No evidence of any systematic bias against manuscripts by women in the peer review process of 145 scholarly journals

Flaminio Squazzoni[a], Giangiacomo Bravo[b], Pierpaolo Dondio[c], Mike Farjam[d], Ana Marusic[e], Bahar Mehmani[f] Michael Willis[g], Aliaksandr Birukou[h] Francisco Grimaldo[i]

Abstract

This article examines gender bias in peer review with complete data on 145 journals in various fields of research, including about 1.7 million authors and 740,000 referees. We reconstructed three possible sources of bias, i.e., the editorial selection of referees, referee recommendations, and editorial decisions, and examined all their possible relationships. In line with previous research, we found that editors were sensitive to gender homophily in that they tended to match authors and referee by gender systematically. Results showed that in general manuscripts written by women as solo authors or co-authored by women are treated even more favorably by referees and editors. This is especially so in biomedicine and health journals, whereas women were treated relatively less favorably in social science & humanities journals, i.e., the field in which the ratio of female authors was the highest in our sample. Although with some caveat, our findings suggest that peer review and editorial processes in scholarly journals do not penalize manuscripts by women. However, considering the complex social nature of gender prejudices, journals should increase gender diversity among reviewers and editors as a means of correcting signals potentially biasing the perceptions of authors and referees.

Keywords: peer review; gender bias; scholarly journals; editors; referees

1 Introduction

Scholarly journals are often blamed for gender discrimination in publications [1–3]. The fact that women have less prestigious academic positions and more problematic careers is often attributed to the publication gap [4, 5]. In the current hyper-competitive academic environment, this gap undermines academic prestige and resource allocation, with negative implications not only on the individual careers of women but also on the composition and evolution of the scientific community [6–8]. However, there is no consensus on the relation between the gender gap in publication rates between men and women and gender bias in peer review and journal editorial processes [9]. On the one hand, recent research has shown that women are systematically less involved in peer review and

[a]Corresponding author: flaminio.squazzoni@unimi.it. Department of Social and Political Sciences, University of Milan, Milan, Ital

[b]Department of Social Studies and Centre for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden

[c]Dublin Institute of Technology, Dublin, Ireland

[d]Department of Social Studies and Centre for Data Intensive Sciences and Applications, Linnaeus University, Växjö, Sweden

[e]University of Split School of Medicine, Split, Croatia

[f]STM Journals, Elsevier, Amsterdam, The Netherlands

[g]John Wiley & Sons, Oxford, United Kingdom

[h]Springer Nature, Heidelberg, Germany

[i]Department of Computer Science, University of Valencia, Burjassot, Spain

rarely appointed to prestigious editorial positions, dramatically affecting their recognition [1, 2, 10]. On the other hand, recent reports from journals in specific fields suggest that editorial processes do not discriminate against women [11, 12]. Unfortunately, findings are controversial [13, 14] especially because research is case-specific and has never been performed at a scale sufficient to provide insights on different fields of research [15–17].

Our study aims to improve our understanding of this *gender gap-gender bias* link by providing the first in-depth analysis of peer review and editorial processes in a large sample of scholarly journals. Thanks to an agreement on data sharing with some of the largest scholarly publishers [18], we collected complete and fully comparable temporal data on 145 scholarly journals, including almost 350,000 submissions by about 1.7 million authors and more than 760,000 reviews performed by about 740,000 referees (see Materials and Methods). These data allowed us to fully reconstruct three possible sources of bias (i.e., the editorial selection of referees, referee recommendations, and editorial decisions), and examine their possible relationships, while controlling for important confounding factors, such as journals' field of research, impact factor and peer review model.

2 Results

Table 1 shows the distribution of journals by fields of research in our sample, the proportion of women among authors and other summary statistics. Our data confirmed previous research on gender disparities in manuscript submissions and peer reviewing [1, 2, 10–12], with 75% of men among submission authors and 79% of men among referees. As expected, we found differences between journals from different research fields, with the most balanced rate in social science journals (38% women as authors and referees) and a greater gap in physics and related fields (19% women as authors and 16% as referees). In addition, women are less involved in peer review compared to their authorship rate in all domains except for social sciences (Tab 1), although this could simply reflect differences in the gender composition of the potential pool of authors and referees.

	Biomedicine & Health	Life Sciences	Physical Sciences	Social sciences & Humanities
Number of journals	55	24	50	16
N. of submissions	113421	31331	184315	19051
Desk rejections (proportion)*	0.25	0.34	0.40	0.41
First-round rejections (proportion)	0.46	0.35	0.41	0.50
Final rejections (proportion)	0.59	0.48	0.48	0.62
Women authors (proportion)	0.32	0.28	0.19	0.38
Women referees (proportion)	0.25	0.21	0.16	0.38

Table 1: Number of journals and frequency distribution of selected sample characteristics by field of research. *Data on desk rejections were available only from a sub-sample of journals (61 journals from one publisher).

Figure 1 shows an overview of the distribution of the final editorial decisions on manuscripts per gender of the first and last author and field of research. This picture suggests a certain degree of diversity among fields, e.g., manuscripts by women would be accepted more frequently in biomedicine, health science, and social science journals, less in life science journals. However, these descriptive statistics do not allow to consider the potential effect of important covariates — such as the journal's impact factor, the number of co-authors, and the review scores — , which would be essential to disentangle potential sources of bias during the editorial and the peer review process.

