

Tracking onscreen gender and role bias over time

Will Radford

Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
will.radford@xrce.xerox.com

Matthias Gallé

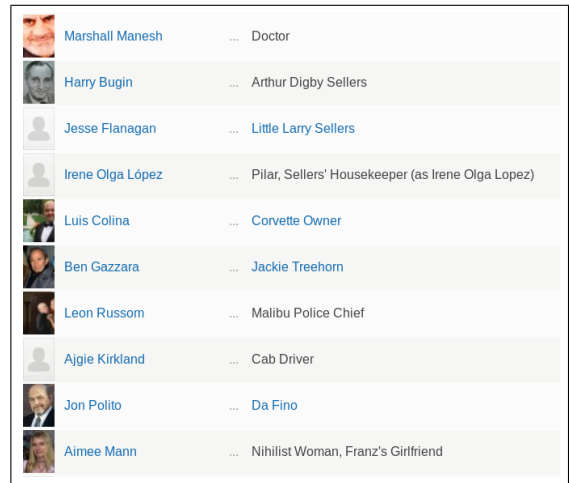
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
matthias.galle@xrce.xerox.com

Abstract

Film and television play an important role in popular culture. However studies that require watching and annotating video are time-consuming and expensive to run at scale. We explore information mined from media database cast lists to explore the evolution of different roles over time. 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. Finally, we investigate how gender role biases differ between film and television. We propose these methodologies are a useful adjunct to traditional analysis that allow researchers to explore the relationship between online and onscreen gender depictions.

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 issues. 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. For example, box office figures do not allow detailed analysis of cinematography.













	Marshall Manesh	... Doctor
	Harry Bugin	... Arthur Digby Sellers
	Jesse Flanagan	... Little Larry Sellers
	Irene Olga López	... Pilar, Sellers' Housekeeper (as Irene Olga Lopez)
	Luis Colina	... Corvette Owner
	Ben Gazzara	... Jackie Treehorn
	Leon Russom	... Malibu Police Chief
	Aigie Kirkland	... Cab Driver
	Jon Polito	... Da Fino
	Aimee Mann	... Nihilist Woman, Franz's Girlfriend

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

Our research question is whether web science can provide viable proxies that let us answer interesting social science research questions at scale. 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 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 three factors from the data: productions are listed with their release date, male and female performers are distinguished in the data,

¹Alexa ranking 49 (global), 24 (US) as of 22/1/15.

²www.imdb.com/title/tt0118715

and unnamed characters are usually listed by their role or profession. This lets us 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 data³. We then explore what roles are found onscreen and how they change over time in Section 4. In Section 5, we examine how roles interact with gender over time and how this compares to real-world gender distributions in Section 6. Finally, in Section 7, we investigate how gender roles vary with the medium on which they are displayed. We believe that web science methodologies can augment traditional manual analysis to enable comparison of online and onscreen gender depictions.

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 [14]. Moreover, Wood notes stereotypical 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 [5], 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 of gender media research questions require manual analysis. In their study of screen portrayals and media employment, Smith et al. consider 26 225 characters⁴ from the 600 top-grossing films from 2007–2013 [12]. 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 student 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 [4]. This consisted of analysing 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 [10]. Other resources for gender information have been gathered from the US Census and automatically processed web text [1, 2]. A possible application for gender data is in coreference resolution [11], 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 hope 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. Web science methodologies, such as those used to study scanned books [8], suggest useful starting points. 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 volume significantly impacts box office sales, rather than content and ratings [6]. The authors attribute this to an indicator of underlying word-of-mouth information

³Code at <https://github.com/wejradford/castminer>

⁴4506 of these were speaking roles.

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 produce⁵ 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`⁶ and applied several cleaning phases. 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. 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, television and game, where film also includes “straight to video” films. We aggregate roles by year and calculate a gender distribution for each role r and year y . Specifically, $p(F|r, y)$ is the count of records with role r in year y by a performer from the ac-

tresses 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 performers can be misclassified and added to the wrong list file, or records listed with incorrect years. We would expect this to be the result of data entry error and focus our analysis on those with higher count, as to avoid this hopefully rare occurrence. 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 distinguish between the production country, which rules out potentially interesting national comparisons and language processing. 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 extracting roles for main characters would require linking to external structured (e.g. Freebase) or unstructured plot data (e.g. Wikipedia). Moreover, central characters are more important, but it’s not immediately clear how to weight their influence so we believe that our approach is a pragmatic compromise. If we were able to map to media country, the language-dependent processing would be possible. This might include mapping `host` and `hostess` using stemming, but this comes 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 [9].

