Gender and Academic Publications

Visualize a gender gap in business & economics

Malik Stromberg and Franziska Müller

19/03/2021

University of Tübingen

Outline

1. Project Idea
2. Data Acquisition
3. Classification
4. App

Project Idea

Project Idea

Gender Ratio and Academic Publications

1. Is there a gender gap in academic publications?
2. How has the gender ratio changed over time?
3. Are there differences between different areas?
   (within social sciences)

→ Algorithm for gender classification based on first names

Limitations - conceptually

- Focus on the area Business and Economics
- Binary gender classification
- Data: choose to take data from 1960 till 2020

Data Acquisition

Publication Data

Publication Data Set

- Infos regarding publications: author name, year, journal, ...
- Note: Focus on the area Economics and Business
  our program would also work for other disciplines
- Representative data set of as many researchers as possible
- Final data set structure: One observation per author and distinct publication
- *Challenges:* Full names of authors, completeness of data set

Publication Data

Characteristics of Data Sets	*google scholar* third party supplier e.g. Serp API [9]	*Web of Science* [1]	Crossref *[3]*
Open data	Paywall	Yes	Yes
Completeness	Yes	Deficient [5]	Yes
Full surnames	No	No	Yes
Main disadvantage	Limited access	No complete surnames	Unstable Server
Additional Info		R Client	R Client

Table 1: Comparison of different APIs for publication data.

Crossref

General Info

- Non-profit platform to improve scholarly communication as well as find, cite, link, assess and reuse research outputs [2]
- 2,192 Members in 2018: Organizations and Publishers [2]

Challenges and solutions

- Unstable server
  $\rightarrow$ Using loop for each field within an area, adjusting pause
- Completeness publications? $\rightarrow$ Yes [5]
- Noticeable less publications in 2012 and 2015
  $\rightarrow$ Contacted Crossref: in some cases category information has only been updated sporadically
- Very few data points in the early years (starting in 1852)
  $\rightarrow$ Split the data set (1960-2020)

Publication Data

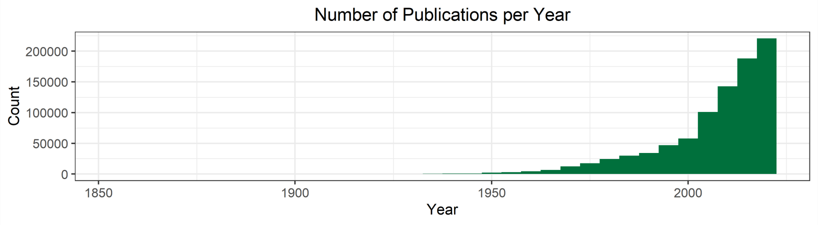

Figure 1: Absolute number of publications per year.

Available Data

- Publications from 1852 till 2022
- 2,222 different journals
- 1,490,879 different articles
- 2,987,756 data points

Final Data

- Publications from 1960 till 2020
- 2,220 different journals
- 1,421,933 different papers
- 2,876,192 data points

Name Data

Name Data Set

- Surnames and respective gender
- Preferred with capacity term
- *Challenges:* Gender neutral names (*Kim*), country-specific assignments (*Andrea*), e.g. Asian names

Name Data

Characteristics	Own Data Set	*Larivière et. al. 2013 [6]*	API *gender-api.com* [7]	R package *genderizeR* [15]
Main data source	US Census [8] 1880-2018	US Census, human coders	Census Data, public social profiles	genderize.io (publ. soc. profiles), human coders
Open data	Yes	No	Paywall	Yes
Reproducible	Yes	No	Yes	Yes
Updates	[No]	No	NA	No
Prob. prediction	Yes	Not entirely	Yes	Yes
Global Reach	Country-specific	Yes	Yes (engl.)	Country-specific
Main disadvantage	Selection bias, e.g. Asian names	Data not available	Limited access, Cooperation possible but with restrictions	Incorrect names

Table 2: Comparison of different data sets and potential sources (excerpt).

Details main data set

- circa 351.65 Mio names in total
- 98,400 unique names

Extending the data set - Human coders

Could human encoding help enriching the data set?