To examine these processes more systematically, we performed robust statistical analysis within

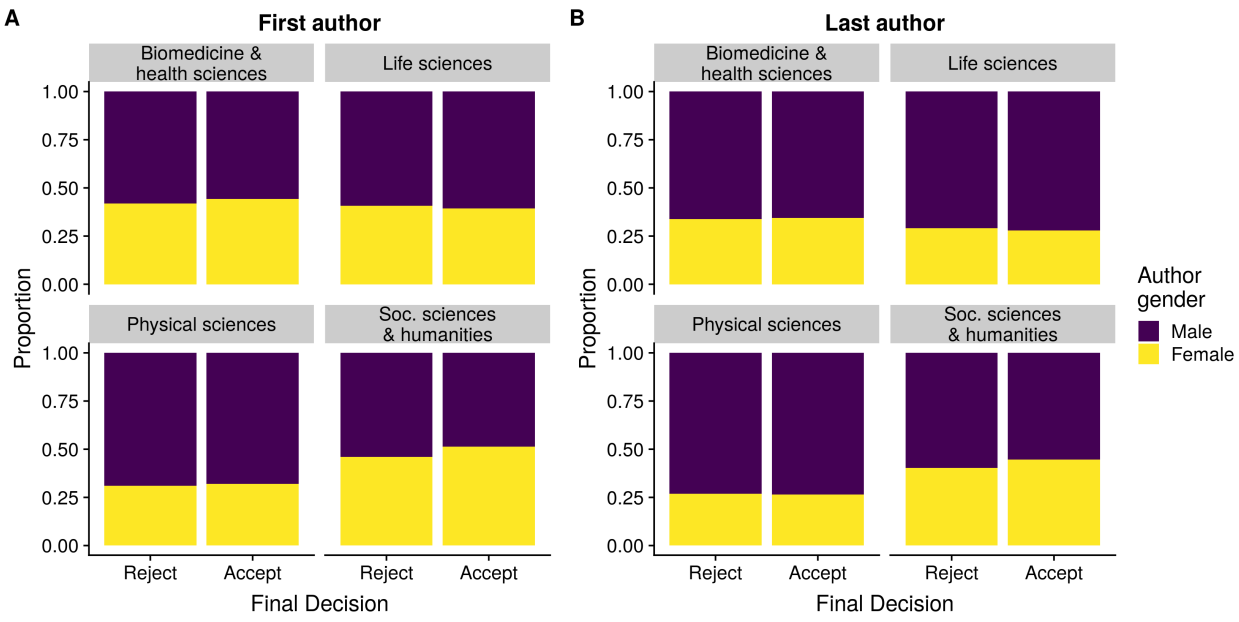

Figure 1: Distribution of final editorial decisions by gender of the first (A) and last author (B) .

a Bayesian framework and estimated different models on the dataset (see Materials and Methods). We first looked at the editorial process by considering each of the following steps separately: (i) the editorial selection of referees, (ii) the referee recommendations, and (iii) the editorial decision on the manuscript. All these steps included specific actions performed either by referees or editors that could reveal a bias. Following previous research and data availability, we considered both the position of women in the author list (i.e., whether they were first or last authors) and the women proportion among the authors as main predictors [1, 4, 14], while controlling for the proportion of women among the referees, the impact factor of the journal, the number of authors in each manuscript, and the type of peer review adopted by the journal [19, 20].

Given that the effect of many of these variables, and crucially of the first and last authors' gender, is likely to be different in each field of research, we estimated separated models for each field. This allowed us to consider field specificities, including the journal prestige and potential diversity of evaluation standards, through in-depth data that have never been available before in this type of research [9]. We then built a Bayesian-learning network model [21] to estimate the effect of complex interactions more systematically and understand the extent and persistence of gender bias across all steps of the editorial process (see Materials and Methods). Regarding the editorial selection of referees (step i), we found that in all fields of research, editors systematically assigned manuscripts with a higher proportion of women among the authors more preferably to female referees (SI, Tab. 1). This is consistent with [1, 22] and was confirmed after controlling for the number of authors in the manuscript, the journal's impact factor, and type of peer review model (single vs. double blind).

Furthermore (step ii), we found that female authors received systematically more positive reviews in biomedicine and health sciences, as well as in social sciences, whereas were less positively treated in life sciences (weak statistical effect) and physical sciences journals (strong statistical effect). Female referees tended to provide more positive recommendations than males in all fields but physical sciences. This effect was consistent after controlling for all other variables (SI, Tab. 2). The fact that our model could only explain a small fraction of the outcome variance (between 4% and 11%,

depending on the field of research), though many model coefficients were significant, suggests that other manuscript characteristics which we could not measure — e.g., its quality and content — had the strongest effect on referee recommendations. This effect was independent of any editorial matching or referee selection options.

To check whether our results were robust to alternative specifications of our gender variable, we estimated two further models that considered (i) whether a woman was first or last author of a manuscript (SI, Tab. 3) and (ii) whether effects were different for our five mutually-exclusive groups of authors: a man sole author, a woman sole author, all-male teams, all-female teams, and co-ed teams of authors (SI, Tab. 4) [4]. In general, our results show that the author gender did not have a consistent effect, although we found the emergence of complex patterns of interaction when the scientific field of journals and the specific composition of author groups were taken into account (for a more systematic analysis of these complex interactions, see the Bayesian-learning network presented below).

Regarding the final editorial decisions (step iii), we found that manuscripts with a higher proportion of women among authors were accepted more frequently in biomedical, health sciences and physical sciences journals (strong statistical effect), whereas no evidence of any effect of the gender variable was found in life sciences and social sciences journals. Note that in case of biomedical and physical sciences journals, the positive effect was robust across variation of contexts and controlling for the referee recommendations and the journal's field of research (Tab. 2). Furthermore, considering the review scores (for a synthesis of referee recommendations, see Materials and Methods), our models were able to explain over 80% of the outcome variance.

Alternative specifications of the gender variable did not lead to any systematic evidence of a bias against women in the peer review process, although resulting in less clear-cut results than in previous models (Tab. 3). When we considered the gender of the first author, we found that manuscripts by women were more favorably treated in physical sciences journals (strong statistical effect) and less in life sciences journals (weak statistical effect). Being the last author had no significant effect on acceptance, except for a weak negative effect in case of biomedical and health sciences journals. We did not find any systematic bias against women across journals and disciplines when considering the four author groups mentioned above (SI, Tab. 5).

