Running Title: Is she married and how many children does that CEO have? A look at gender bias in Wikipedia Infoboxes

Abstract

Introduction

Gender bias in Wikipedia, especially with respect to representation, coverage and article length has been previously attributed to both the prevalence of male contributors to Wikipedia, as well as inherited bias from source material such as traditional encyclopedias (Reagle & Rhue, 2011). The lack of female authors is thought to bias both the representation (in both the number of Wikipedia entries as well as the length of each individual article) and the characterization of Wikipedia articles on women (Graells-Garrido, Lalmas, & Menczer, 2015). While Wikipedia articles have increasingly become the crowdsourced online encyclopedia it set out to be, less attention has been paid to solely the information found in the Wikipedia Infobox.

This study examines the type and amount of data found in Wikipedia's Infobox in an attempt to provide a more nuanced examination of the gender gap. It examines the degree to which infobox data is gendered. For example, how often women's infoboxes contain information on spouses, children and other family members as well as what percentage of a women's infobox information is solely dedicated to such information.

Method

Wikipedia does not list gender within either its infobox or its categorical data, thus, to determine gender identity, we counted pronouns. There are obvious limitations to this -- we did not filter for nonbinary classifications (e.g., Nicky Case, whose Wikipedia entry details that they prefer gender-neutral terminology) or people who have switched classifications (e.g., Caitlyn Jenner, who used to be classified as he/him/his but is now known through she/her/hers pronouns).

Reagle, J., & Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communications 5, 1138-1158*

I. Wikipedia Infobox

A. 9.25.2020 First cut data, from 2017-08-20 snapshot of all [[Category:Living people]]

	Articles	Infobox Spouse	Infobox Children	Infobox Height
Actor	34443	14062 (40.8%)	7513 (21.8%)	3296 (9.6%)
Actress	29350	12919 (44.0%)	6942 (23.7%)	3530 (12.0%)

1. Exclude redirects (only actual articles)
2. Condition on [[Category:Living people]]
3. Rows correspond to "actor" and "actress" appearing as a category for the article, as a case-insensitive substring. Note that some articles have both (actor is *sometimes* gender neutral)
4. spouse/children/height also detected as arbitrary substring from infobox element.

B. Career Highlights vs. Family and Children: Men are “career” focused. Women are “family” focused (spouse/partner/children)

1. A male person of interest is far more likely to have an information box that is focused on career statistics, while a female is far more likely to have words and links to their “spouse” or “children”. **<u>What percentage of the Info Box is “career”-oriented vs. “family”-oriented?</u>**
    - **a) <u>Infobox characters devoted to spouse</u>**
    - **b) <u>Infobox characters devoted to children</u>**
    - **c) <u>Infobox characters devoted to spouse + children</u>**
    - **d) <u>(vs.) Total word count</u>**

2. a few examples:

    a) Billy Joel (singer) and Christie Brinkley (model)
    - (1) They’re both celebrities with multiple marriages
    - (2) They’re both entertainer/business people with net worth of 50 million+
    - (3) Both have been married four times.
    - (4) Christie’s info page lists four marriages, links to one of the children she shares with Billy Joel; Billy’s only mentions the marriages when you scroll down and read about his personal life, does not mention marriage or children within his information.

    b) Halle Berry/Eric Benet or Halle Berry/David Justice

    c) Giselle Bundchen/Tom Brady

    d) Even if we don’t count just links, a woman’s “**partner”** is often considered important enough of a detail to be added (see: Tyra Banks or Demi Lovato-- and with Demi, another example of she links to her former fiancee, who is far less successful/well known)

C. For Women: Emphasis on Physical Characteristics (more?)

1. Number of times “height” “hair color” “eye color” or “eye colour” mentioned
    - a) Infobox characters devoted to “height” as percentage of total infobox characters
    - b) Infobox characters devoted to each of the above as percentage of total infobox characters
2. “Height” Gong Li is another example where it links to one of her spouses (but not vice versa) and also, it lists Gong Li’s **height** as well within the infobox.

3. Working on whether this is biased or not. Difficult to check male vs. female models here (Tyson Beckford has height listed and even additional details, but personal information is below, so even though his physical characteristics matter, they matter less?)