- *Theory:* classifying unknown names by human
- Human coders: Decide by own knowledge,
  enriched by web searches (e.g. Google images)
- *Practical Problems:*
  - Efficiency and resources
  - Capacity terms, country-specific
    $\rightarrow$ would need many coders

$\rightarrow$ Some rare names are not classifiable by human
E.g. Radivoj, Desalegn, Seok $\rightarrow$ Still helpful for intuition

Extending the data set - Further sources

If the name could not been classified, further sources are considered:

1. **Name endings**
   4,712 different endings - e.g. 'ert', 'ine'
2. **API** [7]
   Part of the unclassified names are queried in an API
   e.g. Names: 'Recep' (m), 'Seb' (m), 'Burcu'(f)

Impact data set

- Scientific Journal Rankings (SJR):
  regarding citations in the previous three years
- H Index
  regarding citations over the whole period
- available data from 1999-2019

Classification

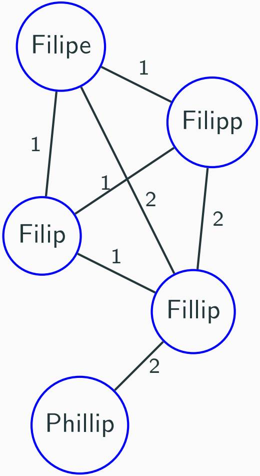

Figure 2: Example for appropriate case for kNN.

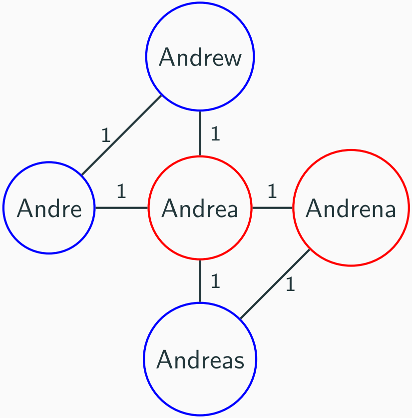

Figure 3: Example for inappropriate case for kNN.

Naive Bayes

$$P\,(gender \mid name) = \frac{P\,(name \mid gender)\,P\,(gender)}{P\,(name)} \tag{1}$$

$$f\,(\,P\,(male \mid name)\,) = \begin{cases} \mathrm{M} & \text{if } P\,(male \mid name) > s \\ \mathrm{F} & \text{if } P\,(male \mid name) < s \\ \mathrm{D} & \text{else} \end{cases} \tag{2}$$

s: decision threshold

Algorithm Overview

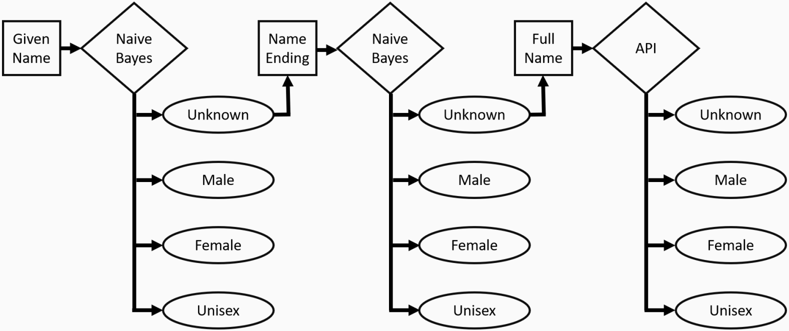

Figure 4: Algorithm overview.

Classification Results

Class	Full Names NB	Names Ending NB	NB	API	All
n	2,697,882	455,521	**2,697,882**	59,741	**2,662,558**
Male	59%	50.53%	**67.53%**	67%	**69.93%**
Female	23.1%	24.38%	**27.21%**	11.64%	**27.84%**
Unisex	1.02%	4.22%	**1.73%**	13.8%	**2.06%**
Unknown	16.88%	20.87%	**3.52%**	7.61%	**0.17%**

Table 3: Classification results with respect to publications and authors.

Sanity Check - Human Coder

In **75%** of the cases the human coder assigns the same label as the Naive Bayes algorithm.
In **94.7%** of the cases the human coder assigns the same label as the Naive Bayes approach when the name is neither unfamiliar to the algorithm nor to human coder.

Name	Algorithm	Human encoder	Suggested Reason
Jorg	Female	Male	Algorithm Error
Lei	Female	Male	Cultural Differences
Jess	Male	Female	Cultural Differences
Georgi	Female	Male	Cultural Differences

Table 4: Comparison of algorithm and human coding (excerpt).

Sanity Check - Data Sources

Classification of a random sample of 5,000 names by NB and API

NB \ API	Male	Female	Unisex	Sum
Male	1,834	292	68	2,194
Female	586	1,094	95	1,775
Unisex	79	41	5	125
Sum	2,499	1,427	168	4,094

Table 5: Confusion matrix for random sample classification.

