

# “Roles for the Boys?” Mining Cast Lists for Gender and Role Distributions 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. Finally, we compare real-life census gender distributions to our web-mediated onscreen gender data. We propose these methodologies are a useful adjunct to traditional analysis that allow researchers to explore the relationship between online and onscreen gender depictions.

## Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing

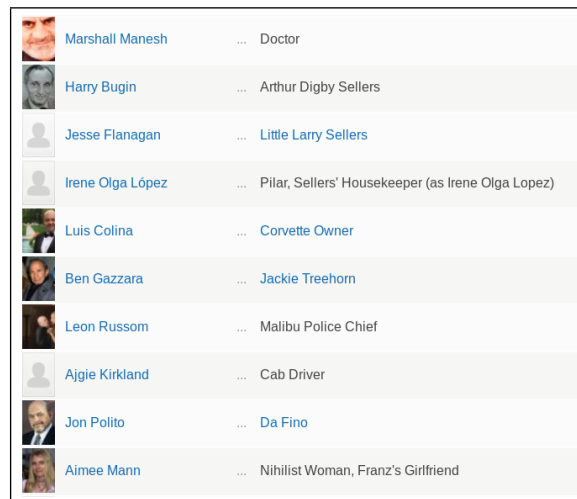
## Keywords

Gender; web science; social science; IMDb; screen media

## 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.

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













|                                                                                     |                  |                                                       |
|-------------------------------------------------------------------------------------|------------------|-------------------------------------------------------|
|    | 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                               |
|    | Ajgie 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”.

popular media website and examine *cast lists*. Figure 1 is a section of the Internet Movie Database (IMDb)<sup>1</sup> cast list from “The Big Lebowski”<sup>2</sup>, 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, 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 data<sup>3</sup>. 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. We believe that web science methodologies can augment traditional manual analysis to enable comparison of online and onscreen gender depictions.

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

<sup>2</sup>[www.imdb.com/title/tt0118715](http://www.imdb.com/title/tt0118715)

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

## 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 [11]. 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 [4], 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 characters<sup>4</sup> from the 600 top-grossing films from 2007–2013 [10]. 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 [3]. 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 [8]. 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 [9], 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 [6], 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-

<sup>4</sup>4 506 of these were speaking roles.

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 [5]. 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<sup>5</sup> 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`<sup>6</sup> and applied several cleaning phases. The files list the performer name and the titles and dates of productions they appear in. Unfortunately, these lists do not distinguish between films, television, so it is difficult to distinguish between media – clearly an important methodological question. We exclude records where the performer is listed using an alternative name (`as ...`), and generate one record per appearance in a film or television episode. We further process records based on the role, filtering roles marked `n/a`, or those that reference selves (e.g. `himself`, `herself` or `themselves`). 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)`). Finally, we remove any text in parentheses and split multi-role characters (e.g. `model/actress`), generating one count for each lower-cased role. 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 actresses list, normalised by the count of all  $r$  and  $y$  records.<sup>7</sup>

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 films and television, and our processing considers a television episode equal to a film.

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

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

<sup>7</sup> $p(M|r, y) = 1 - p(F|r, y)$ .

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

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          |
| mary              | reporter  | host               | various                    | anchor            | housemate       |
| jack              | dancer    | panelist           | various characters         | contestant        | police officer  |
| the girl          | townsman  | announcer          | member of the short circus | musical director  | alex            |
| the wife          | waiter    | sports newsreader  | paul williams              | lexicographer     | laura           |
| the sheriff       | narrator  | weather forecaster | victor newman              | interviewer       | audience member |
| minor role        | barfly    | correspondent      | brady black                | ridge forrester   | david           |
| the husband       | doctor    | correspondent      | jack abbott                | phil              | bar patron      |
| policeman         | bit role  | presenter          | george                     | emcee             | sam             |
| daughter          | bartender | sports reporter    | roman brady                | co-hostess        | sarah           |