Finally, to consider the whole editorial process in which indirect paths of bias may exist and given that complex interactions among variables could affect editorial decisions, we estimated a Bayesian-learning network including all the previous steps of the analysis. After learning coefficients and conditional probabilities through Maximum Likelihood estimation, our model was able to predict with 82% accuracy whether or not a manuscript would ultimately be accepted by the editor (see Materials and Methods). Figure 2 shows that after controlling for all direct and indirect effects of all variables, the effect of authors' gender on referee recommendations depended on the field of research. While recommendations for manuscripts with higher female proportion among authors were slightly more positive in journals in the social and biomedical and health sciences, they were slightly more negative in journals in life and physical sciences. However, note that, even when comparing the extreme cases where manuscripts were authored exclusively by women or men, our model predicted a change in review scores by less than 4%, showing that these effects were minimal.

Although we could not directly estimate the intrinsic quality of manuscripts (if this were possible even only in principle), we used the recommendations of referees as a control variable of the quality and used it to identify a bias of the editorial decision. Our results indicated that there was no systematic bias against women across fields of research. The Bayesian-learning model found that, after controlling for all other variables (including the recommendations), manuscripts by women were more likely to be accepted in journals of all discipline except social sciences, where we did not find any significant gender difference. To quantify the effect of gender, we used the model to

Variable	Biom. & health sc.	Life sc.	Physical sc.	Social sc.
(Intercept)	-6.22	-4.70	-7.07	-5.12
	[-6.63,-5.83]	[-6.05,-3.37]	[-7.97,-6.17]	[-6.07,-4.2]
	1:20000	1:20000	1:20000	1:20000
Women proportion (authors)	0.13	0.05	0.21	-0.07
	[0.02,0.24]	[-0.14,0.24]	[0.11,0.3]	[-0.29,0.16]
	103:1	2:1	20000:1	1:2
Women proportion (referees)	-0.15	-0.04	-0.04	-0.23
	[-0.24,-0.07]	[-0.21,0.12]	[-0.12,0.04]	[-0.45,-0.02]
	1:2856	1:2	1:6	1:59
Referee recommendation	6.02	6.18	6.09	5.82
	[5.91,6.13]	[5.94,6.42]	[6,6.19]	[5.47,6.18]
	20000:1	20000:1	20000:1	20000:1
Agreement	1.21	0.67	0.71	0.20
	[1.09,1.34]	[0.45,0.88]	[0.61,0.8]	[-0.12,0.53]
	20000:1	20000:1	20000:1	8:1
IF	-0.06	-0.14	0.06	-0.14
	[-0.11,0]	[-0.21,-0.07]	[0.02,0.1]	[-0.4,0.11]
	1:57	1:20000	832:1	1:6
N. of authors	0.00	-0.04	0.04	0.01
	[-0.01,0.01]	[-0.05,-0.03]	[0.03,0.05]	[-0.03,0.05]
	2:1	1:20000	20000:1	3:1
N. of referees	-0.18	-0.16	-0.10	-0.30
	[-0.23,-0.14]	[-0.23,-0.09]	[-0.13,-0.07]	[-0.42,-0.18]
	1:20000	1:19999	1:20000	1:20000
PR type: single-blind	0.53	0.12	1.19	1.09
	[0.1,0.96]	[-1.23,1.47]	[0.28,2.11]	[-0.39,2.59]
	105:1	1:1	162:1	14:1
N. of revision rounds	4.09	3.67	3.99	3.76
	[4.04,4.15]	[3.58,3.77]	[3.95,4.04]	[3.62,3.89]
	20000:1	20000:1	20000:1	20000:1

Table 2: Logistic mixed-effects models on the final editorial decision (accept) by field of research using the gender ratio as predictor. Mean estimate, 95% CI, and Bayes factor ($\beta > 0$) are reported for each variable.

predict the final acceptance of all manuscripts in our dataset with the hypothetical scenario that all authors were either male or female. In case of biomedical and health sciences journals, manuscripts written by women were predicted to be 5% more likely to be accepted than manuscripts written by men (women were predicted to be accepted in 45% of cases). While in case of life and physical sciences journals, this probability decreased to 1.5% (for women the prediction was 53% in both fields), in case of social sciences journals, the percentage of the premium was close to zero (with a predicted acceptance of 38% of manuscripts). Interestingly, this suggests that women are treated less favorably exactly in the field of research where the ratio of women among authors is the highest (38% in social sciences vs. 19% in physical sciences). Fig. 3) shows the predicted editor decisions by authors' gender, controlling for different review scores. Finally, the Bayesian-learning network further confirmed that editors are sensitive to gender homophily and tended to match authors and referees by gender systematically.

Although not directly related to the scope of our research, which was specifically the peer review process, we also tested whether desk-rejections by editors could be predicted by the authors' gender. Table 4 shows that manuscripts with a higher proportion of women among authors were desk-