Sanity Check - Data Sources

Name	NB	API	Other Sources
Dorien	Male (94%)	Female (92%)	Unisex [4] [10]
Awn	Female (100%)	Male (88%)	Female [11]
Gabriele	Female (72%)	Male (83%)	Unisex [12]
Jaka	Female (72%)	Male (95%)	Unisex [13]
Sany	Female (98%)	Unisex (M: 54%)	Male [14]

Table 6: Comparison of naive bayes and api classification (excerpt).

Decision Threshold

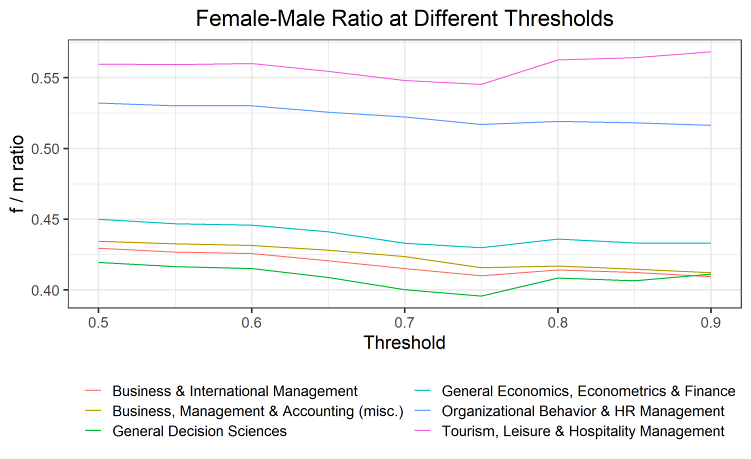

Figure 5: Female-male ratio at different decision thresholds.

Decision Threshold

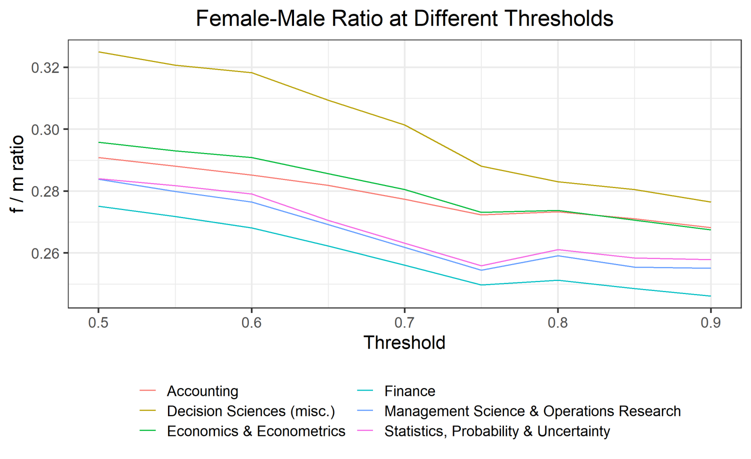

Figure 6: Female-male ratio at different decision thresholds.

Decision Threshold

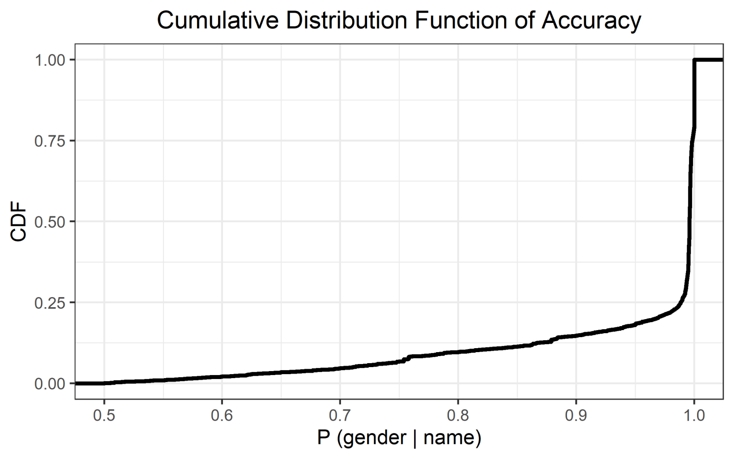

Figure 7: Cumulative distribution function of accuracy.

App

App Data Preparation

→ reduce size of the data set to improve user experience
→ keep loss of information small

Create Measures

- percentage share
- female to male ratio

Aggregation

- overall summary
- year-level
- research field-level
- journal-level

Data Splitting

- different information / different sections
- different global variables: decision threshold

Visualisation

Visualizations act as a campfire around which we gather to tell stories.
— Al Shalloway

Main considerations for our visualisation

- Represent all information in an aesthetic way
- Rather web page layout - Using PagePilling
- Overview general, afterwards more details
- Adaptive user interaction and increasing complexity
- Possible publication

App - Title

Figure 8: Section - Title.