⁵<http://www.newyorker.com/business/currency/hollywoods-big-data-big-deal>

⁶Accessed on 24/10/14 from <http://www.imdb.com/interfaces>.

⁷ $p(M|r, y) = 1 - p(F|r, y)$.

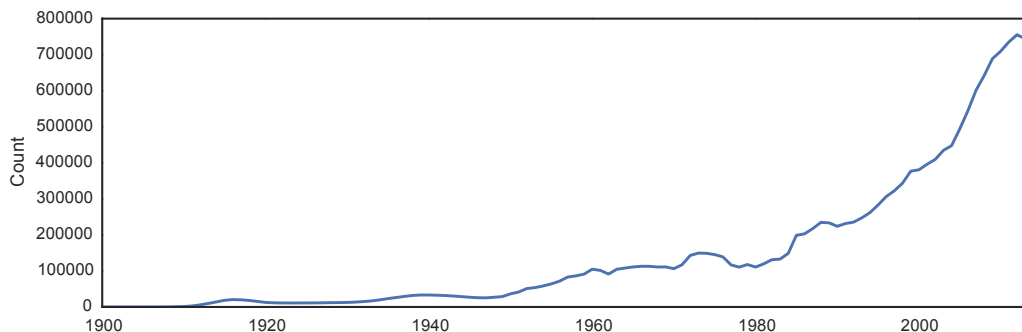


Figure 2: Count of roles over time.

4 Roles

After preprocessing, we retain 18 318 564 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 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 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*). 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*⁸ 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 reflects the recent trend for *zombies*, which typically feature many unnamed zombie characters and thus has a large impact on the count data. We see

⁸This 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

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	correspondent	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.

a continued trend of more first-name roles (*laura*, *david* and the gender-ambiguous *alex* and *sam*), 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 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.

4.2 Role volatility

While this analysis shows when roles became popular, it does not answer questions about decreasing popularity. A related question we studied was which consistently present, but *volatile* roles over time, or which roles changed from popular to unpopular the most often. For this, we modified a popular tool to measure *bursty* features

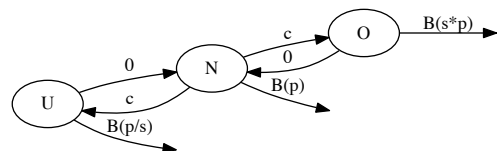


Figure 3: Our model to capture over and under-represented years.

over time [7]. Our modification also permits to model under-representation of roles, as well as over-representation. The input is the relative frequency of role r over years $([r_1, \dots, r_n])$, and the assumption is that the distribution of the given role r for one year follows a binomial distribution. This 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.

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: Most volatile roles in the period 1950–2014

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

This has the additional advantage of being immune to the increasing number of overall roles over years (Figure 2). The normal rate of emission of a role (p) is set to the proportion of that role over all role occurrences, and this probability gets scaled by a parameter s for over-represented years, and scaled-down by s for under-represented years. We model these option as a three state automaton, with Markovian assumptions (see Figure 3), generating respectively 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 in a cost c , defined as in the original model as $\gamma \log(n)$, while returning to the normal distribution is free. For a given role r an optimal state-sequence can then be computed using a standard dynamic algorithm.

This sequence allows a more fine-grained analysis of bursty periods, as it discovers specific periods where a role became more interesting with respect to its base distribution. Here we count the number of times a given path changes state, representing therefore the most variable roles in our dataset. Table 3 shows the roles that changed the most⁹, 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**¹⁰ and **hipster** which all had a sudden spike in recent years.

5 Gender

One of the most valuable characteristics of our 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 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**, **boy**, **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, 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 **psychoterapist** is moderately female. While **psychic**

⁹Calculated over roles occurring more than 500 times in the period 1950–2014, excluding proper names.

¹⁰This may indicate more stories from the female point of view, so include a less-central **boyfriend** role.

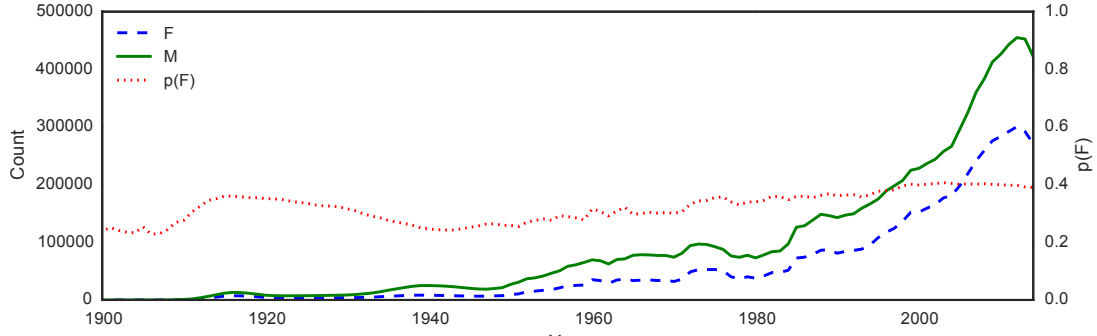


Figure 4: Count of roles from each gender over time, as well as the gender distribution $p(F)$.