Variable	Biom. & health sc.	Life sc.	Physical sc.	Social sc.
(Intercept)	-6.12	-4.50	-7.02	-5.29
	[-6.53,-5.7]	[-5.84,-3.16]	[-7.96,-6.09]	[-6.28,-4.32]
	1:20000	1:20000	1:20000	1:20000
First author woman	0.00	-0.10	0.10	-0.06
	[-0.07,0.07]	[-0.22,0.02]	[0.03,0.16]	[-0.26,0.13]
	1:1	1:18	768:1	1:3
Last author woman	-0.06	-0.05	-0.04	0.04
	[-0.13,0.01]	[-0.18,0.08]	[-0.11,0.02]	[-0.15,0.22]
	1:16	1:3	1:8	2:1
Women proportion (referees)	-0.14	-0.06	-0.03	-0.19
	[-0.23,-0.04]	[-0.25,0.13]	[-0.13,0.07]	[-0.43,0.04]
	1:302	1:3	1:3	1:16
Referee recommendation	6.02	6.25	6.06	5.78
	[5.89,6.14]	[5.97,6.53]	[5.93,6.19]	[5.39,6.18]
	20000:1	20000:1	20000:1	20000:1
Agreement	1.21	0.63	0.65	0.35
	[1.06,1.35]	[0.39,0.89]	[0.52,0.77]	[0,0.71]
	20000:1	20000:1	20000:1	38:1
IF	-0.06	-0.14	0.04	-0.17
	[-0.12,0]	[-0.22,-0.05]	[0,0.09]	[-0.45,0.11]
	1:33	1:1817	28:1	1:8
N. of authors	0.01	-0.04	0.05	0.02
	[0,0.01]	[-0.06,-0.03]	[0.04,0.06]	[-0.02,0.07]
	6:1	1:20000	20000:1	6:1
N. of referees	-0.19	-0.2	-0.14	-0.29
	[-0.24,-0.14]	[-0.29,-0.11]	[-0.18,-0.1]	[-0.42,-0.16]
	1:20000	1:20000	1:20000	1:20000
PR type: single-blind	0.54	0.10	1.34	1.09
	[0.11,0.97]	[-1.24,1.44]	[0.41,2.28]	[-0.41,2.6]
	143:1	1:1	391:1	14:1
N. of revision rounds	4.10	3.71	4.02	3.83
	[4.04,4.16]	[3.6,3.82]	[3.96,4.08]	[3.69,3.99]
	20000:1	20000:1	20000:1	20000:1

Table 3: Logistic mixed effects models on the final editorial decision (accept) by field of research using the first and last author's gender as predictors. Mean estimate, 95% CI, and Bayes factor ($\beta > 0$) are reported for each variable.

rejected more often in life and physical sciences journals and less in all other fields. However, note that the sample of this model was smaller because data on editorial desk rejections were not reported for many journals. This result does not necessarily imply that editors intentionally discriminated authors based on their gender. Manuscripts by women could have been treated differently because they were of different quality or because of any other confounding factors which we could not control with our data. Finally, as expected, we found that rejections were systematically higher in journals with higher impact factor, while the likelihood of a desk-rejection decreased with an increase of the number of manuscript authors.

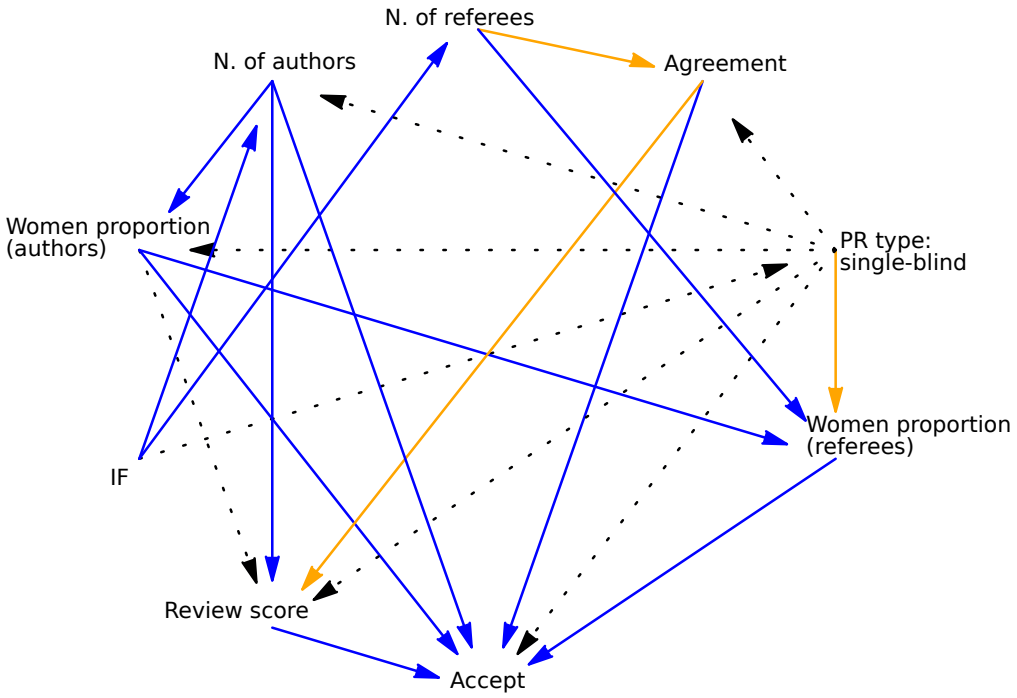

Figure 2: Learned structure of the Bayesian network. For the sake of readability, we did not report the scientific field effect, which was linked to all nodes. Orange arrows indicate a negative relationship, blue positive (dotted black, if the sign depends on the scientific field taken into consideration).

Variable	Biom. & health sc.	Life sc.	Physical sc.	Social sc.
(Intercept)	-0.96 [-1.53,-0.45] 1:20000	-0.89 [-1.15,-0.64] 1:20000	-1.71 [-2.32,-1.23] 1:20000	-0.50 [-1.12,0.06] 1:26
Women proportion (authors)	-0.20 [-0.26,-0.15] 1:20000	0.08 [-0.01,0.16] 24:1	0.25 [0.16,0.35] 20000:1	-0.11 [-0.2,-0.01] 1:75
IF	0.36 [0.33,0.4] 20000:1	0.15 [0.12,0.17] 20000:1	0.53 [0.45,0.6] 20000:1	0.29 [0.2,0.37] 20000:1
N. of authors	-0.06 [-0.07,-0.06] 1:20000	-0.06 [-0.07,-0.05] 1:20000	-0.11 [-0.13,-0.09] 1:20000	-0.13 [-0.15,-0.11] 1:20000

Table 4: Logistic mixed effects model on desk-rejections by field of research. Mean estimate, 95% CI, and Bayes factor ($\beta > 0$) are reported for each variable.

