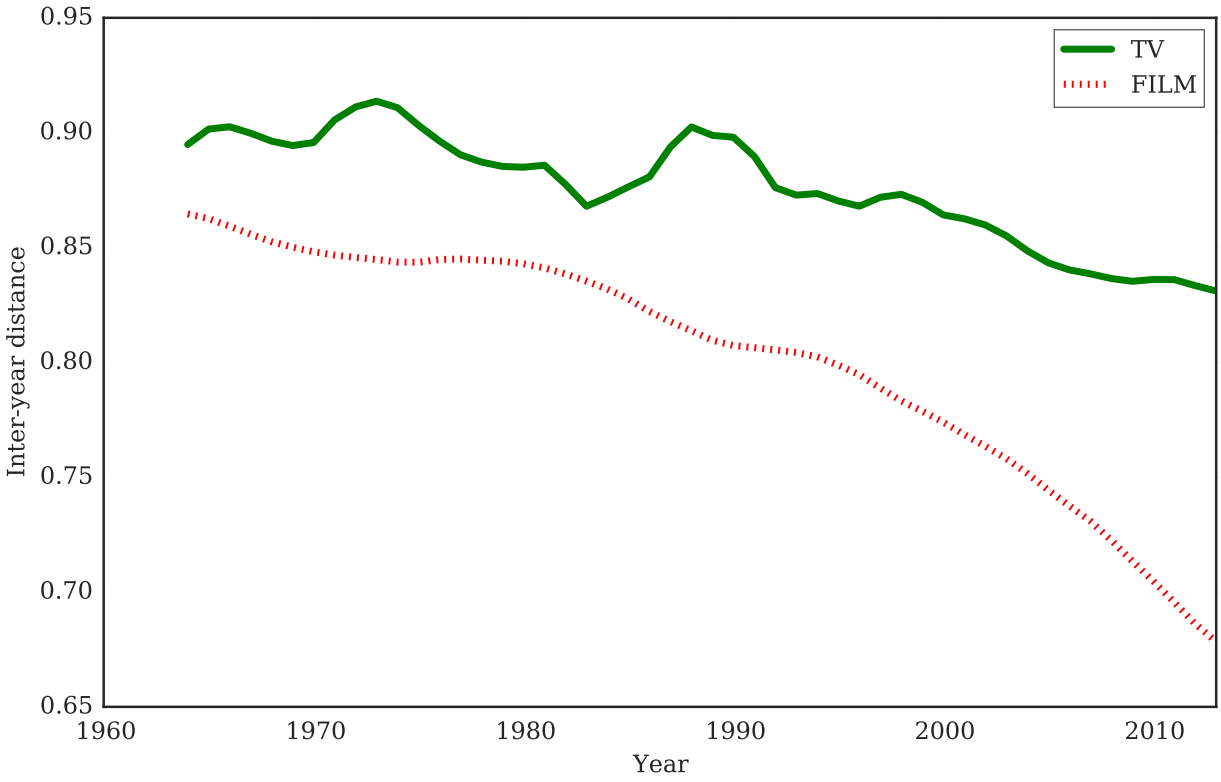


Figure 9: Year-on-year difference between role distributions for film and television.

8 Discussion

The methodology presented in this work has allowed us to study a large, long-term collection of user-generated web content to help answer questions about gender representation. Here we present improvements specific to this study, then discuss the methodology more generally. The first aspects of future work relate to extracting less noisy data and adding more dimensions for analysis. This includes linguistic analysis to aggregate role synonyms (e.g. filmmaker, director), many of which are multi-word expressions. Including genre information may reveal interesting disparities on the gender proportion in them. The production company countries identified in Section 6 may help identify the language of the production, but this may have to be inferred to some extent.

Other future work is more directed to limitations of the dataset itself, and its ability to shed light on the questions we have. Our current model emphasizes the importance of secondary characters and treats them equally. Extracting their roles from other data sources such as plot summaries or reviews would allow us to include major character roles and may motivate a "central role" weighting scheme. If we combined this with an accurate importance metric for a character within a production, we could move beyond our simplifying assumption that cast membership is equivalent to on-screen time. However, it should be noted that focusing on secondary roles has the advantage of focusing the analysis on the underlying distribution of roles, decisions which may not be taken consciously. Obviously, a lot of variables are discussed when deciding on who to cast as the major roles: decisions on secondary roles are made much quicker and can therefore convey better what are the assumptions on how the world works, as well as conveying that message to the audience.

User-generated content is inherently noisy, and so finding good ways to compare this against external datasets for reference and validation is critical from a web science standpoint. We provide exploratory analysis in Figure 6, but further analysis would require matching the informal IMDb and formal OES role ontologies. This depends on mapping between schemas from the noisy IMDb role descriptions and a more structured OES role ontology. We proposed some manual mappings but – while challenging – linguistic analysis could be used to recognize role synonyms and semi-automatically generate those mappings.