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	receptionista	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.

are moderately female, **paranormal investigator** are moderately male. As gender neutral, we can find **swimmer**, **student**, **church member** and **obstetrician**, as well as **margaret thatcher** (but **angela merkel** is moderately female). Note how **computer** and **computer voice** are moderately female.

In [12], 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 actor 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.

Law and corporate professions had around 15% of female representation, which coincides with the

values reported in [12] for Law but not for corporate professions, while the medical domain (doctors) had a female probability of 0.28. In contrast to the results in [12], 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).

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 two roles, where we matched any role containing the query term. On-screen **nurses** have been traditionally almost uniformly female until the 1990s and now one in five

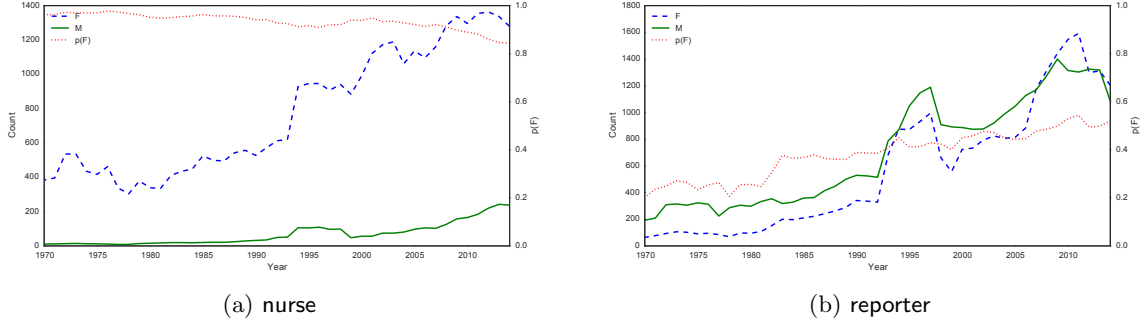


Figure 5: Gender counts and proportions over time for various roles.

Role	F	Role	M
host	123695	host	369824
hostess	74768	narrator	75650
presenter	39540	announcer	58341
newsreader	34114	presenter	51707
model	30282	guest	45998
guest	29263	various	33831
contestant	28565	newsreader	32267
reporter	25858	various characters	31774
nurse	20787	contestant	31349
dancer	19008	reporter	31162
panelist	17801	panelist	25953
various	14491	judge	25012
judge	14114	additional voices	22893
narrator	13700	co-host	22073
co-host	12227	doctor	18248
various characters	12043	policeman	16570
girl	11565	performer	14871
singer	11495	man	13643
woman	11176	bartender	13284
waitress	11094	various roles	12522
correspondent	10686	singer	12439
mother	9983	correspondent	12346
laura	9931	dancer	12163
maria	9871	waiter	11847
additional voices	9648	police officer	11149
performer	8482	cop	10772
sarah	8212	soldier	10168
lisa	8162	david	10078
anna	7977	student	10043
co-hostess	7844	guard	9892
student	7591	detective	9692
mary	6958	paul	9306
rita	6898	tom	9161
alice	6723	sports newsreader	9070
rosa	6719	john	8927
jane	6009	jack	8869
various roles	5921	commentator	8858
julie	5785	townsman	8521
secretary	5682	mike	8508
sara	5548	max	8489
linda	5434	extra	8342
receptionist	5402	frank	8264
extra	5215	boy	8263
eva	5127	mark	8045
marta	5009	tony	7928
jenny	4976	george	7896
sandra	4963	sam	7834
lucy	4918	musician	7793
ana	4857	interviewee	7788
teresa	4809	joe	7778

Table 4: The 50 most frequent female and male roles.

nurses are played by male performers. Conversely, the initial low proportion of onscreen female reporters has risen and the proportion is now relatively even.

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

Table 6: Gender distribution grouped by profession.