3 Conclusions

Although we could not perform a large-scale, across-journal randomized experiment and worked only on existing journal data, our findings indicate that manuscripts submitted by women or co-authored by women are generally not penalized during the peer review process. We found that manuscripts by all women or cross-gender teams of authors had even a higher probability of success in many cases. This is especially so in journals in biomedicine, health and physical sciences. However, considering that we did not have an objective or pre-defined estimation of the quality of manuscripts (if any)

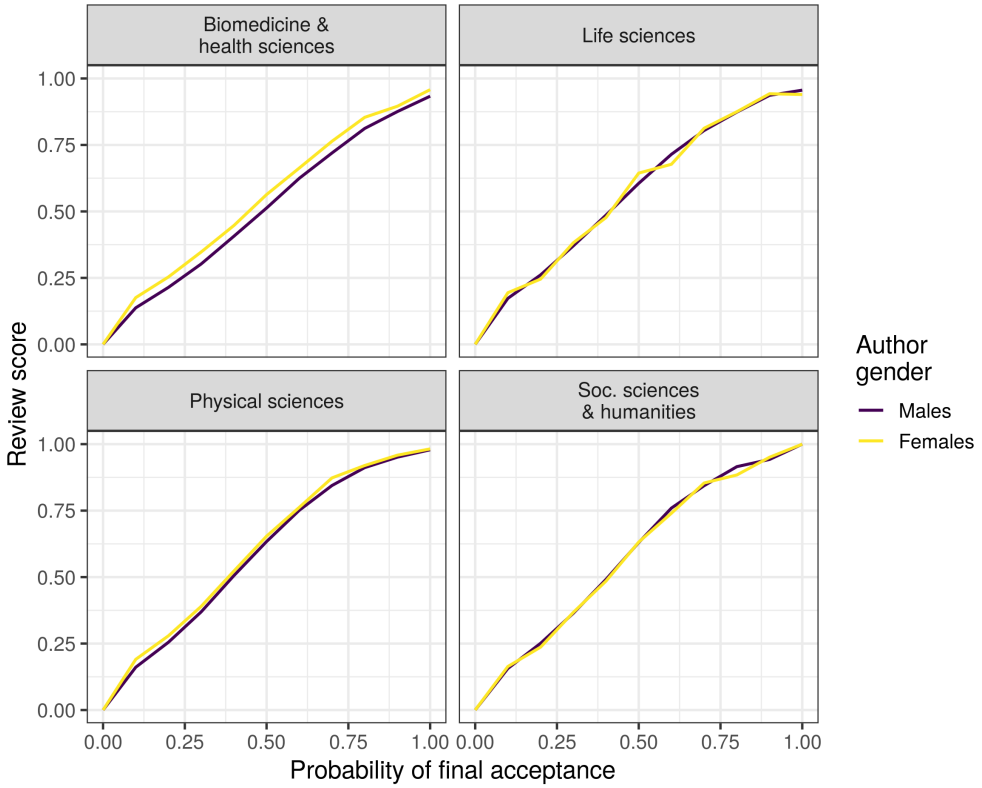

Figure 3: Bayesian network predictions of the rejection probability by author gender, referee recommendation score panels and field of research.

and could use only referee recommendations as an indication, this positive inclination by referees and editors could simply reflect some intrinsic characteristics of manuscripts. Previous research suggested that women could be inclined to invest more in their manuscripts to prevent expected editorial bias [4,23], which could in principle also explain why they submit fewer manuscripts [11,12]. This seems to reflect a consistent social mechanism, which has been also found in other competitive contexts, such as politics: when women expect to be discriminated in a competitive setting, they will be more determined, invest more in developing their abilities, and will eventually outperform men [24]. In this respect, the fact that manuscripts by cross-gender teams of authors received systematically more positive treatments in our sample could even reveal an exploitation opportunity by men, who benefit from collaborating with female colleagues.

Unfortunately, while the potential positive effect of higher inclusion of women in scientific networks has been found also in other studies [4, 25, 26], our dataset did not permit us to control for distorsions in the potential pool of authors and referees available in each journal, age cohorts or other (institutional/personal) status characteristics to understand if these potentially positive effects penalize older women and/or authors from less prestigious institutions [2]. It is worth noting that besides the lack of an objective measure of the quality of manuscripts, which is problematic and probably even impossible to establish consistently across fields, the size of the unobserved variables in our models is not small. Some of them could be at least potentially minimized with extensive data search, e.g, the effect of authors' academic affiliation, others are impossible to capture, e.g., the role of authors' seniority and reputation, especially considering the scale and the across-field nature of our dataset. For instance, it is extremely difficult to estimate the gender composition of

various communities to calculate the potential pool of authors and referees in each journal, while we do not have robust proxies of authors' investment in manuscripts to estimate gender differences in submissions and prolificity.

In any case, our findings do not mean that peer review and journals are free from biases. For instance, the reputation of certain authors, the institutional prestige of their academic affiliation, not to mention authors' ethnicity or the type of research submitted could influence the process, and these factors could also have gender implications [20, 27, 28]. On the one hand, the fact that women would submit and review less has negative implications on academic careers of women as this influences merit and promotion [29]. Here, data on the demographic composition of each disciplinary community, data on request and willingness to review at the journal level could help to complete our picture. On the other hand, these distortions could reflect built-in gendered norms and expectations, which could persist and be reproduced either consciously or not, even when their expected 'true' effects have disappeared [23,30]. Considering the persistent and usually non-acknowledged obstacles that women still face in hyper-competitive science [31], these expectations can be consistent even if the editorial processes of a set of journals are not objectively biased against women. Here, it is possible that the fact that we found that editors show a preference for matching authors and referees by gender could be explained exactly as an implicit means to counter-balance these expectations.

