GENDER TRENDS IN COMPUTER SCIENCE AUTHORSHIP

A PREPRINT

Lucy Lu Wang, Gabriel Stanovsky, Luca Weihs, Oren Etzioni
Allen Institute for Artificial Intelligence
Seattle, Washington, USA

June 20, 2019

ABSTRACT

A comprehensive and up-to-date analysis of Computer Science literature (2.87 million papers through 2018) reveals that, if current trends continue, parity between the number of male and female authors will not be reached in this century. Under our most optimistic projection models, gender parity is forecast to be reached by 2100, and significantly later under more realistic assumptions. In contrast, parity is projected to be reached within two to three decades in the biomedical literature. Finally, our analysis of collaboration trends in Computer Science reveals decreasing rates of collaboration between authors of different genders.

1 Introduction

This paper presents a comprehensive and up-to-date analysis of gender trends in the Computer Science literature (ranging from 1970 through 2018).[1] Specifically, we aim to address the following questions regarding gender and authorship in the Computer Science literature:

- How is gender balance among authors changing over time?
- When might gender parity be reached among authors?
- How is gender associated with co-authorship?

We answer these questions by performing an automated study of literature meta-data from Computer Science conferences and journals (2.87 millions papers), utilizing data from the Semantic Scholar academic search engine.[2] To provide a basis for comparison, we also analyze papers from the top 1,000 Medline journals by citation count (11.63 million papers), and compare the trends observed in Computer Science to those in the biomedical literature.

2 Data

Corpus	Total papers (millions)	Total author-paper units (millions)	Average author per paper	Unique first names
Computer Science	2.87	8.24	2.87	186116
Medline	11.63	47.66	4.10	439981

Table 1: Corpus statistics for Computer Science and Medline.

Our analysis was performed over the Computer Science and Medline corpora and their meta-data. The corpora contain papers published between 1970 and June 2018, and associated metadata such as title, abstract, authors, publication venue, and year of publication. Summary statistics for both corpora are given in Table 1. The Computer Science

[1] We acknowledge that gender is not binary, but for the sake of this large-scale study–we adopt a simplified view of gender as binary and rely on first names as an approximate proxy for the author's gender.

[2] https://www.semanticscholar.org/

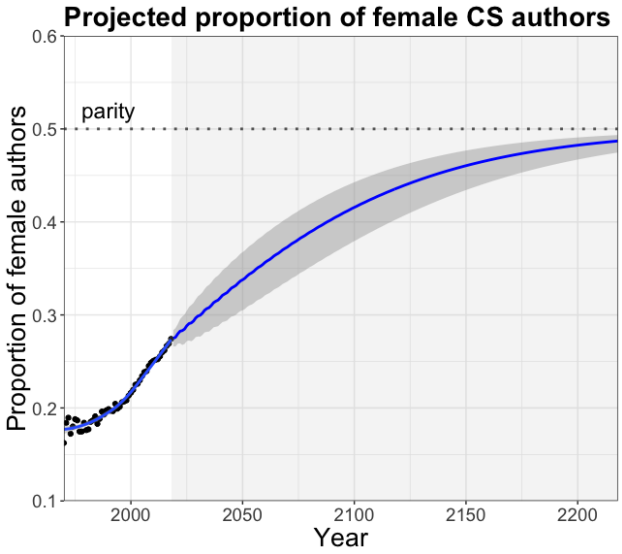

Figure 1: At current rates of growth, the proportion of female authors is predicted to reach 0.45 around 2137 (95% Confidence Interval: [2109, 2172]). The trend line is given by an ARIMA projection with 95% confidence intervals.

corpus consists of 2.87 million papers retrieved from conferences and journals in Computer Science. Publication and author metadata are automatically derived by Semantic Scholar from DBLP.[3] The Medline corpus consists of 11.63 million papers from the top 1000 Medline-indexed journals as determined using overall citation count.