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

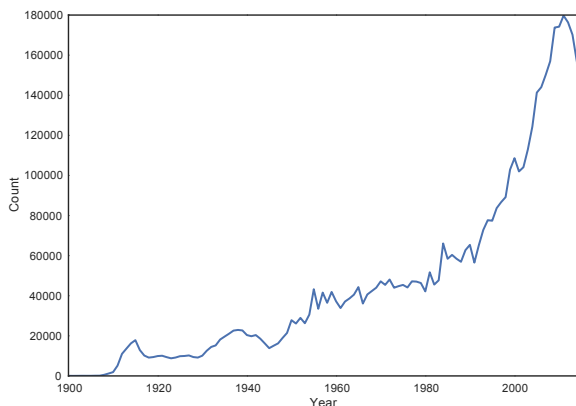


Figure 2: Count of roles over time.

This skews the data in favour of television and future work may be able to map to other resources to tease them apart. Likewise, 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 [7].

After preprocessing, we retain 15 468 002 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.<sup>8</sup>

## 4. ROLES

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*,

<sup>8</sup>We consider all data for counts, but graphs do not show data after 2014.

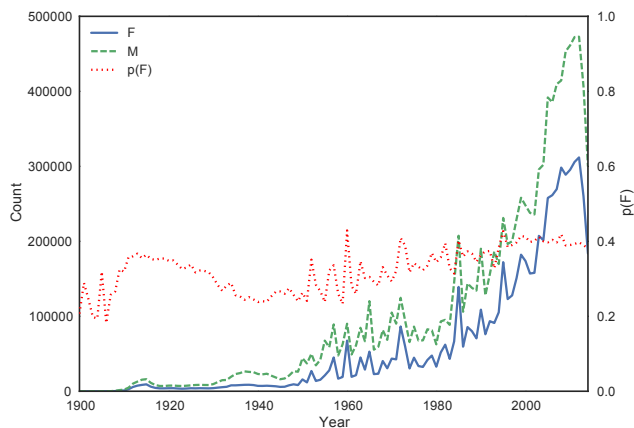


Figure 3: Count of roles from each gender over time, as well as the gender distribution  $p(F)$ .

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 (corresponal 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<sup>9</sup> 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 thus far from the two decades from 2000 reflects the recent trend for zombie characters in television, driven in part by the success of productions such as “The Walking Dead”, which typically feature many unnamed 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), and roles that reflect current naming conventions (police officer rather than policeman, the generic role mother and bar patron rather than the earlier bar fly). One concern with this method is that by only considering roles that have not been seen in a previous top 50, then we may find that the listed roles are low rank or count with respect to the overall

<sup>9</sup>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](http://en.wikipedia.org/wiki/Victor_Newman)