Here, our findings suggest that promoting more gender diversity in editorial teams and pools of referees could help scholarly journals to correct ambiguous indications about gender bias to potential authors and referees and so stimulate female inclusion and participation [32–34]. While diversity is beneficial for science and innovation per se [33], in this case, it would also be a signal that could contribute to modify the social construction of gender categories in science and help scholarly journals to increase submission rates by women. Unfortunately, our research could not examine these complex expectations and norms characterizing academic life across all its spectrum, including academic choices of priorities and specialties [35,36]. Studies capable of combining academic standards of promotion and the effect of author prestige and institutional affiliation on editorial process in scholarly journals are required to examine the complex nexus of gender discrimination (and even other sources of bias) in academia [23], including reconstructing the *gender gap-gender bias* link in a comprehensive manner. However, this calls for the problem of data availability [37]. While data sharing on editorial processes of journals should be encouraged more systematically on a large scale with collaboration between publishers and independent research groups [15, 18, 38–40], examining structural mechanisms that determine academic opportunities requires data integration from various sources (i.e., funding agencies, academic institutions, and scholarly citation databases). Only collaboration efforts on data sharing by various stakeholders will help us to grasp all pieces of this gender puzzle.

A Materials and methods

A.1 Data overview

Our dataset included internal data of 157 scholarly journals between 2010 and 2016, of which 61 were in biomedicine and health, 50 in physical sciences (including engineering and computer science), 24 in life sciences and 22 in social sciences and humanities. Details on journal selection and protocol for data sharing are provided in the SI file. Data consisted of all actions or events performed by one of the journal editors, such as inviting referees, receiving reviews or deciding about manuscripts. They included 753,909 submitted manuscripts, of which 389,431 (51.7%) were sent out to referees.

To ensure better comparability of peer review and editorial standards, in our analyses we only considered journals included in the Journal Citation Report based on the Web of Science and having

assigned an impact factor (98% of our observations, see SI, Fig. 1). The resulting dataset included 145 journals and 348,223 submissions. Because of a few missing observations in the data the actual numbers of complete observation used in the analysis were 348,118 (Tab. 1). These included a total of 1,689,944 authors and 745,693 referees, with an average of 2.1 completed reviews per manuscript.

The dataset includes the following variables:
Manuscript ID Unique manuscript identifier
SubmissionDate Initial submission date
JournalID Unique journal identifier
ScientificArea Journal's field of research (scientific area)
PRType Peer review type
IFRounded Journal's impact factor rounded to integer (this was to ensure journal's anonymity)
nAuthors Number of authors
NumRounds Number of review rounds
Agreement Referee agreement score
nRev Number of referees
RevScore Review score
AutRatFem Ratio of female authors
RevRatFem Ratio of female referees
FirstAuthorGender Gender of the first author
LastAuthorGender Gender of the last author
FinalDecision Final editorial decision

The number of manuscripts reviewed by these journals was approximately constant over time, with about 50,000 editorial decisions per year, and a majority of records from physics and biomedicine and health journals (SI Fig. 2).

Given that we aimed to focus on the peer review process, we considered each submitted manuscript as our unit of analysis. Statistics showed that the proportion of accepted paper varies across scientific fields: from 51.9 % in life sciences to 37.7% in social sciences (SI Fig. 3).

Referee recommendations were combined so that a *review* and an *agreement* score were calculated for each manuscript [19]. The former was bounded in the $[0, 1]$ interval, independently of the number of referees, with higher values reflecting more positive referee recommendations. Following [19], the agreement score was calculated in the same interval, with higher values meaning a stronger agreement between referee recommendations [19].

More specifically, in order to calculate review scores, we first re-coded each referee recommendation (which sometimes appeared as non-standard expressions in our database) in a standard ordinal scale *reject*, *major revisions*, *minor revisions*, *accept*. We then derived the set of all possible unique combinations of recommendations for each manuscript (from now on, the 'potential recommendation set'). Using this set, we counted the number of combinations that were clearly less favourable ($\#worse$) or more favourable ($\#better$) than that actually received by the manuscript (e.g.,{accept, accept} was clearly better than {reject, reject}). Finally, we calculated the score of each manuscript as follows:

$$reviewScore = \frac{\#worse}{\#better + \#worse} \tag{1}$$

Note that while [19] calculated a *disagreement* score, here we assumed an *agreement* score for each manuscript, i.e., one minus the number of referee recommendations that should be changed to reach a perfect agreement between referees divided by the number of referees assigned to the manuscript. This permitted full comparability between manuscripts receiving a different number of reviews.

A.2 Data availability

Our dataset is made available as a SI file (an SV source file) with all records required to rerun our analysis.

A.3 Statistical analysis

We estimated our mixed effects models using the *R* 3.6.1 platform [41]. Our plots were generated using the *ggplot2* package on the same platform. In all linear and logistic mixed-effect models, we included random effects for journals. We tested all model specifications including nested random effects for journals by considering the potential distortions due to sampling by publishers and found no effect on results. Note that due to the compliance with the data sharing protocol, we did not report details here to avoid journal identification. Mixed effects models were estimated using the *brms* package [42] and are the outcome of four independent chains, each including 10,000 iterations (5000 burn-in + 5000 sampling). To ensure that the estimates are reliable, we checked that all scale reduction factors ($\hat{\mathrm{R}}$) [43] were below 1.01. In each table, we reported the coefficients' mean estimates, 95% credible intervals (CI), and the Bayes factor corresponding to the hypothesis $\beta > 0$. The interpretation of Bayes factors was done following the recommendations in [44]. To compute the proportion of variance explained by the models (pseudo-R^2) we used the approach proposed in [45]. All models used flat priors with a zero mean for all model parameters.

A.4 Bayesian network

Our analysis followed a previous study on network effects on peer review in four journals [19]. Building a Bayesian network was pivotal to model complex interactions between variables and potential indirect paths of bias [21]. We selected this method over alternative machine learning techniques (e.g., neural networks) as it allowed to generate a directed acyclic graph that was more appropriate to examine the structure of relations characterizing the editorial process. Furthermore, this graph permitted us to calculate the probability of an event (e.g., a rejection) depending on the value of other variables of interest (e.g., all authors being male).