We believe our main contribution is a demonstration of how a combination of natural language processing and data mining techniques can be used on top of large-scale user-generated content to provide insights into questions of societal value. As an example, Table 5 provides some very clear insights about which roles are generally portrayed by which gender, and poses obvious questions on the impact this may have on the viewers, most notably on younger generations. Our comparison with the work of Smith et al Smith *et al.*, 2014 shows some interesting agreement and disagreement in gender representations of roles in different areas. The latter can either be attributed to the different focus (major role vs secondary role), quality of data (manually annotated vs massive user generated) or scale; stressing the importance of having complementary methodologies. Having decided our research questions, we have tried to automate the data-extraction process as much as possible. The released code allows to re-run the experiments on newer snapshots of IMDb and OES, assuming their formats do not change substantially. While we have found great value in the IMDb data, it is less obviously a source of user-generated content than Wikipedia or Twitter. However, its combination of long timespan and scale make it compelling for analysis, and we hope this encourages more work on more obscure, but large-scale user-generate content.

9 Conclusion

This paper presents methodologies for mining information about onscreen media gender from cast lists. We use 18 million actor-role pairs over more than 50 years to study temporal evolution of roles, their gender and media distribution and their relationship to roles in the real world. Despite the noise inherent in user-generated data, we assert that large-scale screen production metadata is a useful proxy for framing and answering questions about the evolution of roles over time, and how gender balances evolve. We propose that these methodologies make for a compelling adjunct to traditional manual analyses and can help study how onscreen media is reflected onto the web, and eventually, how the web influences onscreen media.

References

Bergsma, S. (2005). "Automatic Acquisition of Gender Information for Anaphora Resolution". In: *Proceedings of the 18th Conference of the Canadian Society for Computational Studies of Intelligence (Canadian AI'2005)*. Victoria, B.C., Canada. 342–353.

Bergsma, S. and D. Lin (July 2006). "Bootstrapping Path-Based Pronoun Resolution". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics. 33–40. DOI: 10.3115/1220175.1220180. URL: http://www.aclweb.org/anthology/P06-1005.

Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distributions". *Bulletin of the Calcutta Mathematical Society*. 35: 99–109.

Boyle, K. (May 2014). "Gender, comedy and reviewing culture on the Internet Movie Database". *Participations: Journal of Audience & Reception Studies*. 11(1): 31–49. URL: www.participations.org/Volume%2011/Issue%201/3.pdf.

Collins, R. L. (2011). "Content Analysis of Gender Roles in Media: Where Are We Now and Where Should We Go?" *Sex Roles*. 64(3–4): 290–298. URL: http://rd.springer.com/article/10.1007/s11199-010-9929-5.

Duan, W., B. Gu, and A. B. Whinston (Oct. 1, 2009). "Do online reviews matter? - An empirical investigation of panel data." *Decision Support Systems*. 45(4): 1007–1016. URL: http://dblp.uni-trier.de/db/journals/dss/dss45.html#DuanGW08.

Hoffart, J., M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum (July 2011). "Robust Disambiguation of Named Entities in Text". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics. 782–792. URL: http://www.aclweb.org/anthology/D11-1072.

Kleinberg, J. (2002). "Bursty and Hierarchical Structure in Streams". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '02*. Edmonton, Alberta, Canada: ACM. 91–101. ISBN: 1-58113-567-X. DOI: 10.1145/775047.775061. URL: http://doi.acm.org/10.1145/775047.775061.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden (2011). "Quantitative Analysis of Culture Using Millions of Digitized Books". *Science*. 331(6014): 176–182. DOI: 10.1126/science.1199644. eprint: http://www.sciencemag.org/content/331/6014/176.full.pdf. URL: http://www.sciencemag.org/content/331/6014/176.abstract.

Miller, G. A. (1995). "WordNet: A Lexical Database for English". *Communications of the ACM*. 38: 39–41.

Pang, B., L. Lee, and S. Vaithyanathan (July 2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 79–86. DOI: 10.3115/1118693.1118704. URL: http://www.aclweb.org/anthology/W02-1011.

Pradhan, S., L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue (June 2011). "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. Portland, Oregon, USA: Association for Computational Linguistics. 1–27. URL: http://www.aclweb.org/anthology/W11-1901.

Ramakrishna, A., N. Malandrakis, E. Staruk, and S. Narayanan (Sept. 2015). "A quantitative analysis of gender differences in movies using psycholinguistic normatives". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics. 1996–2001. URL: http://aclweb.org/anthology/D15-1234.

Smith, S. L., M. Choueiti, and K. Pieper (2014). "Gender Inequality in Popular Films: Examining On Screen Portrayals and Behind-the-Scenes Employment Patterns in Motion Pictures Released between 2007-2013". http://annenberg.usc.edu/pages/~/media/MDSCI/Gender_Inequality_in_500_Popular_Films_-_Smith_2013.ashx. Accessed: 22/1/15.

Talaika, A., J. Biega, A. Amarilli, and F. M. Suchanek (2010). "IBEX: Harvesting Entities from the Web Using Unique Identifiers". In: *Proceedings of the 18th International Workshop on Web and Databases. WebDB'15*. Melbourne, VIC, Australia: ACM. 13–19. ISBN: 978-1-4503-3627-7. DOI: 10.1145/2767109.2767116. URL: http://doi.acm.org/10.1145/2767109.2767116.

United States Department of Labor (2015). "Occupational Employment Statistics". accessed July 2015, www.bls.gov/oes/.

Wood, J. T. (1994). "Gendered Media: The Influence of Media on Views of Gender". In: *Gendered Lives: Communication, Gender and Culture*. Cengage Learning. Chap. 9. 231–244.