The author list is extracted from all publications and compiled into a list of first names. We use Gender API[4] to perform gender lookup for each name. Gender API is a large online database of known name-gender relationships derived by linking publicly available governmental data with social media profiles in various countries. For each name, Gender API outputs the predicted binary gender (*female* or *male*), along with the accuracy associated with the prediction and the number of samples used to arrive at that determination. Authors for whom only first initials were available (less than 0.5% of all authors in our corpora) were excluded from analysis.

Because many names are gender-ambiguous, we use the accuracy returned by Gender API to represent each author as a composite of male and female. For example, the first name Matthew is determined to be male with an accuracy score of 100, the maximum. This result is unambiguous. The name Taylor, however, is determined to be female but only

[3] https://dblp.uni-trier.de/
[4] https://gender-api.com/

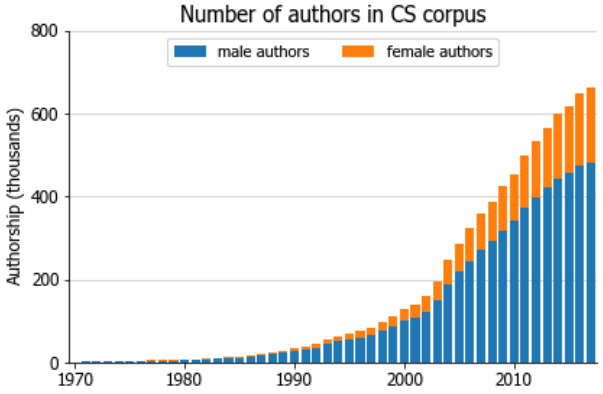

Figure 2: The total number of male and female authors in the Computer Science corpus over time.

receives an accuracy score of 55. The accuracy is used to generate a two probabilities for each name, (m, f), where m is the probability of the associated author being male, f is the probability of the associated author being female, where $m + f = 1$. In this example, each author with the first name Matthew will be represented with the probability tuple $(1.0, 0.0)$, and each author with the first name Taylor will be represented as $(0.45, 0.55)$.

Most papers are authored by more than one individual. For the purposes of our analysis, each author-paper pair is treated as one unit. A single-author paper yields one author-paper pair; a three-author paper yields three author-paper pairs and so on. In the Computer Science corpus, the average number of authors is 2.87 per paper. Average authors per paper increased from approximately 1.4 per paper in 1970 to approximately 3.5 in 2018.

3 Analyses

We perform two types of analysis on this data. First, we analyze publication trends, examining the proportion of female authors over time (Section 3.1). To identify when gender parity may be reached, we project the proportion of female authors based on current trends. Here, we define parity as the proportion of female authors falling within 10% of 0.5, within the range of 0.45-0.55. Second, we study the interactions between authors in the community through co-authorship as reflected in our data (Section 3.2).

3.1 Authorship analysis

The proportion of female authors over time is used to determine the trend towards gender parity. The number of female authors in any year is computed as the sum of probabilities f over the author-paper units of that year, and the number of male authors is correspondingly generated as the sum of probabilities m. The proportion of female authors for each year F_t is computed as the number of female author-paper units divided by the total number of author-paper units for the corresponding year. We compute projections by performing an autoregressive integrated moving average (ARIMA) analysis, a commonly-used method for creating forecasting models [1]. We use the auto ARIMA function in the R 'forecast' package [2], which automates the selection of ARIMA model order, with a preference for simple models with lower order.

The growth in gender proportion should observe logistic behavior, where a stable equilibrium will eventually be reached in gender balance. We first apply σ_α^{-1}, the inverse of the α-scaled sigmoid (or logit) function $\sigma_\alpha(x) = \alpha/(1 + \exp(-x))$, to map the gender proportion into the real line so that the data is more amenable to linear approximation. We call α the expected equilibrium proportion parameter. This transform generates $y_t = \sigma_\alpha^{-1}(F_t)$, where F_t is the proportion of female authors per year. We then fit a non-seasonal ARIMA model with parameters p, d, and q for the transformed process y_t represented by the following equation:

$$\phi_p(B)(1 - B^d)y_t = c + \theta_q(B)\varepsilon_t \tag{1}$$

where B is the backshift operator, which shifts by one to the previous time point, and ε_t is zero-centered, normally distributed noise [2].