App - Data Overview

Figure 9: Section - Data overview.

App - Gender Distribution Overview

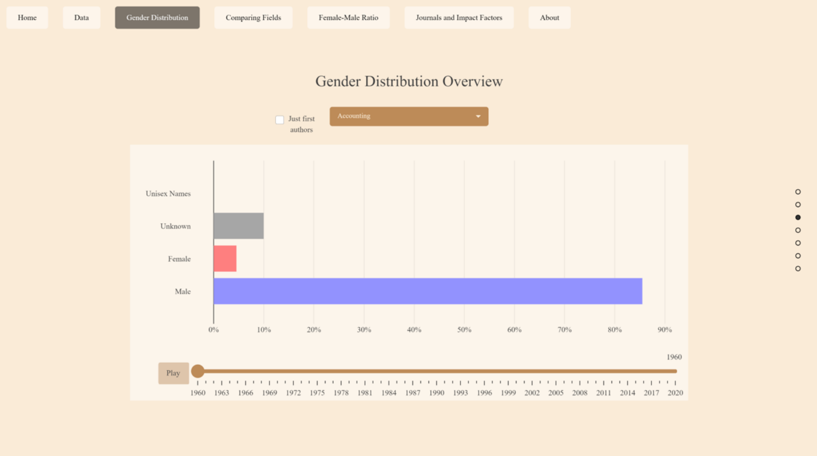

Figure 10: Section - Gender distribution overview.

App - Comparing Different Fields

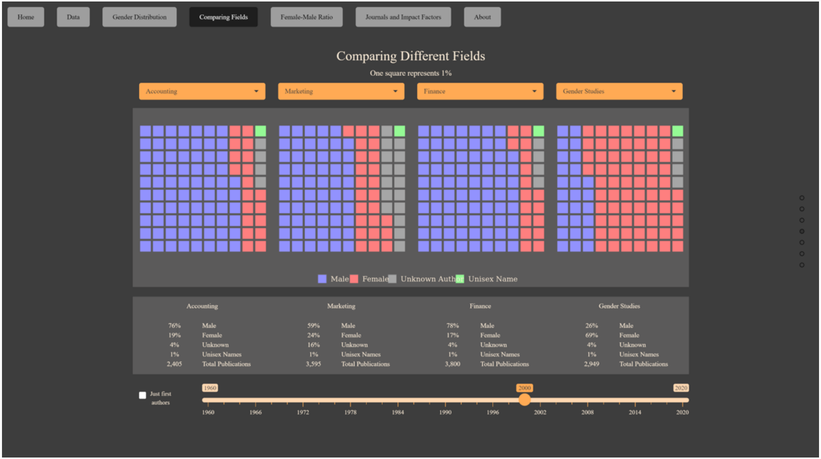

Figure 11: Section - Comparing different fields.

App - Comparison Between Research Fields Over Time

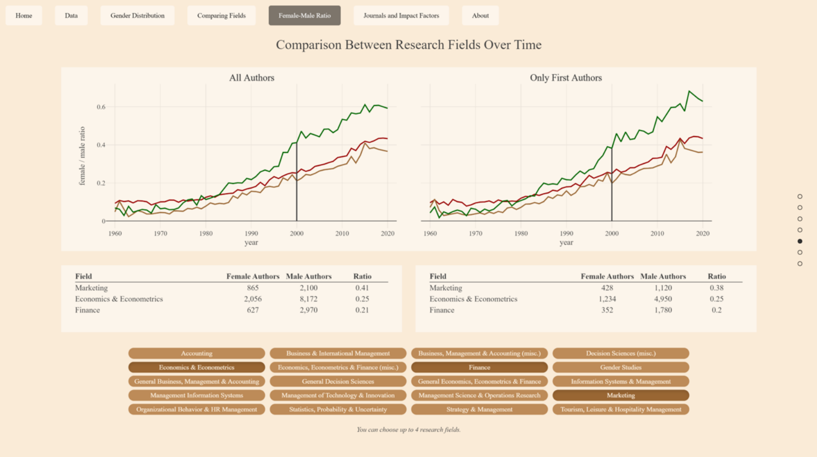

Figure 12: Section - Comparison between research fields over time.

