Investigating the Effects of Gender Bias on GitHub

Nasif Imtiaz[1], Justin Middleton[1], Joymallya Chakraborty[1], Neill Robson[1], Gina Bai[1], and Emerson Murphy-Hill[*2]

[1]Department of Computer Science, North Carolina State University
[2]Google, LLC
{simtiaz, jamiddl2, jchakra, nlrobson, rbai2}@ncsu.edu, emersonm@google.com

***Abstract*—Diversity, including gender diversity, is valued by many software development organizations, yet the field remains dominated by men. One reason for this lack of diversity is gender bias. In this paper, we study the *effects* of that bias by using an existing framework derived from the gender studies literature. We adapt the four main effects proposed in the framework by posing hypotheses about how they might manifest on GitHub, then evaluate those hypotheses quantitatively. While our results show that effects of gender bias are largely invisible on the GitHub platform itself, there are still signals of women concentrating their work in fewer places and being more restrained in communication than men.**

***Index Terms*—GitHub, gender, open source**

I. Introduction

Gender diverse teams can be more productive than homogeneous teams. For example, Hoogendoorn and colleagues found that teams with a gender balance performed better in a simulated startup than teams that were dominated by men [22]. In a software engineering context, Vasilescu and colleagues study of GitHub found that gender diversity is a significant and positive predictor of software development productivity [44].

Unfortunately, increasing gender diversity in software engineering teams is challenging because so few women participate in computer science. According to the most recent Taulbee survey by the Computing Research Association, only about 18.1% of bachelor's degrees are awarded to women [50]. In a 2013 survey over 200 open source developers, only 11.2% of the respondents were women [4]. In an annual survey on Stack Overflow, a question-answer site for programmers, only 7.6% of the 64,227 developers surveyed were women [39].

One of the reasons for women's underrepresentation in software engineering is gender bias, that is, prejudice based on gender. In Nafus's interviews with open source developers, she found that "sexist behavior is…as constant as it is extreme" [30]. In a quantitative retrospective study, Terrell and colleagues showed that when women portray themselves as women in their GitHub profiles, their acceptance rate is lower than men's, yet when their gender is not visible, their acceptance rate is higher than men's [40]. The existence of bias in the profession should be unsurprising; as Heilman finds, gender bias is generally higher in occupations and fields dominated by men [20].

To solve the problem of gender bias in software development, we must first understand it and its effects. This paper contributes the first study that quantitatively investigates these effects by examining the publicly visible behavior of men and women on the open source ecosystem GitHub. To focus our investigation, we examined the four effects of gender bias described in Williams and Dempsey's work [47]. We chose to follow Williams and Dempsey's framework because it is based on both their own interviews and a meta-review of the literature. The four effects that we examine are: *Prove-It-Again*, when women must provide more evidence than men to demonstrate competence, as women are often measured at a stricter standard; *Tightrope*, when women behave in a restrained manner to avoid backlash; *Maternal Wall*, when women who are mothers have their commitment and competence questioned; and *Tug of War*, when women are discouraging to other women. In this paper, we operationalize each of these effects quantitatively and measure how they play out in an open source software engineering context. While this paper represents a contribution to software engineering, it also represents a contribution to gender studies more broadly because GitHub presents an unparalleled opportunity to understand how people of different genders interact at a large scale online collaborative work platform.

II. Related Work

Numerous researchers have studied gender discrimination in different workplaces and gender-group differences in job performance. While a full review of all gender studies work here is prohibitive due to space limitations, we review the major meta-analyses here. Davison and colleagues' meta-analysis has shown that sex-type of a job (e.g., nurse is female sex-typed) affects employees' performance ratings and favors the sex it is typecast to [13]. It also shows that when job-relevant information about the applicants is not provided, women become subject to even greater discrimination. Roth and colleagues' meta-analysis, however, showed that women generally score higher in job performance ratings, although they still lag behind men in promotions [36]. The issue of women being rarely found at the highest level of organizations in the US and Europe was also addressed by Heilman, who explains how gender stereotypes lead to biased judgments against women and impede their career progress [20]. The

[*]This author performed the work while an Associate Professor at North Carolina State University.

present paper builds on this work by studying gender issues in the context of software development.

Other researchers have studied gender issues in software development. Vasilescu and colleagues' study of Stack Overflow, a question-answer site for programmers, found that women disengage from the platform sooner than men, even though they have similar activity levels [43]. Our paper builds on this study by investigating the effects of bias in the context of GitHub using Williams and Dempsey's existing framework.

Nafus, in her qualitative study, finds that despite having a reputation for being open and democratic, the complex construct of "openness" in open source software development leads to a male monopoly [30]. Nafus' study points out the existence of sexist behavior in open source, which may play a part in women's low participation [16]. Compared to Nafus' work, our study takes a quantitative and confirmatory approach, rather than a qualitative and formative one.

We recently studied gender differences and bias in open source [40], largely by comparing the acceptance rate of pull requests from men and women on GitHub. The authors found that while women overall get their pull requests accepted more often than men do, when the gender of the pull requestor is visible, the trend is reversed. We build directly on Terrell and colleagues' work by reusing their data and several of their analysis techniques. Our study is different in that Terrell and colleagues studied bias directly, whereas in this paper we investigate several dimensions of the effects of that bias.

III. General Methodology

Although we investigate each of the effects of bias using a different methodology, much of our methodology is common across investigations. We describe that methodology here.