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 1995 until 2014 [13]. 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 a 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

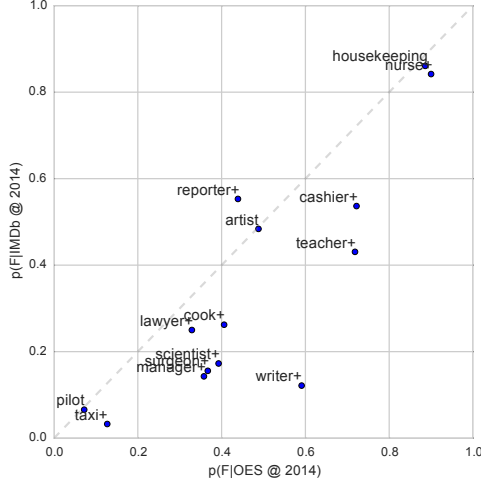


Figure 6: Proportion of female in IMDb and OES. + indicates significant at $p < 0.05$ in a two-tailed, two-proportion Z-test.

under-representation onscreen by female performers. For example, **scientists**, **cashiers**, **nurses** and **managers** 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 reversal in representation for others where female **reporters** are initially under-represented onscreen, then over-represented, and vice-versa for **nurses**. In some cases, female under-representation is becoming exacerbated, as is the case for **surgeons**, who are increasingly less likely to be played by a female performer, despite female surgeons becoming more common in real life.

There are several limitations of this analysis that should be taken into account before drawing strong conclusions. Firstly, 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. Note also that we were not able to retrieve OES data for 2001 and 2002, and

that the ontology used changed significantly after 2003, which may explain some of the curves. Secondly, we do not distinguish between US productions and those from other countries, so comparing with the OES may introduce some noise. Overall, this analysis lets us draw an interesting exploratory counterpoint between onscreen gender representation and real-world figures.

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 distance¹¹ [3], 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 distance between role distributions 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 pro-

¹¹Or Hellinger distance.

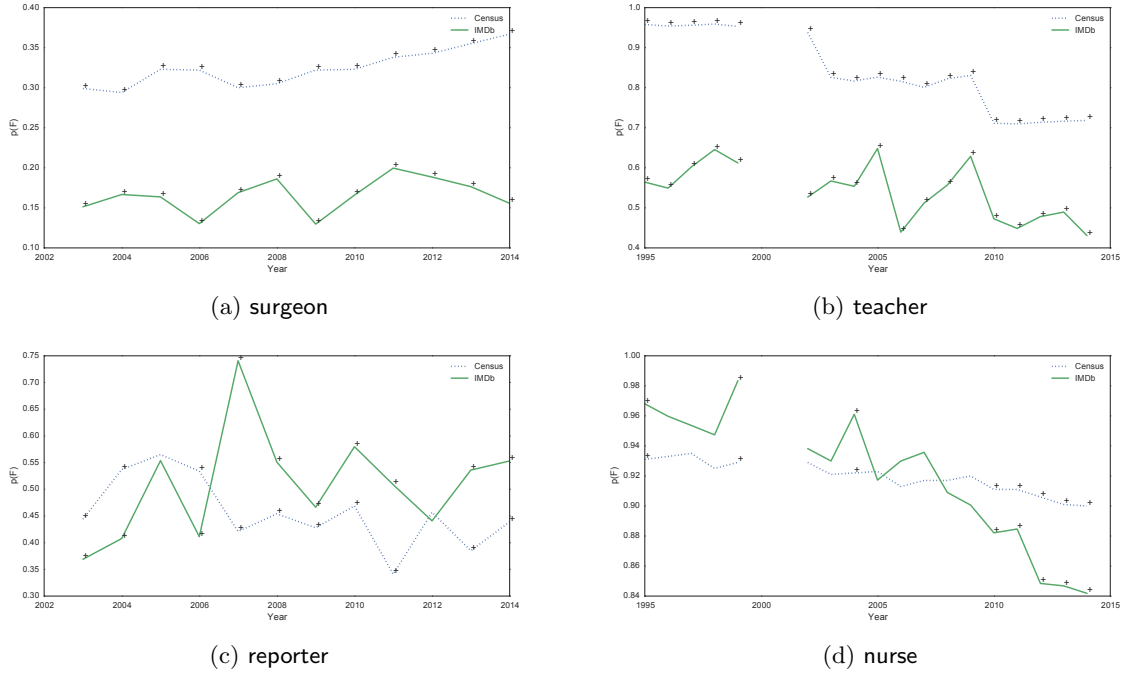


Figure 7: Gender distribution in IMDb and OES over time. + indicates significant at $p < 0.05$ using the two-tailed, two-proportion Z-test. Note that these do not use a rolling mean.