Finally, we obtain the forecast in the original domain using a sigmoid transform over the projected values, applying σ_α to y_t for $t > 2018$. We let $\alpha = 0.5$ so that $\sigma_{0.5}$ has minimum and maximum values of 0 and 0.5 respectively. This constrains the projected values to be between 0 and the expected equilibrium proportion of 0.5. The 95% predictive interval is computed and shown for all projections. Note that α represents the proportion of female authors we expect in the long run. An equilibrium proportion of 0.5 indicates that we expect the authorship makeup to eventually stabilize at around 50% men and 50% women. An equilibrium proportion of 0.9 indicates that we expect the authorship makeup to eventually stabilize at around 10% men and 90% women. Trends toward equality suggest that the former is more plausible than the latter. As is further elaborated in Section 4.1, we perform a sensitivity analysis to determine the effect of the selected α parameter on the year in which parity is expected to be reached.

3.2 Co-authorship analysis

Co-authorship is computed for each unique pair of author-paper pairs for each paper. If a paper has n authors, $\binom{n}{2}$ co-author pairs are generated. Given a co-author pair (n_1, n_2) and associated gender probabilities:

$$\begin{array}{lcl} n_1 & \rightarrow & (m_1, f_1) \\ n_2 & \rightarrow & (m_2, f_2) \end{array} \tag{2}$$

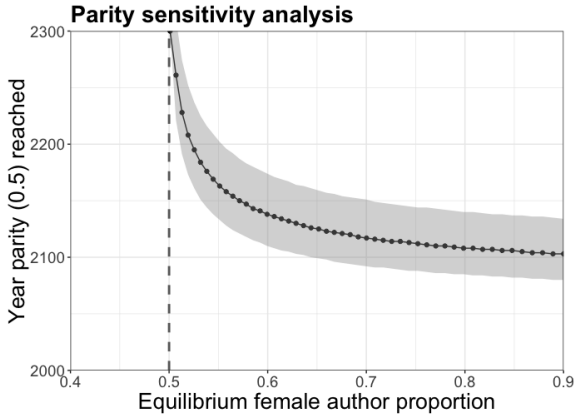

Figure 3: The equilibrium female author proportion parameter affects the year that parity is reached. The expected year for reaching exact parity (the first year in which the female author proportion equals or exceeds 0.5) is shown along with 95% confidence intervals.

we compute three probabilities, p_{mm}, p_{mf}, and p_{ff}, corresponding to the possible gender combinations, i.e., between two male authors, one male and one female author, and two female authors respectively. These probabilities are:

$$\begin{aligned} p_{mm} &= m_1 m_2 \\ p_{mf} &= m_1 f_2 + f_1 m_2 \\ p_{ff} &= f_1 f_2 \end{aligned} \tag{3}$$

where $p_{mm} + p_{mf} + p_{ff} = 1$. The total number of male-male, male-female, and female-female co-author pairs for each year is computed by summing each of the three above probabilities over all co-authorship pairs of that year.

We then assess the number of same-gender and different-gender collaborations over time. The results are measured as a deviation from the expected, where the expected co-authorships are determined by sampling from the numbers of female and male authors active in a given year, assuming the same number of collaborations per year as observed in our data. The total number of extra or missing collaborations is computed as the difference between the observed counts of each type of collaboration and the expected value. To show rates of change, we also compute the ratio between observed and expected collaborations (O/E) of each type.

4 Results

The 2.87 million papers in the Computer Science corpus yield 8.24 million author-paper units.

4.1 Authorship trends

Figure 2 shows the number of female and male authors over time. The total number of authors is increasing over time, along with the proportion of female authors.