A. Research Setting: GitHub

We conducted our research on GitHub, a collaborative software development platform [17]. GitHub uses a pull-based development model, where any user can contribute changes to any public project [18]. Projects on GitHub can be owned by individual developers, or by an organization. Each user has a profile page where they can optionally list personal information and where their history of project contributions are automatically updated.

B. Dataset

We began with our prior dataset [40], which builds on GHTorrent data [17] mined from public GitHub data from June 7, 2010 to April 1, 2015. Our prior dataset corrected some unreliable data from GHTorrent, and added gender data by linking GitHub users' email addresses to their public Google+ social media profiles, which contained self-reported genders for about 35% of GitHub users. The final dataset that we used contained 152,534 PRs from 20,926 women and 3,135,384 PRs from 308,062 men. Of the users who made at least one pull request, the average woman made an average of 7.3 PRs ($median = 2$), while an average man made 10.2 ($median = 2$).

For this paper, we further enriched our prior dataset by mining web pages to obtain discussion comments in pull request and issue threads. For users with an associated gender, we collected information on their profile pages, such as biography, company, and website URL. We also found that the interface and the authentication requirements of Google+ have changed since 2015, and hence we could not update gender data for new GitHub users. Therefore, the core part of our dataset remains the same that in our prior paper [40].

C. Measures and Measure Validation

Williams and Dempsey discuss the four effects of gender bias in various work contexts, but open source development is not one of them. Consequently, in this paper we create new quantitative proxy measures, adapted for the context of GitHub. Where possible, we aim to increase the construct validity of these measures in two ways: by providing manual validation by inspection and by triangulating on each bias effect by using multiple hypotheses.

IV. Prove-It-Again

The first effect of bias that we investigate is what Williams and Dempsey call Prove-it-Again, where a member of a group that doesn't align with stereotypes is measured at a *stricter* standard than those who do align with the stereotypes and consequently has to provide more evidence to demonstrate competence [6], [47]. Given a similar quality of work, work by contributors whose perceived gender is male is preferred [25], [29]. This effect is more prominent in roles that are stereotypically typecast to men [13], such as software development.

A. Hypotheses for GitHub

To triangulate the Prove-It-Again effect on GitHub, we pose and test four hypotheses.

1) Pull Request Description Length: One way that Prove-It-Again may manifest on GitHub is via pull request (PR) descriptions. On GitHub, developers submit PRs and reviewers decide if the contribution can be merged into the code base. When submitting a PR, developers can provide a description that justifies the PR [49]. Thus, we hypothesize that women's descriptions may be longer, as a method of justifying their change:

> ***Hypothesis PIA 1:*** Women provide longer pull request descriptions than men.

To validate that longer descriptions indicate more substantial justifications, we performed a manual analysis.[1] First, through a pilot analysis, we devised a binary coding scheme to determine if a sentence relates to justifying the PR. Next, we coded each sentence in 50 random non-empty English PR descriptions. As a result, we found a high Pearson correlation ($r = 0.98$, $p < .001$) between the total character count of the PR description and the character count of the sentences that justify the PR. For example, in the following PR description: the first sentence describes the change made in the PR,

[1] All coding schemes: https://figshare.com/s/de8412ceac89a3687510

the second sentence justifies how the change will improve usability, and the third sentence exemplifies a new use case that the change will handle.

> *Change the prompt function to just use the git command vs. an absolute path to /usr/bin/git. This lets the prompt work regardless of how git was installed on the machine or to where. Macports for example puts it in /opt/local/bin*

2) Pull Request Scrutiny and Push-Back: The Prove-It-Again pattern also suggests that women may face more push-back than men, where push-back means resistance to accepting a pull request, whether warranted or gratuitous. We hypothesize such push-back may take several forms:

> ***Hypothesis PIA 2:*** Women's pull requests generate more discussion, receive more change suggestions, and take more time to get accepted.

We validated these measures as follows. First, we validated that follow up comments indicate push-back. We used a pilot analysis to create a binary coding scheme to determine whether a comment relates to push-back. Next, we coded each comment in 50 random PRs that had at least one comment, where a comment could be a general one or a comment on a specific line of code. We then measured the Pearson correlation between the total number of comments and the comments relating to push-back; the correlation was low ($r = 0.32$, $p < .05$). However, we noticed that comments on specific lines of code were more likely to contain push-back. So we again sampled 50 random PRs with at least one code comment and found an almost perfect correlation ($r = .9996$, $p < .001$). Thus, henceforth we use only comments on specific lines of code as a measure of push-back.

Second, we validated that the number of follow-up PR commits indicate the amount of push-back. A follow-up PR commit is made *after* the developer has already submitted the PR for review but *before* the PR is accepted. The challenge with this measure is that follow-up commits may be made in response to reviewer push-back, but might also be made because the developer decided to proactively change the PR on their own accord. To distinguish these two, we manually inspected 50 random PRs that had at least one commit after the PR was opened. We found 30 PRs had commits that were apparently made in response to reviewer comments. In 7 PRs, we could not locate an explicit suggestion, indicating they may be proactive commits. Two PRs had changes based on automated tools' feedback. Two PRs had commits from different developers than the developer who submitted the PR. The remaining PRs had commits not related to push-back, such as merge commits. From this validation, we conclude that a reviewer likely pushed back on a PR if it had at least one commit that was made by the developer who submitted the PR and was made after at least one review comment.