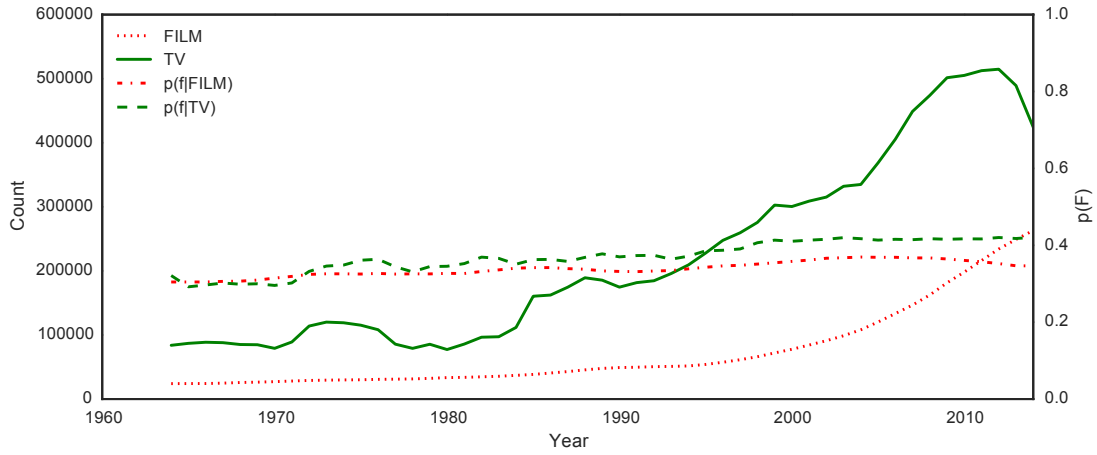


Figure 8: Count and proportion of female roles in film and television.

duction 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 7 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 pronounced 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)

Role	Film		M	Role	Television		M
	F	Role			F	Role	
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 7: The 20 most frequent female and male roles across film and television.

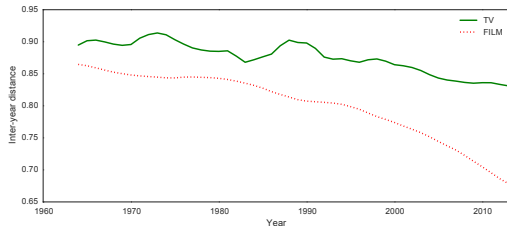


Figure 9: Year-on-year difference between role distributions for film and television.

or explicitly gendered roles (e.g., **policeman**, **waitress**), while the latter is more balanced with both males and females in common television roles.

8 Conclusion

Future work would concentrate on refining the data processing and adding useful structure for more rigorous statistical analysis. This includes linguistic analysis to aggregate role synonyms, many of which are multi-word expressions. Discriminating between genres may reveal interesting disparities on the gender proportion in them. Identifying a production country would also be useful for analysis and language identification. The IMDb data release does not report this information directly and it would have to be inferred. Our current model emphasises 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. Contrasting on-screen gender representation with real-life data has the greatest potential 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 paper presents methodologies for mining information about onscreen media gender from cast lists. 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 the 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

- [1] S. Bergsma. 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)*, pages 342–353, 2005.
- [2] S. Bergsma and D. Lin. 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*, pages 33–40, Sydney, Australia, July

2006. Association for Computational Linguistics.
- [3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.
 - [4] K. Boyle. Gender, comedy and reviewing culture on the internet movie database. *Participations: Journal of Audience & Reception Studies*, 11:31–49, May 2014.
 - [5] R. L. Collins. Content analysis of gender roles in media: Where are we now and where should we go? *Sex Roles*, 64:290–298, 2011.
 - [6] W. Duan, B. Gu, and A. B. Whinston. Do online reviews matter? - an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, 2008.
 - [7] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’02, pages 91–101, New York, NY, USA, 2002. ACM.
 - [8] J.-B. Michel, 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. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.
 - [9] G. A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.
 - [10] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics, July 2002.
 - [11] S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
 - [12] S. L. Smith, M. Choueiti, and K. Pieper. 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, 2014. Accessed: 22/1/15.
 - [13] United States Department of Labor. Occupational employment statistics, 2015. accessed July 2015, www.bls.gov/oes/.
 - [14] J. T. Wood. Gendered media: The influence of media on views of gender. In *Gendered Lives: Communication, Gender and Culture*, chapter 9, pages 231–244. Cengage Learning, 1994.