Figure 1 shows the projected proportion of female authors in the Computer Science corpus. The projected growth in female author proportion is computed using ARIMA, with model order $(p, d, q) = (2, 1, 2)$. Residuals of the fit line appear normally distributed, and are not significant under the Shapiro-Wilk Normality Test (W = 0.98, p-value = 0.68) [3]. Based on these projections, the proportion of female authors in Computer Science is predicted to reach 0.45 around 2137 (95% CI: [2109, 2172]), more than 115 years from now.

Figure 3 shows a sensitivity analysis over the equilibrium female author proportion parameter α. This analysis shows the year in which parity is first reached at each equilibrium proportion; note that when $\alpha = 0.5$, exact 50/50 parity is, by definition, never attained in finite time. We therefore report the time at which the female author proportion surpasses 0.45, within 10% of exact parity. When the equilibrium proportion is expected to favor women over men (above 0.5), the year in which parity is reached occurs earlier. Even with the aggressive projection that women will

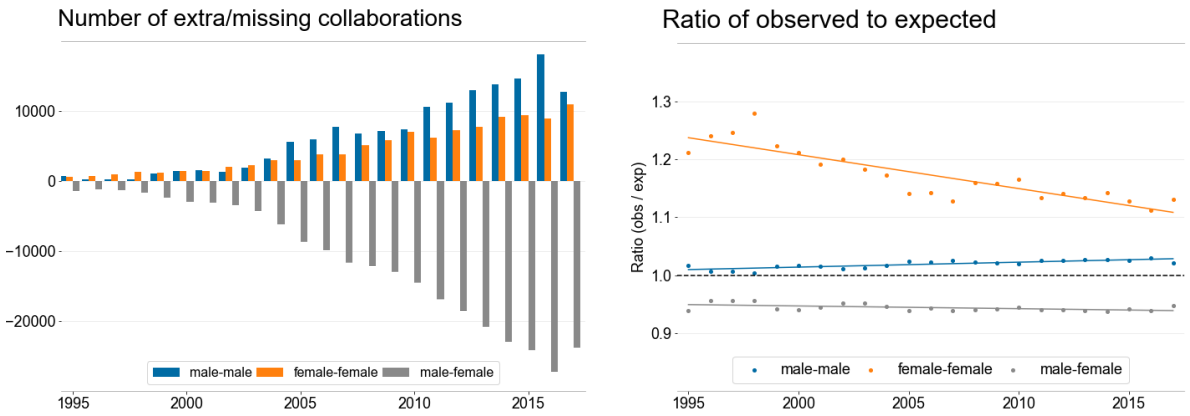

Figure 4: The difference (*left*) and ratio (*right*) between observed and expected same- and different-gender co-authorships in Computer Science since 1995.

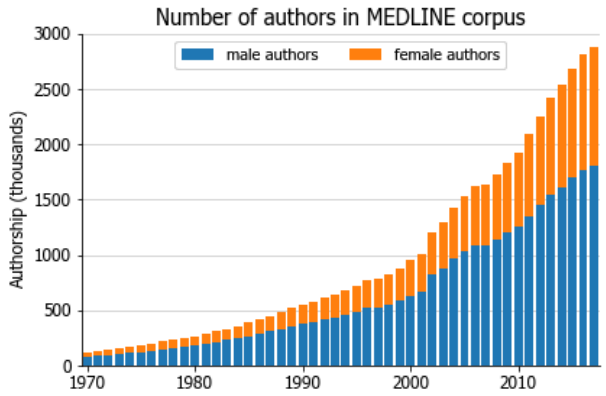

Figure 5: The total numbers of female and male authors in the Medline corpus.

eventually author 90% of all publications, the expected year in which parity will be reached at current rates of growth is still around 2100.

4.2 Co-authorship trends

The number of same- and different-gender co-authorships in Computer Science were computed for each year. Figure 4 shows the number of extra and missing same- and different-gendered collaborations since 1995. There are more same-gender co-authorships than would be expected among both men and women, and less different-gender co-authorships than would be expected. In recent years, more than 20,000 different-gender collaborations per year were missing when compared to expected numbers.