3) Profile Signals: Women on GitHub might Prove-It-Again outside of their PRs, such as through their profiles. Prior research suggests that many developers consciously manage their self-image to promote their work [11] and that GitHub reviewers explore new contributors' profiles to seek information when evaluating a pull request from an unknown person [28]. Developers may also try to establish the authenticity of their profiles to build trust in a work platform like GitHub. Thus, we hypothesize:

> ***Hypothesis PIA 3:*** Women put more signals of competence on their profile than men.

We measure these signals through a profile's website URL, company name, and "bio" (that is, a short biography). However, these features can also be used for other purposes, such as humor, so we validated that the metric is largely used to highlight one's competence. To do so, we manually inspected 50 random men's and 50 random women's profiles that contained non-empty features:

- Out of 100 website URLs, we found 68 listed a personal website, 6 listed the user's social media profile, and 5 listed the user's LinkedIn profile. The remaining 21 profiles pointed to links that were broken or did not contain an apparent competence signal.
- Out of 100 organization names, 70 included real company names, 14 included real school names, and 11 included GitHub organizations. 5 profiles had company names that we could not verify the existence of or were listed as "private".
- Out of 100 bios, we categorized 71 as signals of specific technical skills or organizational membership, where 16 indicated general competencies such as "coder" or "software engineer". 13 bios did not convey any signal of competence. We found that longer bios generally include a greater number of signals about the users' competence.

We conclude that company, website URL, and bio can be largely interpreted as signals of competence to other users.

4) Pull Request Concentration: We also investigated a second-order effect of Prove-It-Again. When the Prove-It-Again effect is present, one strategy that women might take, consciously or subconsciously, is to focus their efforts on known environments so that they do not have to go through proving themselves repeatedly in new environments [47]. On GitHub, we can evaluate this second-order effect by investigating how women spread their pull requests across projects and organizations:

> ***Hypothesis PIA 4:*** Women's pull requests remain more concentrated in fewer projects and fewer organizations than men's.

B. Methodology

As Hypotheses PIA 1 and 2 deal with the pull request evaluation process, we restrict our analyses to pull requests coming from *outsiders* [40] who do not have direct commit permission to the project, because such outsiders are more likely to have to demonstrate competence than people already trusted by the project. The rationale derives from Tsay and colleagues' finding that outsiders are less likely to have their PRs accepted [42] and Terrell and colleagues' finding that gender bias affects outsiders [40]. PRs submitted by outsiders with gender information yields 46,581 pull requests for women and 1,326,253 pull requests for men.

For Hypothesis PIA 1, to determine if women provide longer descriptions to justify their PRs than men, we measure length as the character count of a PR description.

For Hypothesis PIA 2, we restrict our analysis to outsider pull requests from users whose gender is likely visible to other users, using Terrell and colleagues' technique of selecting only users who (a) have chosen their own profile image, and (b) an automated tool can confidently guess their gender based on their display name [40]. We add this restriction because a pull request reviewer must know the pull requestor's gender for the Prove-It-Again bias to materialize according to Hypothesis PIA 2. From PRs coming from gender-identifiable outsiders, we count the *review comments* (comments on specific lines of code), *review commits* (commits from the pull requester after the PR was opened and at least one review comment was posted), and the time taken to accept if the PR is eventually merged. We calculated the acceptance time as the time difference between when the pull request was first opened and its last merge state.

For Hypothesis PIA 3, we analyzed the proportion of men and women who have put information in their bio, company, and website URL field on their profile, that is — they did not leave these fields empty. We also measured the number of characters of the bio.

For Hypothesis PIA 4, we measure each user's pull request concentration, across projects and organizations. For projects, we divide the number of PRs a user has made by the number of distinct projects that these PRs are made to. The higher this ratio, the more concentrated the developer's work. For organizations, we analyzed how many pull requests from a user are submitted to a project that is owned by an organization. We then divide the number of pull requests that are submitted to any organization by the number of unique organizations submitted to. However, the spread of pull requests may not accurately represent users' activity as pull requests vary in how much code change they contain. Therefore, we also collect data on how many lines of code (LOC) are changed for each PR (subtracting the number of lines deleted from the number of lines added). We then measure each users' concentration of net LOC change across projects and organizations in the same way as before. The spread of pull request can also depend on a user's long-term presence and reputation on the platform. To control such confounding variables and validate that gender indeed is a significant predictor of concentration, we created a multiple regression model where we add seniority (number of days since the user has joined GitHub), and number of followers as variables along with gender predicting the PR concentration for each user. Thus, we test if gender indeed is a significant predictor in determining PR concentration.

To evaluate the differences, we use Mann-Whitney-Wilcoxon test for Hypotheses PIA 1, 2 and 4 and a Chi-square test for Hypothesis PIA 3. If a hypothesis is found to be supported with a significant difference between men and women, we also present the corresponding effect size (Cliff's Delta).

TABLE I
CHARACTERISTICS OF PRs FOR WOMEN AND MEN

	Women	**Men**
No Description Provided (merged)	47.4%	29.9%
No Description Provided (non-merged)	42.8%	25.0%
Median character count of provided description (merged)	103	164
Median character count of provided description (non-merged)	118	193
PRs with review comments (merged)	7.0%	7.0%
PRs with review comments (non-merged)	11.0%	8.2%
PRs with review commits (merged)	9.0%	9.7%
PRs with review commits (non-merged)	11.1%	11.2%
Median hours for PR merge	1.4	6.0

TABLE II
GITHUB PROFILE CHARACTERISTICS FOR WOMEN AND MEN

	Women	**Men**
Total Profiles	35 676	529 253
Bio Provided	7.3%	11.7%
Median Character Count of Provided Bio	37	37
URL Provided	17.9%	31.2%
Company Provided	18.7%	29.7%

C. Results

Table I shows that Hypothesis PIA 1 is not supported, that is, women do not provide longer descriptions when they submit PRs. In the first two rows, we see that women submit pull requests without any description more often than men ($p < .001$). Likewise, the next two rows show that women's pull requests often contain shorter descriptions than men ($p < .001$).