The Bayesian network was estimated using the *bnlearn* package. We first trained the network on a random sample of 80% of all available manuscripts, while the other 20% were used as independent test data for model validation. Note that all nodes corresponded to the variables used in the statistical models presented in the main text. The structure of the Bayesian network and the direction of influence were learned through various constraint- and score-based structure learning algorithms. All algorithms resulted in structurally similar graphs, which were then aggregated in one network by including all links learned by at least 70% of structure learning algorithms. Figure 2 shows the resulting network. Note that we only imposed restrictions on the structure learning algorithms such that links pointing from the referee recommendation score and the editorial decision to any of the other nodes were not allowed, as were any links that were chronologically impossible.

It is worth noting here that our data were imbalanced in respect to certain variables considered in the Bayesian network. This is the case of the lower amounts of women among submission authors and the over-representation of manuscripts from physical sciences. On the one hand, this in principle implies that the learned structure of the network cannot be fully generalized to all manuscripts. However, all model diagnostics showed that these imbalances did not affect our results (SI Tab. 6). Therefore, we decided not to re-balance data manually, which would have been difficult given the amount of variables characterizing our dataset and in any case would have led to losing information.

A.5 Gender determination

The method used for gender determination was inspired by previous research [1,8,46]. We followed a standard disambiguation algorithm recently validated on a dataset of scientist names extracted from the WoS database and tested with the same time window used in our study [47].

Gender was assigned to each individual record following a multi-stage gender inference procedure consisting of three steps, in order of priority. First, we performed a preliminary gender determination using, when available, gender salutation (i.e., Mr, Mrs, Ms...). Secondly, we queried the Python package gender-guesser about the extracted first names and country of origin, if any, to corroborate gender classification. In order to maximize accuracy, we did not follow gender-guesser for names classified as mostly_male, mostly_female, andy (androgynous) or unknown (name not found). Previous research shows that gender-guesser achieves the lowest mis-classification rate and minimizes gender bias [47]. We then queried the best performer gender inference service, Gender API (https://gender-api.com/), and used the returned gender whenever we found a minimum of 62 samples with, at least, 57% accuracy. These confidence parameters for Gender API permitted us to comply with the optimal values ensuring that the rate of mis-classified names did not exceed 5% (see Benchmark 2 in [47]).

As a result, we were able to identify the gender of 82% of referees and 77% of authors (SI, Tab. 7). The remaining scientists were assigned an unknown gender, a proportion which is in line with up-to-date non-classification rates for names of scientists found in literature [47]. Note that this is a robust achievement because it implies that a human coder would hardly be able to identify these uncertain gender cases, thereby potentially introducing further bias, if involved.

Gender attribution in our three-step gender determination procedure was mostly obtained from gender-guesser (SI, Tab. 8), which is currently the best tool to assign names by origin. We assigned 57% of authors and 63% of referees their gender from this library, which also showed a fraction of mis-classification under 5% (see Table 6 in [47]). Note that the validation performed by [47] limited mis-classification to 1.5% for European names, 3.6% for African names and 6.4% for Asian names (see Table 5 in [47]). We followed Gender API to assign the gender to 13% of referees and 16% of authors. The percentage of mis-classification of this gender service was 2.1% for European names, 4.7% for African names and 11.2% for Asian names (see Table 5 in [47]). Finally, salutation was used to identify the gender to 4% authors and 6% referees.

Acknowledgements This work was supported by the TD1306 COST Action "New Frontiers of Peer Review". Access to data was possible thanks to the "PEERE Protocol for Data Sharing", co-signed by all involved partners on 1 March 2017. We would like to thank the IT office staff of all partners for support on initial data extraction. The analysis was carried out exploiting the Linnaeus University Centre for Data Intensive Sciences and Applications high-performance computing facility. This work was also partially supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) under project RTI2018-095820-B-I00. A preliminary version of the manuscript received confidential comments by Joan Marsh and Alessandra Marengoni.

Supplementary Information
Data sharing protocol
Supplementary Figures 1 to 3
Supplementary Tables 1 to 8

Competing interests BM, AB and MW declare a competing interest, being currently employed respectively as Reviewer Experience Lead at Elsevier, Executive Editor in the Computer Science team at Springer Nature, and Senior Manager, Peer Review at John Wiley and Sons. Neither of them had access to the database, elaborated any version of the dataset or were involved in data analysis.

Author contributions FS designed the study, wrote and revised the manuscript. FG coordinated data collection, wrote and revised the manuscript. PD collected and prepared data, and revised the manuscript. GB designed and performed the analysis, wrote and revised the manuscript. MF designed the analysis and wrote and revised the manuscript. AM contributed to the study design, wrote and revised the manuscript. MW and BM contributed to the study design, provided data and revised the manuscript. AB contributed to the study design and revised the manuscript.

References

[1] Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. Research: Gender bias in scholarly peer review. *eLife*, 6:e21718, 2017.

[2] Jory Lerback and Brooks Hanson. Journals invite too few women to referee. *Nature*, 541(455–457), 2017.

[3] Aileen Fyfe and Camilla Mørk Røstvik. How female fellows fared at the Royal Society. *Nature*, 555(159-161), 2018.

[4] Dawn Langan Teele and Kathleen Thelen. Gender in the journals: Publication patterns in political science. *PS: Political Science and Politics*, 50(2):433–447, 2017.

[5] Katherine Weisshaar. Publish and perish? an assessment of gender gaps in promotion to tenure in academia. *Social Forces*, 96(2):529–560, 2017.

[6] Mary Frank Fox. Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science*, 35(1):131–150, 2005.

[7] Erin Leahey. Not by productivity alone: How visibility and specialization contribute to academic earnings. *American Sociological Review*, 72(4):533–561, 2007.