| Role               | F       | Role               | M       |
|--------------------|---------|--------------------|---------|
| host               | 123 775 | host               | 370 187 |
| hostess            | 74 856  | narrator           | 75 736  |
| presenter          | 39 551  | announcer          | 58 356  |
| newsreader         | 34 145  | presenter          | 51 762  |
| model              | 30 289  | guest              | 46 107  |
| guest              | 29 296  | various            | 33 917  |
| contestant         | 28 651  | newsreader         | 32 289  |
| reporter           | 25 911  | various characters | 31 785  |
| nurse              | 20 852  | contestant         | 31 739  |
| dancer             | 19 039  | reporter           | 31 190  |
| panelist           | 17 820  | panelist           | 25 999  |
| various            | 14 541  | judge              | 25 036  |
| judge              | 14 123  | additional voices  | 22 906  |
| narrator           | 13 714  | co-host            | 22 177  |
| co-host            | 12 314  | doctor             | 18 299  |
| various characters | 12 047  | policeman          | 16 590  |
| girl               | 11 595  | performer          | 15 964  |
| singer             | 11 509  | man                | 13 680  |
| woman              | 11 197  | bartender          | 13 327  |
| waitress           | 11 147  | various roles      | 12 522  |
| correspondent      | 10 691  | singer             | 12 463  |
| mother             | 10 009  | correspondent      | 12 356  |
| laura              | 9 930   | dancer             | 12 173  |
| maria              | 9 860   | musical guest      | 11 937  |
| additional         | 9 652   | waiter             | 11 876  |
| performer          | 8 582   | police officer     | 11 206  |
| sarah              | 8 235   | cop                | 10 812  |
| lisa               | 8 122   | soldier            | 10 185  |
| anna               | 8 002   | david              | 10 087  |
| co-hostess         | 7 847   | student            | 10 070  |
| student            | 7 624   | guard              | 9 906   |
| mary               | 6 960   | detective          | 9 720   |
| rita               | 6 908   | paul               | 9 315   |
| alice              | 6 744   | tom                | 9 124   |
| rosa               | 6 730   | sports newsreader  | 9 078   |
| jane               | 6 022   | john               | 9 068   |
| various roles      | 5 922   | jack               | 8 978   |
| julie              | 5 790   | commentator        | 8 864   |
| secretary          | 5 692   | mike               | 8 536   |
| sara               | 5 546   | townsman           | 8 522   |
| linda              | 5 427   | max                | 8 508   |
| receptionist       | 5 419   | extra              | 8 363   |
| extra              | 5 221   | frank              | 8 281   |
| eva                | 5 135   | boy                | 8 271   |
| marta              | 5 013   | mark               | 7 999   |
| jenny              | 5 002   | tony               | 7 936   |
| sandra             | 4 930   | george             | 7 895   |
| ana                | 4 860   | musician           | 7 840   |
| teresa             | 4 800   | interviewee        | 7 822   |
| clara              | 4 775   | joe                | 7 803   |

Table 3: The 50 most frequent female and male roles.

roles (i.e. as per Table 1). The lowest rank was 40 (sarah in 2000–2020) and the lowest count was 614 (bartender in 1920–1940).

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.

## 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 3 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. The higher female proportion around 1920 may reflect the fact that records correspond to film, not television, but this is difficult to establish without taking extra metadata into account.

Table 3 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.

| Strongly male     |        | Moderately male   |        | Gender neutral   |        | Moderately female |        | Strongly female            |        |
|-------------------|--------|-------------------|--------|------------------|--------|-------------------|--------|----------------------------|--------|
| Role              | $p(F)$ | Role              | $p(F)$ | Role             | $p(F)$ | Role              | $p(F)$ | Role                       | $p(F)$ |
| delivery man      | 0.00   | band              | 0.05   | emt              | 0.17   | corresponsal      | 0.35   | member of the short circus | 0.60   |
| color commentator | 0.00   | little boy        | 0.05   | player           | 0.18   | center square     | 0.35   | secretary                  | 0.88   |
| father            | 0.00   | basketball player | 0.07   | additional voice | 0.20   | patient           | 0.35   | mother                     | 0.93   |
| boyfriend         | 0.00   | biker             | 0.07   | trainer          | 0.22   | co-host           | 0.36   | nurse                      | 0.94   |
| policeman         | 0.00   | moderator         | 0.09   | host             | 0.25   | hotel guest       | 0.36   | old woman                  | 0.96   |
| musical director  | 0.00   | coroner           | 0.10   | mentor           | 0.26   | office worker     | 0.40   | model                      | 0.97   |
| truck driver      | 0.01   | fbi agent         | 0.10   | guest co-host    | 0.27   | news anchor       | 0.42   | actress                    | 0.98   |
| inspector         | 0.02   | bailliff          | 0.11   | inmate           | 0.28   | android           | 0.43   | maid                       | 0.98   |
| monk              | 0.02   | bartender         | 0.13   | passerby         | 0.29   | candidate         | 0.44   | stewardess                 | 0.99   |
| soldier           | 0.02   | staff humorist    | 0.14   | journalist       | 0.31   | participant       | 0.47   | secretaria                 | 1.00   |

Table 4: Examples of common roles with different gender distributions.