Table I also shows that Hypothesis PIA 2 is not supported. The 5th and 6th rows show that women and men have about the same proportion of PRs with review comments when the PR is eventually merged; however women have a higher proportion when the PR is not merged ($p < .001$). The 7th and 8th rows show a similar pattern, where women have fewer review commits when the PR is merged ($p = .02$); but for non-merged PRs, there is no significant difference. The pattern suggests that women go through relatively heavier push-back than men when the PRs are not eventually accepted. The hypothesis is not supported by the last row; women's pull requests are merged substantially faster than men's.

Table II shows that Hypothesis PIA 3 is also not supported; women are significantly less likely to write a bio, list their company, or list a website URL on their profile ($p < .001$ for each). When users decide to write a bio, there is no significant difference in the length of the text between men and women ($p = 0.88$).

Table III shows results for Hypothesis PIA 4, which is supported. As the table shows, women have a higher concentration of pull requests in both projects ($p < .001, d = .02$) and organizations than men ($p < .001, d = .05$). The result is further supported by a higher concentration of code changes by women in both projects ($p < .001, d = .03$) and organizations

TABLE III
PULL REQUEST CONCENTRATION FOR WOMEN AND MEN

			Women	Men
Concentration of PRs	Across Projects*	Mean	2.9	2.4
		Median	1.0	1.0
	Across Organizations*	Mean	5.8	4.8
		Median	2.0	1.5
Concentration of Net Lines of Code Changed	Across Projects*	Mean	3418	2175
		Median	26	25
	Across Organizations*	Mean	5258	2681
		Median	47	29

($p < .001, d = .07$) than men. Moreover, women contribute to fewer organizations and fewer projects than men ($p < .001$), consistent with prior work about project contributions [40]. While most women do not submit pull requests to any organization, the median man contributes 14.3% of his pull requests to organizations. Our regression model also shows gender is a significant predictor for PR concentration both in projects and organizations ($p < .001$ for each).

Hypotheses PIA 1 to 3 were not supported but Hypothesis PIA 4 was supported. We conclude that while women do not provide more evidence to demonstrate competence, they do concentrate their work across fewer projects and organizations, compared to men.

V. TIGHTROPE

The term Tightrope is most commonly associated with the circus, where a performer walks on a stretched rope. By analogy, the Tightrope effect refers to the narrow band of socially acceptable behavior for women. Behavior at the extreme ends, for example being too impolite or too polite, can have negative repercussions for women. In contrast, men tend to be forgiven for extreme behavior, sometimes even rewarded for showing extreme masculine behavior. To quote one woman interviewed by Williams and Dempsey [47],

> *In the chat rooms around Silicon Valley, from the time I arrived and until long after I left HP, I was routinely referred to as a "bimbo" or a "bitch" – too soft or too hard, and presumptuous, besides.*

Prior work demonstrates Tightrope in workplace. Burgess and colleagues showed how descriptive (how women are) and prescriptive (how women should be) components of gender stereotypes impose artificial restrictions on women's behavior [8]. Bowles and colleagues found that women are penalized when initiating a negotiation [7]. Consequently, women face an impression-management dilemma because there often is a negative reaction to expressing agency [37].

A. Hypotheses for GitHub

We examine the Tightrope effect by measuring how women express themselves in comments on GitHub. The GitHub documentation designates issue and pull requests as the appropriate place for project-related discussions among developers [2]. If the Tightrope effect is present, we hypothesize that women will show more restraint in showing emotions and affects when they participate in such discussions. Thus, the Tightrope effect would predict that women would refrain from being too polite:

> ***Hypothesis TR 1:*** Women avoid showing explicit politeness more than men.

At the other end, the Tightrope effect would also predict that women are less likely to use strong expletives to avoid impoliteness, as the social norms disapprove the use of profane words by women [34]. Thus, we hypothesize:

> ***Hypothesis TR 2:*** Women avoid profane words more than men.

The Tightrope effect would likewise predict women are less extreme in how they show their emotions through written sentiments:

> ***Hypothesis TR 3:*** Women are more neutral in showing sentiment than men on GitHub.

Prior work has shown that the ideograms are also a good indicator of sentiment. Hogenboom and colleagues found that, "whenever emoticons are used, their associated sentiment dominates the sentiment conveyed by textual cues and forms a good proxy for intended sentiment" [21]. The Tightrope effect would predict that women are both less likely to use positive or negative ideograms (such as a smiley or angry face), and more likely to use neutral ideograms (such as a neutral face):

> ***Hypothesis TR 4:*** Compared to men, women use more sentiment-neutral ideograms.

We measure a more direct demonstration of Tightrope, where Williams and Dempsey find that women are expected to behave neither too masculine nor too feminine in the workplace:

> ***Hypothesis TR 5:*** Women are more likely than men to avoid showing stereotypical masculine and stereotypical feminine traits.