The observed to expected ratio shows pessimistic collaboration trends. Although both men and women are more likely to collaborate with authors of their own gender (positive O/E), the degree of same-gender preference is declining among female authors but increasing among male authors. At the same time, the different-gender collaboration gap (O/E < 1.0) is increasing in size, such that in recent years, only around 94% of expected different-gender collaborations are observed. In other words, although there are more opportunities for cross-gender collaboration in recent years (due to the increase in female scientists working in the field), the observed number of cross-gender collaborations has not increased as expected.

4.3 Comparison with Medline

The Medline corpus of 11.63 million papers yield 47.66 million author-paper units. Figure 5 shows the number of female and male authors in the Medline corpus. Figure 6 shows the projected proportion of female authors forecast using ARIMA, with model order $(p, d, q) = (0, 1, 0)$. A discontinuity can be observed in the Medline corpus data in 2002. This is due to the requirement of full author names in Medline-indexed records beginning in publication year 2002 [4]. The drop in proportion in 2002 shows that Medline journals not using full names for authors contributed to the false appearance of a high representation of female authors prior to 2002. Consequently, the ARIMA projection is computed using only the proportion data since 2002. The projection forecasts the proportion of female authors to surpass 0.45 around 2048 (95% CI: [2045, 2051]), a bit over 25 years from now. Because of the large number of authors in the Medline corpus since 2002, the confidence intervals for this projection are quite narrow.

5 Discussion

Our analysis of the Computer Science literature reveals persistent patterns of inequality in gender and academic authorship. Although gender balance is improving, progress is slower than we had hoped.

5.1 Limitations

Inferring gender from names is imperfect, and all gender-inference tools are subject to biases. Several studies have described and measured the differences between these services [5, 6]. Based on results in Santamaría and Mihaljević, Gender API has the lowest overall error rate but was slightly biased toward under-representation of females in their evaluation, in other words, the number of women estimated may be slightly lower than in reality. However, this bias may be offset by our sampling bias, since the population of Computer Science authors is unlikely to be an unbiased sample of the general population, or the subset of the general population whose names were used to construct the database behind Gender API. We attempted to mitigate some of these biases by treating the output of Gender API as probabilistic. To assess the accuracy of Gender API, we validated the predictions for the 50 names most commonly predicated as male and 50 names most commonly predicted as female, and found them to match our expectations.

The proportion of authors with high uncertainty Gender API results has also grown in our corpus over time. As evident in Figure 7, our average confidence in gender prediction decreased from about 90% in 1970 to 85% in 2018. While Gender API's average prediction confidence on our corpus is still high, this trend may pose a challenge for similar analyses in the future. Upon inspection of the data, we attribute this to the growing number of East Asian authors

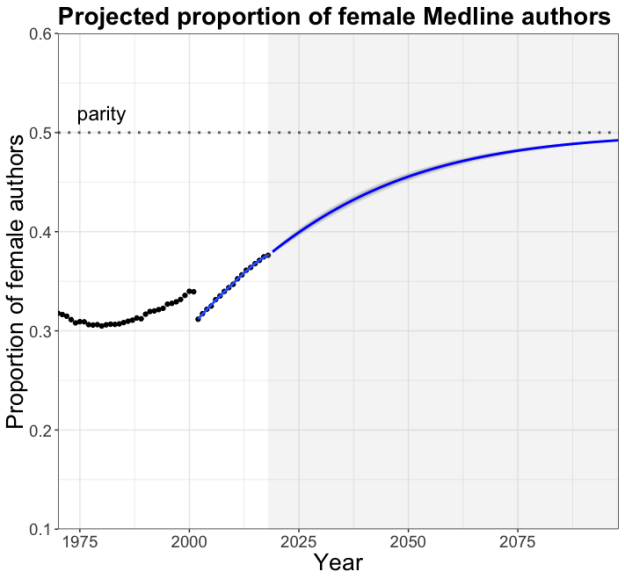

Figure 6: The proportion of female authors in the Medline corpus is projected to surpass 0.45 around 2048 (95% Confidence Interval: [2045, 2051]) based on ARIMA projections. The 95% confidence interval around the projection is plotted, but the error is small.