[8] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. Global gender disparities in science. *Nature*, 504(211–213), 2013.

[9] Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013.

[10] Charles W. Fox, C. Sean Burns, and Jennifer A. Meyer. Editor and reviewer gender influence the peer review process but not peer review outcomes at an ecology journal. *Functional Ecology*, 30(1):140–153, 2016.

[11] Thomas König and Guido Ropers. Gender and editorial outcomes at the american political science review. *PS: Political Science and Politics*, page 1–5, 2018.

[12] Carissa L. Tudor and Deborah J. Yashar. Gender and the editorial process: World politics, 2007–2017. *PS: Political Science and Politics*, page 1–11, 2018.

[13] Edwards Hannah A., Schroeder Julia, and Dugdale Hannah L. Gender differences in authorships are not associated with publication bias in an evolutionary journal. *PLOS ONE*, 8(13):1–6, 2018.

[14] Fox Charles W. and Paine C. E. Timothy. Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecology and Evolution*, 9(6):2045–7758, 2019.

[15] Drummond Rennie. Let's make peer review scientific. *Nature*, 535(31–33), 2016.

[16] Francisco Grimaldo, Ana Marusic, and Flaminio Squazzoni. Fragments of peer review: A quantitative analysis of the literature (1969-2015). *PLOS ONE*, 13(2):1–14, 02 2018.

[17] Drummond Rennie and Annette Flanagin. Three decades of peer review congresses. *JAMA*, 319(4):350–353, 2018.

[18] Flaminio Squazzoni, Francisco Grimaldo, and Ana Marusic. Publishing: Journals could share peer-review data. *Nature*, 546(352), 2017.

[19] Giangiacomo Bravo, Mike Farjam, Francisco Grimaldo, Aliaksandr Birukou, and Flaminio Squazzoni. Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics*, 12(1):101–112, 2018.

[20] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[21] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2–3):131–163, 1997.

[22] Dakota Murray, Kyle Siler, Vincent Larivière, Wei Mun Chan, Andrew M. Collings, Jennifer Raymond, and Cassidy R. Sugimoto. Author-reviewer homophily in peer review. *bioRxiv*, 2019.

[23] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[24] S. F. Anzia and C. R. Berry. The jackie (and jill) robinson effect: Why do congresswomen outperform congressmen? *American Journal of Political Science*, 55(3):478–493, 2011.

[25] L. G. Campbell, S. Mehtani, M. E. Dozier, and J. Rinehart. Gender-heterogeneous working groups produce higher quality science. *PLOS ONE*, 8(10):1–6, 10 2013.

[26] Brink Marieke and Benschop Yvonne. Gender in academic networking: The role of gatekeepers in professorial recruitment. *Journal of Management Studies*, 51(3):460–492, 2014.

[27] M. W. Nielsen, S. Alegria, L. Börjeson, H. Etzkowitz, H. J. Falk-Krzesinski, A. Joshi, E. Leahey, L. Smith-Doerr, A. W. Woolley, and L. Schiebinger. Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences*, 114(8):1740–1742, 2017.

[28] 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):1–20, 04 2018.

[29] Jamie Lundine, Ivy Lynn Bourgeault, Jocalyn Clark, Shirin Heidari, and Dina Balabanova. The gendered system of academic publishing. *The Lancet*, 391(10132):1754–1756, 2018.

[30] Michał Krawczyk and Magdalena Smyk. Authors' gender affects rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment. *European Economic Review*, 90:326–335, 2016.

[31] S. Lundberg and J. Stearns. Women in economics: Stalled progress. *Journal of Economic Perspectives*, 33(1):3–22, February 2019.

[32] M. W. Nielsen, J. P. Andersen, L. Schiebinger, and J. W. Schneider. One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Human Nature Behavior*, 1:791–796, 2017.

[33] M. W. Nielsen, C. W. Bloch, and L. Schiebinger. Making gender diversity work for scientific discovery and innovation. *Human Nature Behavior*, 2:726–734, 2018.

[34] M. R. Berenbaum. Speaking of gender bias. *Proceedings of the National Academy of Sciences*, 116(17):8086–8088, 2019.

[35] Erin Cech, Brian Rubineau, Susan Silbey, and Caroll Seron. Professional role confidence and gendered persistence in engineering. *American Sociological Review*, 76(5):641–666, 2011.

[36] 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):1–6, 07 2013.

[37] Flaminio Squazzoni, Ana Marusic, Marco Seeber, Bahar Menhami, Michael Willis, Phil Hurst, and Aliaksandr et al Birukou. Data sharing and research on peer review: A call to action. *SocArXiv*, 2019.

[38] JP Tennant, JM Dugan, D Graziotin, DC Jacques, F Waldner, D Mietchen, Y Elkhatib, L B. Collister, CK Pikas, T Crick, P Masuzzo, A Caravaggi, DR Berg, KE Niemeyer, T Ross-Hellauer, S Mannheimer, L Rigling, DS Katz, B Greshake Tzovaras, J Pacheco-Mendoza, N Fatima, M Poblet, M Isaakidis, DE Irawan, S Renaut, CR Madan, L Matthias, J Nøargaard K, DP O'Donnell, C Neylon, S Kearns, M Selvaraju, and J Colomb. A multi-disciplinary perspective on emergent and future innovations in peer review. *F1000Research*, 6:1151, 2017.

[39] Ginger Pinholster. Journals and funders confront implicit bias in peer review. *Science*, 352(6289):1067–1068, 2016.

[40] Giangiacomo Bravo, Francisco Grimaldo, Emilia López-Iñesta, Bahar Mehmani, and Flaminio Squazzoni. The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications*, 10(1):322, January 2019.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[42] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 2017.

[43] Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[44] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[45] Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for bayesian regression models. *The American Statistician*, 73(3):307–309, May 2019.

[46] 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 *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 53–54, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

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