B. Methodology

From the pull request comments that we had a gender for, we removed code snippets and URLs, because they can get incorrectly processed in our automated analyses. We also left out users who have fewer than 100 comments to remove floor and ceiling effects because we count neutrality rate on a per-person basis. This filtering left us with comments from 1581 women. To obtain a comparable sample of men's comments, we randomly choose 1581 men that also made at least 100 comments. Through this filtering, we get 1581 women users who have 653,031 comments in a total and an equal number of randomly selected men users who have 958,062 comments in total. This sample data is used for Hypothesis TR 1 to 4.

1) Politeness Analysis: We used a tool developed by Danescu-Niculescu-Mizil and colleagues [12] to measure politeness in our GitHub comments, which has been used in recent software engineering studies [31], [32]. This tool was found to have a fair agreement with human rating over GitHub comments [23]. Using this tool, we rated each comment as either polite or neutral as per the same approach followed in the prior work [23]. We chose not to use the impoliteness ratings from the tool as prior work has shown them to be relatively unreliable [23].

2) Profane Words Usage Analysis: For Hypothesis TR 2, we automatically analyzed comments for profane words. For a list, we used a crowdsourced and curated list of 349 profane words [1] used in prior work [26], [38]. If a comment contains at least one profane word, we rate it as a profane comment. Otherwise, we rate it as a neutral comment.

To validate that the profane words from the list actually are used to denote profanity in comments, we manually inspected 20 random samples (10 from men, 10 from women) for each of the 7 words that made up 97% of the detected profanity. For 4 words, we judged all 20 samples to be profanity. For 2 words, we found 18 out of 20 for each of them to be used as profanity. However, for the word 'ho', we found that it was a false positive for all the 20 uses; it appeared to be used in either non-English comments or as a part of the phrase "hey ho". Therefore, we decided to exclude "ho" from our list for the actual analysis. Other profanity that we did not inspect may have had false positives, but because of their low frequency of use, we judged that such false positives would have limited impact.

3) Sentiment Analysis: To evaluate Hypothesis TR 3, we applied automated sentiment analysis, which focuses on classifying the sentiment of text as to their polarity (positive, negative, or neutral) [33]. We use a tool named Senti4SD [9], which has the most accurate sentiment analysis over GitHub comments, compared to similar tools [23].

4) Ideograms Usage Analysis: For Hypothesis TR 4, we evaluated the use of ideograms within GitHub comments, where users can use emojis and emoticons. For emojis, GitHub offers a continuously expanding list of emojis, but we chose to study the 194 "people" GitHub emojis as defined by de Souza [14]. For emoticons, we use Wang and colleagues' list of the 34 most frequently used emoticons [45]. Next, we manually rated these 228 ideograms based on sentiment. For this rating, we used four human coders, and took a majority vote approach to determine the final rating. Finally, we used our ideogram ratings to obtain an overall sentiment rating for each comment. If a comment contains no ideogram, or only neutral ideograms, we rate the comment as neutral. In cases where the comments have any polar ideograms, we rate the comments by majority, or neutral in case of equal number of opposite ideograms.

5) Stereotypical Gendered Behavior Analysis: To triangulate the results from automated analyses in measuring the Tightrope effect, we perform a manual analysis to find out if women are more restrained in showing stereotypical gendered traits on GitHub. To analyze this, we took 550 comments from men, and 550 comments from women each coming from a different user. Three human coders, consisting of two men and one woman from the authors of this study, rated the tone of each comment as either stereotypically masculine (e.g. exhibiting achievement-orientation), stereotypically feminine (e.g. exhibiting concern for others), or neutral.

TABLE IV
TIGHTROPE ANALYSES FOR WOMEN AND MEN

	Women	**Men**
Median Neutrality Rate for Politeness*	66.0%	62.2%
Percentage of Users never using Profanity*	77.6%	64.5%
Median Neutrality Rate for Sentiment*	68.6%	64.3%
Median Neutrality Rate in use of Ideogram with Sentiment*	95.2%	94.4%
Neutrality Rate in showing stereotypical Gendered Behavior	71.5%	73.0%

We used Heilman's definition and traits of stereotypical masculine and feminine behaviors, as a guiding material for this manual analysis [20]. Heliman also claims these definitions of gender stereotypes to be consistent across culture, time and context. The three coders sat together for a pilot analysis over 100 comments, where they discussed the definitions presented by Heilman and their own understanding of the GitHub platform to rate each comment and come to a common comprehension. Then they rated 1100 comments individually. The coders had access to the web page of the respective PR/issue threads that the comments come from so that they can do the rating keeping the underlying context of the discussion in mind. To not bias coders' judgment by observing the GitHub users' genders or project membership status, we modified an existing open source web browser extension[2] to hide all the images and names of all the GitHub users and their membership status in any GitHub Pull Request or Issue page.

6) Neutrality Rate: For Hypotheses TR 1-4, we calculate a "neutrality rate" for each user: the percentage of comments that are neutral. We then compare men's and women's neutrality rates using a Mann-Whitney-Wilcoxon test, and measure the effect size (Cliff's Delta) if a significant difference is found. For Hypothesis TR 5, we use Chi-squared test to compare men's and women's rate of neutral comments.

C. Results

Table IV shows the results of our analysis, with an asterisk (*) indicating statistical significance. Note that our tools failed to compute a rating for 17% of the data for politeness analysis, and 10% of the data for sentiment analysis. A manual inspection suggests that most of these comments are code that our heuristics could not remove, machine-generated messages, or non-English languages. However, in this analysis we report on about 1500 users from each gender for whom we had more than 100 valid comments.