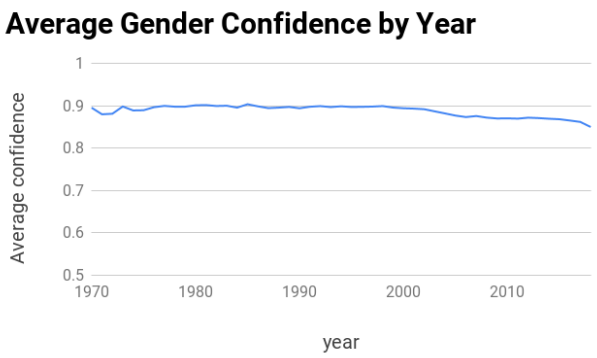

Figure 7: The average author gender confidence of Gender API on the Computer Science corpus, per publication year.

publishing in recent years. East Asian first names, especially when subject to romanization, can be quite gender ambiguous. We believe that by representing each author as a composite of male and female based on probability, we offset some of the issues associated with the increasing numbers of ambiguous names in our corpus over time. However, the authors of Computer Science literature are unlikely to be an unbiased sub-sample of the broader population, and this assumption may introduce some error into our analysis.

We also recognize the limitations of using author-paper pairs as our units of measure. We do not distinguish between a person who is a single author on a paper, and a person who co-authors with many others. This biases our data by over-weighting authors in papers with more authors. Similarly, in our analysis of collaboration, we take each combination of authors for a paper as a collaborating pair, which over-weights again papers with more authors. In the Computer Science corpus, we observe an increase in the average authors per paper over time, growing to approximately 3.5 authors per paper in 2018. However, Computer Science papers are still generally authored by smaller groups of individuals in the lower single digits, and we believe the bias introduced by our usage of author-paper pairs or collaborating author pairs to be minimal.

Each author is also weighted equivalently in our analysis. We acknowledge the special recognition extended to first authors, last authors, and single authors, and previous studies have already shown the distinctions between these groups [7].

5.2 Previous work

Gender bias is a well documented and studied issue in academia. Studies have shown that existent and perceived gender bias may affect many aspects of career and academic success, including but not limited to a woman's choice of college major [8], crediting in scientific publications [9], access to mentorship [10, 11], and opportunities for collaboration [12]. All these factors and more can lead to biased representation of women in certain fields of study.

With the increasing digitization of scholarly communication and availability of publication-related metadata, scholars have been better able to quantify inequality in authorship. A 2012 analysis of 1.8 million papers from JSTOR, a large multi-disciplinary repository of academic literature, revealed that although gender gaps are shrinking in academic publications, women were found to be significantly underrepresented as last and single authors [7]. Elsevier, the largest publisher of academic manuscripts, in an analysis of data from Scopus and ScienceDirect, reported the presence of gender imbalance among authors and inconsistent trends towards equal representation among different fields [12]. A study in early 2018 confirmed continuing gender disparities among Nature Index journals, commonly considered some of the most reputable sources of academic literature, and in particular, limited representation of women among last authors, who are often perceived as more senior [13].

A study of gender bias in authorship conducted by Holman et al. projected the closing of the gender gap in various fields based on current trends [14]. Through analyzing 9.1 million articles from PubMed, the authors projected that gender parity would be reached in around 20 years in certain biomedical fields such as Molecular Biology, Medicine, or Biochemistry. Holman et al.'s analysis of a small corpus of Computer Science pre-prints from arXiv show that gender parity in Computer Science will be reached in more than 100 years from the present [14].

Major strides have been made to reduce gender disparities. The presence of an overall structure of sexism in academia continues to be debated [15, 16, 17], but many academic institutions recognize the problem and have sought to equalize admissions and hiring procedures. Evidence of movement toward equal representation in hiring and publication has been observed in some controlled settings [18, 19, 20]. How these observations translate into systemic changes remain to be seen. It is clear, however, that the rate of change in reducing the gender gap may be insufficient in many fields for parity to occur within several generations [14].