4. **<u>How many pictures or pictures/100 words? (Clark Gable 21, Spencer Tracy 17, Audrey Hepburn 23, Katherine Herburn 22 -- difficult to track)</u>**

D. Is this Independent of Success/Popularity? **(Seems like it?)**

1. Kim Kardashian vs. Kris Humphries (all three of her marriages are listed, she has a far higher net worth ~900 million, ongoing success/fame, for Kris, his career highlight (within the infobox) comes before he was recruited, and yet his infobox emphasizes his career statistics and accomplishments, no mention of personal life)

2. Methodology for Popularity:
    - **a) Total word count? Percentage of total word count that is listed under “personal life”**
    - b) Is it linking to Google or a Wikipedia page? (not gettable -- even if we query it directly afterwards?)
    - c) Google Search Hits?
    - d) Instagram/Twitter Followers
    - e) Box office data (might not be easy to scrape and noisy)

E. Education:

1. Would like to look into how often “education” or “alma mater” comes up for men vs. women.

2. Anecdotally: Again, a larger percentage of Kylie Jenner’s box is her relationship with Travis Scott and her children. Travis Scott also has career/education, etc.

F. How do gay men and lesbian women get classified/impacted?

1. Like Billy Porter: Gay icon, education is listed, partner is not. Caitlyn Jenner (now listed as a “she” spouses, etc. emphasized)

II. Wikipedia Information

A. Who’s Contributing (we don’t know if it’s really possible to track the gender of the person adding the contribution. Other studies though would seem to suggest that it wouldn’t matter, women would perhaps also focus more on the female-oriented information as well)

1. Who is entering the data (gender breakdown) -- lots of articles on this, not sure how to get at the data since it’s based on usernames.

2. Number of contributors? Is there an impact?

III. Miscellaneous

A. How do we remove people who are single?

B. I believe that it's a bias that can get fixed in time. Anecdotally, the older Hollywood stars tend to be a little more even in terms of listing of spouses, though the men still have larger infoboxes where the emphasis is more accomplishment driven. Not sure if this is worth looking into.

C. Unsure if there's a way to track this but there are small things where I don't think we can farm for but are true: Frédéric Joliot-Curie won the Nobel jointly with his wife Irène Joliot-Curie (it's listed in her infobox as "jointly" but in his it's just listed as an award, though again, if you read it, it will say jointly). *I feel like husbands/wives are great case studies here, but likely too much for this project for now.*

D. **Too hard to track for now:** women who are never married vs. men who were never married and the percentage of text devoted to that (and again, sexuality complications: James Buchanan has a whole section, as the only bachelor president, has a section on just that, Rosalind Franklin has a section explaining controversies, affairs, and never married)

E. **Very separately:** I'm VERY interested in what PERCENTAGE of a non-white person's wikipedia information is devoted to talk of race. Often it's classified under "personal information" but I believe there is almost an expectation that a black actor is going to be talking about race and a white actor is not. Often quotes, etc. linked to various controversies. (Like the world Black will come up 13 times on Morgan Freeman's page, some of that is the links below, wouldn't know how to clean that, and once it's someone's last name… so there's suddenly a subclassification of all of his achievements because these were black plays.)

F. Actors/Actresses easier first grab, not sure how to grab others: Sandra Day O'Connor, Sonia Sotomayor, RBG all have spouses listed. Elena Kagan does not (but has never married). Small percentage of men have it listed (though again this could be a time thing, as in perhaps we're moving towards having them all). With politicians I'm fairly sure it's skewed, not worth grabbing now.

Technical Notes:

- Anything visible as structured data when looking at the wikitext (go to article, click "Edit") is potentially easy to do
- *Categorizing* general terms is hard, but using categorizations that are already there is easy
  - Things like: [[Category:Useless Trivia]]
- There's an official-ish sql interface to Wikipedia that may be helpful:
  - https://quarry.wmflabs.org/
  - Requires a wikipedia login, I think that's a minor issue
- Plan to create a SQLite3 database
  - Make it easy to answer variations on the above questions as they evolve
  - Kids contribute feature extraction code

  - Primary table has one column per “feature”:
    - Gender (possibly multiple, if we have more than one method)
    - Categories
    - Infobox (extract -> JSON? Something?)
    - Order of personal/professional info
    - Word count
    - Picture count
    - Superficial content
      - Height/weight
    - Possibly sentiment analysis
      - Types of words used: “controversy”, “sex”
    - Spouse
      - Spouse article id useful when present
      - Should also support lookup by name (sometimes unlinked)
  - Produce both a “full” table and a “sampled” table
    - Smaller table for iterating / spot checking
    - If we’re truly being good, we should do our exploration mostly on the sampled data!
- Have a compute node that should be able to host full uncompressed wikipedia + some derived dbs in memory
  - Single snapshot only
- Have the following snapshot dates:
  - 20170820
  - 20181020
  - 20190101
  - 20200401
  - 20200901
  - May be able to find more if we search (like late 2019)
  - New ones come out ~monthly