Hypothesis TR 1 is supported; women are less likely to show explicit politeness ($p < .001, d = .13$). Hypothesis TR 2 is also supported; as indicated in the second row, fewer women use profanity than men ($p < .001, d = .12$). Although there were very few instances of profane comments, the most used profanity across genders was *damn* (accounting for 18% of profanity used), *hell* (17%), *shit* (14%), *fuck* (8%), and *ass*

[2] https://github.com/nasifimtiazohi/blind-reviews

(6%). Hypothesis TR 3 is also supported, as women are less likely to show positive or negative sentiment ($p < .001, d = .15$). We also found Hypothesis TR 4 to be supported ($p < .001, d = .06$), which confirms the Hypothesis TR 3 result we get from Senti4SD as another measure for sentiment.

Hypothesis TR 5 is not supported. The human coders rated 488 comments from women, and 522 comments from men validly. The rest of the comments were rated invalid if they were non-English, had a broken GitHub page, or the commenter had mentioned their identity within the comments. We found 349 neutral, 86 feminine and 53 masculine comments from women; and 381 neutral, 69 feminine and 72 masculine comments from men. We found no statistically significant difference in neutrality rate between men and women.

We find 4 out of 5 hypotheses are supported; the data suggests the Tightrope effect is largely evident, for which women tend to be more restrained than men in general.

VI. Maternal Wall

Williams and Dempsey define the Maternal Wall effect as the bias against mothers wherein their coworkers perceive a choice without compromise: either the mothers continue to work and neglect their family, making the mother less likable, or the mother prioritize family over work, making them less reliable in the workplace [46]. Because of this perception, mothers' work and résumés are often held to higher scrutiny, and many working mothers are subtly pushed out of their positions in favor of non-mothers [10]. Prior work has also shown that mothers make less money than women without children and also less than fathers, whose wages tend to actually increase [10].

A. Hypotheses for GitHub

To determine whether the Maternal Wall effect occurs on GitHub, we must first determine how motherhood can be signaled. In a physical workspace, a person's motherhood may be apparent through casual conversation or a visible "baby bump," but it is less clear for virtual environments. Prior work provides no suggestions on how parenthood might be signaled on a developers' platform like GitHub. We estimate that only about .01% of GitHub comments refer to a developer's child; we performed a convenience sample of 100,000 comments on GitHub to find that about 1% of those comments contained the strings "child," "baby," or "kid," and of those, 99% of 200 manually inspected comments use the terms to refer to programmatic elements or diminutive references to other developers. However, GitHub still supports other ways of personal expression, such as a user's choice of profile picture, as is common on other social media [48]. We, therefore, inspect the profile pictures on GitHub to determine if users post pictures of or with children, which would signal their parenthood.

After identifying profiles with relevant pictures, we wanted to measure whether the mothers faced more push-back in pull request evaluation than other developers; however, given the data which we describe in Section VI-B, we did not find a sufficiently large population of mothers with numerous pull requests in to evaluate the Maternal Wall directly. We instead measure a second-order effect of Maternal Wall, wherein we hypothesize mothers are more likely to react to the knowledge of this effect by not revealing the fact that they had children, in fear that their coworkers and employers would view them and their priorities differently [47]. On GitHub, this effect might manifest itself through women being less likely to exhibit their parenthood on their profiles than men. Hence, we hypothesize,

> ***Hypothesis MW 1:*** The proportion of women who display that they are parents on GitHub is lower than the corresponding proportion of men.

B. Methodology

We evaluate our hypothesis by measuring whether women are less likely to include children in their profile pictures. However, we decided against comparing the raw proportions of those who had a picture with a child against those whom did not, since a 2013 survey by Arjona-Reina and colleagues suggests that OSS contributors of different genders are parents at different rates [4]. Instead, we decided to establish a baseline rate of how often OSS contributors demonstrate their parenthood by examining their behavior on general social networking sites which do not necessarily have the professional expectations of GitHub. We consulted the profile pictures of contributors' Google+ profiles to see whether they used a picture that demonstrated their parenthood on that social network, and we could investigate how many of the same contributors also had pictures with children on their GitHub profiles. If women had a bigger difference in the rates at which they used pictures of children between Google+ and GitHub than men had, then we interpreted that as evidence that they were suppressing that information for a more professional environment, which would support our hypothesis.

From our data of GitHub contributors with Google+ accounts, we randomly selected 1,547 men's accounts and 1,582 women's accounts which had a user-uploaded Google+ profile image; the slight discrepancy in sample sizes of accounts was a result of discovering upon investigation that some accounts having been deleted some time after the original data collection. One author then manually inspected every corresponding Google+ and GitHub profile image and marked those which they believed depicted a child, alone or among other people, that could plausibly belong to the owner of the account. The requirement of plausibility eliminated all photos of children that were well-known memes or commercial images, confirmed by a reverse image search of each photo to see whether it originated elsewhere than this account, or contained stylistic elements (e.g. the quality of the image or the fashions contained in it) that suggested that the child depicted was the account's owner many years ago. We acknowledge that not every qualifying image might truly correspond to the user's child; the child could be a niece or nephew, for example, or the account could belong to a talented ten-year-old developer.

However, the photos can still signal an impression of the users' parenthood to a viewer and therefore are justified to be in our dataset.