Female to Male Ratio

$$ratio_{c,t} = \frac{\#female_{c,t}}{\#male_{c,t}} \tag{3}$$

$$weighted\ ratio_t = \frac{\sum_{k=1}^{K} \sum_{j=1}^{n_k} w_k \cdot (impact_{j,\,t}) \cdot \#female_{j,\,t}}{\sum_{k=1}^{K} \sum_{j=1}^{n_k} w_k \cdot (impact_{j,\,t}) \cdot \#male_{j,\,t}} \tag{4}$$

w_k:	weight for group k
$impact_{j,\,t}$:	impact of journal j in year t
$\#female_{j,\,t}$:	number of female authors of journal j in year t
$\#male_{j,\,t}$:	number of male authors of journal j in year t

App - Impact Factor and Gender Gap

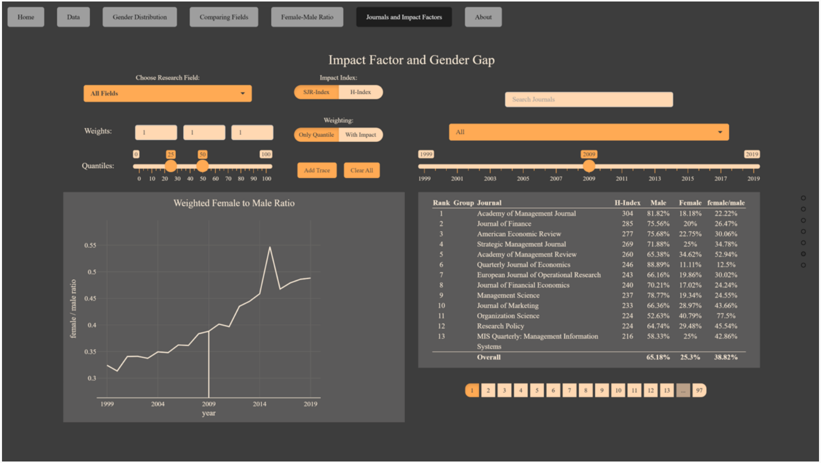

Figure 13: Section - Impact factor and gender gap.

Results and extensions

General Results

- Less women than men (except for gender studies)
- Decreasing gender gap over time till Covid-19
- Covid-19: Less publications by women
- Comparing the largest fields, in marketing the gender gap decreases fastest
- Gender gap tends to increase in impact factor

Possible extensions

- Other research fields
- Forecasting ratios

Thanks!

References

[1] Clarivate. *API sample data*. URL: https://clarivate.com/webofsciencegroup/solutions/api-sample-data/.

[2] Crossref. *Crossref Annual Report & Fact File 2018-19*. URL: https://doi.org/10.13003/y8ygwm5.

[3] Crossref. *General information about Crossref*. URL: https://www.crossref.org/.

[4] *firstname.de*. URL: https://www.firstname.de/Vorname/Dorien/.

[5] Anne-Wil Harzing. "Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?" In: *Scientometrics* 120.1 (2019), pp. 341–349. ISSN: 0138-9130. DOI: 10.1007/s11192-019-03114-y.

[6] Vincent Larivière et al. "Bibliometrics: global gender disparities in science". In: *Nature* 504.7479 (2013), pp. 211–213. DOI: 10.1038/504211a.

[7] Markus Perl. *API gender-api.com*. URL: https://gender-api.com/de/frequently-asked-questions.

[8] Andrew Piechocki. *Social Security Admistration Data 1880-2018*. URL: https://data.world/dpieski/namesgender/workspace/file?filename=names_gender.csv.

[9] Serp API. *Google scholar API*. URL: https://serpapi.com/google-scholar-api.

[10] *vornamen.blog*. URL: https://vornamen.blog/Dorien.

[11] *vornamen.blog*. URL: https://vornamen.blog/Awn.

[12] *vornamen.blog*. URL: https://vornamen.blog/Gabriele.

[13] *vornamen.blog*. URL: https://vornamen.blog/Jaka.

[14] *vornamen.blog*. URL: https://vornamen.blog/Sany.

[15] Kamil Wais. "Gender Prediction Methods Based on First Names with genderizeR". In: *The R Journal* 8.1 (2016), pp. 17–37. URL: https://journal.r-project.org/archive/2016/RJ-2016-002/index.html.

Classification Results

Class	Full Names NB	Names Ending NB	NB	API	All
n	99,845	76,277	**99,845**	8,262	**97,723**
Male	11.05%	49.36%	**48.75%**	56.46%	**54.59%**
Female	12.18%	33.03%	**37.42%**	12.84%	**39.32%**
Unisex	0.37%	4%	**3.43%**	7.7%	**4.15%**
Unknown	76.4%	13.61%	**10.4%**	23%	**1.94%**

Table 7: Classifications results with respect to unique names.