| Profession  | Keywords                                             | $p(F)$ |
|-------------|------------------------------------------------------|--------|
| IT          | software, computer, hacker                           | 0.51   |
| Doctor      | medical, dr, dr., doctor<br>md, physician            | 0.23   |
| Corporate   | corporate, ceo, coo                                  | 0.18   |
| Law         | prosecutor, lawyer                                   | 0.15   |
| Politics    | minister, dictator, parliament<br>senator, president | 0.09   |
| Science     | science, professor<br>priest, priestess, reverend    | 0.09   |
| Religion    | pastor, prior, allamah<br>imam, rabbi, guru, lama    | 0.08   |
| Engineering | bishop, ayatollah, swami<br>engineer                 | 0.05   |

Table 5: Gender distribution grouped by profession.

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 *policemen*, *men*, *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 1000 that did not belong to a list of common names from the US Census. To try and characterise the space of roles, we ordered them by  $p(F)$  and partitioned them into five equal bins and randomly sampled 10 entries from each. Table 4 shows the results: on both extremes there are again gendered roles (*boyfriend*, *actress*), while more towards the middle section some more interesting biases can be observed (*biker* and *basketball player* as male and *secretary* as female). Note that due to the overall higher count of male occurrences, the midpoint of gender distribution is between the “moderately” and “strongly” female classes.

In [10], 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 5, 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 [10]

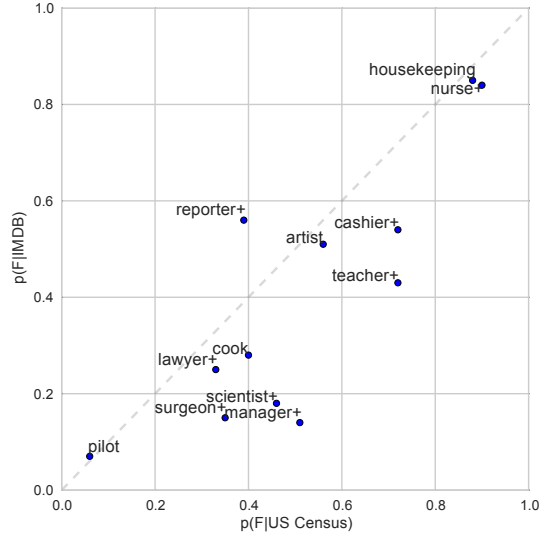


Figure 5: Proportion of female in movie dataset vs. US Census. + indicates significant at  $p < 0.05$  in a two-tailed, two-proportion Z-test.

for Law but not for corporate professions, while the medical domain (doctors) had a female probability of 0.23. In contrast to the results in [10], Religion does not score at the bottom with regards to female presentation (although very low with 0.08). From the professions we selected, Engineering was the lowest (0.05). The highest scoring profession was IT (0.52), which is partly due to the fact that many computer voices were female (*computer* had 460 female occurrences, versus 247 male ones; 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 4 shows the distribution of two roles, 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. Conversely, the initial low proportion of onscreen female reporters has risen and the proportion is now relatively even.

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. Figure 5 shows how onscreen gender distributions