C. Results

In the table below, we list the number of user profiles examined and how many plausibly depicted the account owner's children in the profile picture, whether on both Google+ and GitHub or one and not the other.

		Pictures with Children		
	Profiles	Only G+	Only GH	Both
Men	1,547	35	6	5
Women	1,582	23	5	1

For example, the overall number of Google+ accounts we found with profile pictures depicting children, which is 64 accounts of 3,129, is given by the sums of the "Only G+" and "Both" columns. The accounts included in the "Only G+" column had a GitHub profile picture that did not plausibly depict any of the user's children. We also recorded the reverse case, wherein the user depicted children on GitHub but not Google+, although these cases are not included in the following statistical significance test.

In our sample of men and women who post pictures of children on Google+, 12.5% of men and 4.2% of women also use pictures of children on GitHub. Although this sample suggests women post the pictures less often, a chi-square test of the proportions of "Only G+" to "Both" with Yates' continuity correction, however, suggests that our results cannot conclude a difference in the rates at which the investigated genders signal their parenthood between social network environments ($\chi^2(df = 1, n = 64) = .441, p = .507$), perhaps due to the small sample sizes of the accounts we found with child photos. In the context of William and Dempsey's four effects of gender bias, although we could not discover any definitive effects of the Wall, we nevertheless view this as an important area to probe as manners of self-expression in virtual and collaborative environments evolve.

Hypothesis MW 1 was not supported; we did not find evidence that women avoid posting pictures of children at a statistically significantly higher rates than men.

VII. Tug of War

Williams and Dempsey's Tug of War suggests that in heavily competitive environments, women sometimes discourage other women because either they doubt the competency of other women [27] or they think encouraging others may increase the level of competition [15]. The Tug of War effect is related to the so-called "queen bee" syndrome, where some successful women believe that more junior women should be able to be successful in a harsh work environment without help, because the senior women were successful in the same environment [35]. As an example, in interviews at law firms, Batlan quotes one legal assistant, who said, "Females are harder on their female assistants" [5].

A. Hypotheses for GitHub

If the Tug of War pattern is prevalent on GitHub, we would expect that women may be especially harsh on other women. In the context of pull requests, we use a metric used in a prior work to measure bias [40] to hypothesize:

> ***Hypothesis TOW 1:*** Women are less likely to accept the pull requests created by other women, compared to pull requests created by men.

Similarly, while evaluating another woman's pull request, we would predict that women would give each other more push back on their pull requests:

> ***Hypothesis TOW 2:*** Compared to pull requests made by men, pull requests made by women generate more discussions, receives more change suggestions, and take more time to be evaluated when reviewed by a woman.

While the measure in Hypothesis TOW 1 is validated in a prior work [40], the measures in Hypothesis TOW 2 are validated in Hypothesis PIA 2.

B. Methodology

Here we analyze PRs coming from identifiable men and women, for which the PRs were reviewed by a woman. We excluded PRs which remain open since they have no reviewer. In our analysis for Hypothesis TOW 1, we compare the PR acceptance rate by female PR reviewers when men are the creator versus when women are the creator. For Hypothesis TOW 2, we follow similar methodology to Hypothesis PIA 2.

C. Results

Table V and Table VI show that Hypothesis TOW 1 and TOW2 are not supported. The PRs reviewed by women do not show a statistically significant different acceptance rate between PRs created by men versus those created by women ($\chi^2(df = 1, n = 33,346) = 0.9, p = .317$).

In a follow-up analysis for Hypothesis TOW 1, we consider only those PRs made by project insiders (owners and collaborators) as it arguably more closely matches the officemate

TABLE V
Pull Requests Reviewed by Men and Women

Reviewer	Creator	PR Acceptance Rate
Women	Women	88.4%
Women	Men	88.0%
Men	Women	84.0%
Men	Men	80.1%

TABLE VI
Push-back against PRs by Women Reviewers

	PR creator's Gender	
	Women	Men
PRs with review comments (merged)	4.0%	5.8%
PRs with review comments (non-merged)	8.1%	7.9%
PRs with review commits (merged)	6.2%	7.8%
PRs with review commits (non-merged)	15.7%	8.7%
Median hours for PR merge	0.2	0.9

relationships observed in many Tug of War studies. For this data, the hypothesis is confirmed, as women have a 1.6% lower probability of having their pull request being accepted by a woman than men do ($\chi^2(df = 1, n = 14,799) = 12, p = .001$). A possible explanation behind this result could be the aforementioned "queen bee" syndrome [35].

As for Hypothesis TOW 2, we investigated whether women who create PRs are more scrutinized than men, where women are reviewing in both cases. We find that while women have a lower proportion of PRs with review comments when the PR is merged, they have a higher proportion when the PR is not merged. We also find a similar pattern regarding review commits (commits made based on suggestions from the review). This pattern aligns with what we found in Section IV-C, possibly due to other reasons than the effects we are discussing. Similarly, women's PRs are still resolved more quickly. All the differences in Table VI are found to be significant using Mann-Whitney-Wilcoxon test. So, Hypothesis TOW 2 is not supported.

Hypothesis TOW 1 had conflicting support and Hypothesis TOW 2 was not supported. Therefore, we do not find substantial evidence for the Tug of War effect, where women would discourage other women.