6 Conclusions

We performed a comprehensive analysis of the Computer Science literature (2.87 million papers) to evaluate gender trends among authors. Based on recent trends, the proportion of female authors in Computer Science is forecast to not reach parity in this century, and under more realistic assumptions—it may take far longer. We also observed lower than expected numbers of cross-gender collaborations, with the ratio of observed to expected decreasing over time.

Slow rates of growth in the proportion of female scientists in Computer Science continue to challenge women entering the field. Female scientists may face more challenges finding collaborators than their male counterparts due to the existing gender distribution of authors and observed co-authorship behaviors. We hope that these findings will motivate others in the field to evaluate their relationship to these gender biases and consider ways to improve the status quo.

Acknowledgements

We would like to thank Jonathan Borchardt, Matt Gardner, and Candace Ross for conducting the initial analysis that motivated this project. We would also like to thank Maarten Sap, Noah Smith, and Mark Yatskar for helpful comments on earlier drafts of this paper.

References

[1] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: Forecasting and control*. Prentice Hall, Englewood Cliffs, N.J., 3 edition, 1994.

[2] Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22, 2008.

[3] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

[4] Medline®data changes – 2002. *NLM Tech Bull*, Nov-Dec(323):e11, 2001.

[5] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *WWW*, 2016.

[6] Lucía Prieto Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, 2018.

[7] Jevin D. West, Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom. The role of gender in scholarly authorship. *PloS one*, 8(7):e66212, 2013.

[8] Rachael D. Robnett. Gender bias in STEM fields: variation in prevalence and links to STEM self-concept. *Psychology of Women Quarterly*, 2015.

[9] David F. Feldon, James L. Peugh, Michelle A. Maher, Josipa Roksa, and Colby Tofel-Grehl. Time-to-credit gender inequities of first-year PhD students in the biological sciences. In *CBE life sciences education*, 2017.

[10] Rochelle Decastro, Kent A. Griffith, Peter Anthony Ubel, Abigail J. Stewart, and Reshma Jagsi. Mentoring and the career satisfaction of male and female academic medical faculty. *Academic medicine : journal of the Association of American Medical Colleges*, 89(2):301–11, 2014.

[11] Natalie Schluter. The glass ceiling in NLP. In *EMNLP*, 2018.

[12] Gender in the global research landscape. Technical report, Elsevier, 2017.

[13] Michael H. K. Bendels, Ruth Mueller, Doerthe Brueggmann, and David Alexander Groneberg. Gender disparities in high-quality research revealed by nature index journals. *PloS one*, 13(1):e0189136, 2018.

[14] Luke Holman, Devi Stuart-Fox, and Cindy E. Hauser. The gender gap in science: How long until women are equally represented? *PLoS biology*, 16(4):e2004956, 2018.

[15] Jamie Lundine, Ivy Lynn Bourgeault, Jocalyn Clark, Shirin Heidari, and Dina Balabanova. The gendered system of academic publishing. *The Lancet*, 391(10132):1754–6, 2018.

[16] Jason R Boynton, Kristina Georgiou, Mark Reid, and Andrew Govus. Gender bias in publishing. *The Lancet*, 392(10157):1514–5, 2018.

[17] Jamie Lundine, Ivy Lynn Bourgeault, Jocalyn Clark, Shirin Heidari, and Dina Balabanova. Gender bias in academia. *The Lancet*, 393(10173):741–3, 2019.

[18] W. Mattieu Williams and Stephen J Ceci. National hiring experiments reveal 2:1 faculty preference for women on stem tenure track. *Proceedings of the National Academy of Sciences of the United States of America*, 112(17):5360–5, 2015.

[19] Erin Hengel. Publishing while female. are women held to higher standards? Evidence from peer review. *Cambridge Working Paper Economics*, 1753, 2017.

[20] Stephen J. Ceci and Wendy M. Williams. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8):3157–3162, 2011.