1. Linguistic bias (don’t like the abstract vs. concrete way, but could do)
   a. Socialite
   b. Urban (thanks Pinker)

2. Cultural dimensions though Hofstede?

3. Academia:
   a. Citations on Google Scholar (control within field, omg fields so different) vs. Wikipedia
   b. H-index
   c. “Influenced” within wikipedia infobox
   d. How to control for social media presence (Lior vs. Barbara Wold)
   e. https://www.theatlantic.com/ideas/archive/2020/08/women-scientists-have-evidence-about-sexism-science/615823/
   f. Pair by age? (rough proxy of likelihood of social media presence)

Session below from my summary db, 2017 data. Some fields for reference:

- Sampled all “Living people” with “actor” or “actress” appearing within their categories

- gender : ‘M’ male, ‘F’ female, ‘D’ disagreement, ‘?’ no signal
  - Use actor/actress category, first pronoun, and pronoun counts as signals
  - Settle on ‘D’ if there are any conflicting signals
  - If any signals come up empty, they stay out of voting
- personal_first, career_first - Section with ‘personal’ or ‘career’ in name appear before the other
  - 0 or 1 (both may be zero if neither section exists)
- personal, career - Number of sections with ‘personal’ or ‘career’, resp
- info_*_chars - Number of characters in infobox section of this name

---

```
sqlite> select avg(info_spouse_chars * 1.0 / info_value_chars) as pct_spouse,
gender, count(*) from summary group by gender;
pct_spouse          gender      count(*)
------------------  ----------  ----------
0.0193191522861121  ?           47
0.0537099947147733  D           714
0.0527697909240428  F           29112
0.0390361677940769  M           31816
sqlite> select gender, count(*), avg(personal_first), avg(career_first) from
summary group by gender;
gender      count(*)    avg(personal_first)  avg(career_first)
----------  ----------  -------------------  ------------------
?           47          0.0                  0.0212765957446809
D           714         0.0896358543417367   0.488795518207283
F           29112       0.0867683429513603   0.463623248145095
M           31816       0.0769738496354036   0.447730701533819
sqlite> select gender, count(*), avg(personal_first), avg(career_first) from
summary where personal>0 and career>0 group by gender;
gender      count(*)    avg(personal_first)  avg(career_first)
----------  ----------  -------------------  -----------------
D           168         0.136904761904762    0.863095238095238
F           6797        0.151243195527439    0.848756804472561
M           6496        0.1564039408867      0.8435960591133
sqlite> select gender, count(*), avg(personal>0), avg(career>0) from summary
group by gender;
gender      count(*)    avg(personal>0)  avg(career>0)
----------  ----------  ---------------  ------------------
?           47          0.0              0.0212765957446809
D           714         0.2927170868347  0.521008403361345
F           29112       0.2849340478153  0.498935147018412
M           31816       0.2492142318330  0.479664319839075
sqlite> select gender, count(*), avg(height), avg(weight) from summary group by
gender;
Error: no such column: height
sqlite> select gender, count(*), avg(info_height_chars), avg(info_weight_chars)
from summary group by gender;
gender      count(*)    avg(info_height_chars)  avg(info_weight_chars)
```

```
----------  ----------  ----------------------  ----------------------
?           47          6.93617021276596        2.61702127659574
D           714         7.02380952380952        2.7843137254902
F           29112       4.41285380599066        1.11287441604837
M           31816       2.09853532813679        0.727149861704803
sqlite> select gender, count(*), avg(info_height_chars>0),
avg(info_weight_chars>0) from summary group by gender;
gender      count(*)    avg(info_height_chars>0)  avg(info_weight_chars>0)
----------  ----------  ------------------------  ------------------------
?           47          0.127659574468085         0.0638297872340425
D           714         0.106442577030812         0.0518207282913165
F           29112       0.0892758999725199        0.0251099203077769
M           31816       0.0558838320341966        0.0157153633392004
```