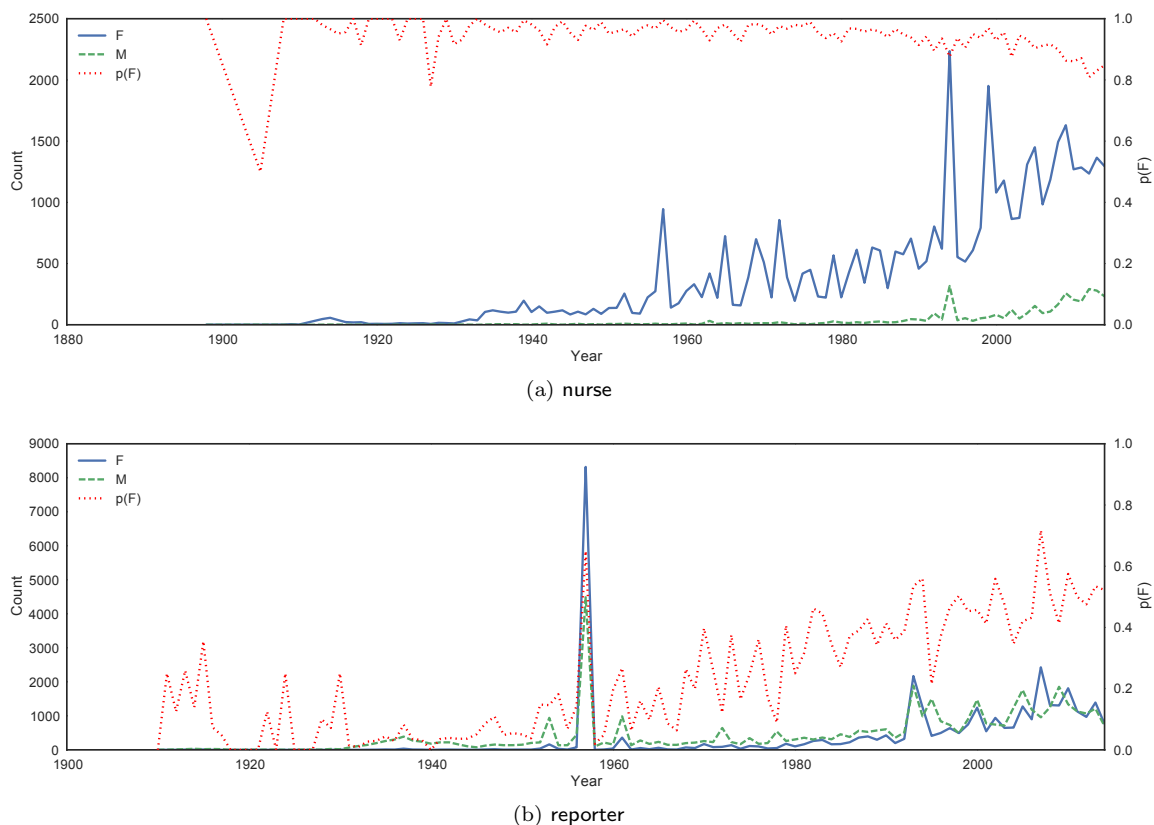


Figure 4: Gender counts and proportions over time for various roles.

map to those listed in the US Census<sup>10</sup>. In both cases, the data was restricted to 2014. Intuitively, points on the diagonal line have a portrayal consistent with the census 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, *scientists*, *cashiers*, *nurses* and *managers* are played more frequently by male performers than their census counterparts. There are several limitations of this analysis that should be taken into account before drawing strong conclusions. Firstly, comparing user-generated roles with strict census 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. Secondly, we do not distinguish between US productions and those from other countries, so comparing with the US Census may introduce some noise. Overall, this analysis lets us draw an interesting exploratory counterpoint between onscreen gender representation and real-world figures.

## 6. 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.

<sup>10</sup><http://www.bls.gov/cps/cpsaat11.pdf>

Discriminating between media types (film, television) and 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 5, but further analysis would require matching the informal IMDb and formal census 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.

## 7. ACKNOWLEDGEMENTS

The authors wish to thank XRCE colleagues and Kellie Webster for thoughtful early feedback.



## 8. 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] K. Boyle. Gender, comedy and reviewing culture on the internet movie database. *Participations: Journal of Audience & Reception Studies*, 11:31–49, May 2014.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] G. A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.
- [8] 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.
- [9] 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.
- [10] 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://annenbergl.usc.edu/pages/~media/MDSCI/Gender\\_Inequality\\_in\\_500\\_Popular\\_Films\\_-\\_Smith\\_2013.ashx](http://annenbergl.usc.edu/pages/~media/MDSCI/Gender_Inequality_in_500_Popular_Films_-_Smith_2013.ashx), 2014. Accessed: 22/1/15.
- [11] 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.