VIII. Discussion

In Section IV, we found no visible signs of women providing more evidence to demonstrate their competence than men. In fact, we see their PRs getting merged substantially faster than men, even with shorter descriptions. One explanation is that women may interact with reviewers outside of GitHub more often than men, which would leave behind less public data for us to analyze. Similarly, less push-back may indicate that women are contributing more to their familiar environment, which aligns with our finding that women have higher concentrations of work across fewer projects and organizations. Hence the result we observed can be an indirect outcome of the Prove-It-Again effect which leads women to prefer work in familiar places to avoid the Prove-It-Again cost in new projects and teams. Future researchers who wish to analyze the communication pattern in open source may consider addressing the limitation of working solely with data mined from virtual platforms like GitHub and look for what other channels developers use to communicate outside the platform their project is hosted on, i.e. mailing lists, IRC, slack channels, and even face-to-face communication. Another explanation for our Prove-It-Again result could be that women, although lower in number in GitHub, are more competent than men overall [40] and therefore, their work does not face as much push-back as men. The high number of male users on GitHub could mean they include a larger variety of competence which is why we find Prove-It-Again affects men more than women on GitHub.

On the other hand, our hypotheses on Tightrope effects in Section V were largely supported. We see women taking a more measured approach in expressing their emotion, using cues of politeness and avoiding profanity more. A reader might attribute these results to "natural" expressiveness differences between genders, but prior research suggests that emotional expressiveness is strongly shaped by social expectations [19]. The effect may give rise to unfairness where one group of community members must moderate their behavior and withhold valuable contributions for the sake of social expectations. Interventions like adopting a "code of conduct" [3], may help mitigate Tightrope effects by helping communities articulate acceptable behaviors for all members. Technical interventions like Behavior Bot[3] may help nudge members into compliance with codes of conduct. For a non-technical intervention, Williams and Dempsey suggest forming a "posse" of coworkers to celebrate each others' accomplishments, since women are less to brag about their own accomplishments, a traditionally masculine behavior. Similarly, posses within open source communities might likewise mitigate Tightrope effects through rewarding software development accomplishments. Further research is necessary to examine the effectiveness of codes of conduct, behavioral bots, and posses on Tightrope effects in open source.

Our Hypothesis for Section VI about the Maternal Wall was not supported. If we interpret the conservation of the null hypothesis to indicate that the Maternal Wall does not make the same effect on GitHub, one possible explanation is that mothers may not expect that virtual work will experience bias because it can be performed anywhere, unlike work in traditional workplaces. Thus, profile pictures may not necessarily be viewed as a sufficiently salient signal of a users' motherhood status, so women are less concerned about posting them. While we were unable to evaluate our first-order hypothesis about the Maternal Wall—that pictures of children in profile pictures harm pull request acceptance rates for women but helps it for men—we hope as the GitHub community grows, future researchers will be able to evaluate it with additional data.

In Section VII, our hypotheses about the Tug of War effect were mostly not confirmed. Although Hypothesis TOW 1 is supported if we restrict our analysis to only insiders, there might be other confounding factors behind this as our data set becomes smaller as we apply many filters during this analysis (e.g. women reviewers, insider PR creators). Future researcher can explore this subject more to investigate if Tug of War effect manifest itself within big teams and organizations.

IX. Threats to Validity

One threat to the internal validity of our findings is whether the gender data is accurate. In the data that we adopted from Terrell and colleagues [40], developers could have misrepresented their genders on their Google+ accounts, or there could be bot accounts with arbitrarily assigned genders with a Google+ profile. In these cases, we would have included the pull request success under the wrong gender.

[3] https://github.com/behaviorbot

Additionally, since we gathered this data over a period of time, some of the underlying data might have changed while we were carrying out the collection. For example, about a year passed in between when genders were collected by Terrell and colleagues, and when we analyzed profile pictures to investigate the Maternal Wall bias. In that time, the number of child pictures might have changed. This threat would not affect all of our population, other than an unknown proportion that altered their profiles in that interval.

Kalliamvakou and colleagues' discussion of GitHub mining underscores the threat of data incompleteness that also threatens the internal validity of our results [24]. For example, communication on GitHub represents only a part of the overall project discussions, as developers might also discuss projects on organization-specific websites or other social media. We don't have insight into these external conversations and so we omit their effects on pull request evaluation. Additionally, some pull requests are merged off-site, and their status on GitHub are not reflective of their actual integration into the project. However, for much of our analyses, we considered only project outsiders, which we argue increases the likelihood that the interactions on Github are more reflective of the reality of the projects for our subjects and that the repositories under examination are not mere toy projects for which pull requests are not considered seriously.

Furthermore, inaccuracies in the tools we used might yield inaccurate results. For example, the sentiment and politeness analysis tools we used are fairly reliable. Imtiaz and colleagues evaluated the tools we used on a test data set of GitHub comments and measured an agreement rate (Weighted Kohen's Kappa) of .33 for Senti4SD and .39 for the politeness tool with the human rating [23]. Although these agreement rates are deemed as "fair", their results cannot be solely depended on to come to a final conclusion. Hence we triangulate the tools' results with three other analyses. We are also more confident in our sentiment analysis as Hypothesis TR 3 and Hypothesis TR 4 individually gives the same result on sentiment through different methods. Furthermore, we ran another cross-check with a commonly used tool named SentiStrength [41]. SentiStrength also confirms our findings, as it gives women (55.0%) a higher neutrality rate than men (51.3%) with a significant difference ($p < .001$). While we did not have an alternate politeness tool to run a cross-check of our results, Hypothesis TR 2 can be counted as a proxy to measure impoliteness and thus complements the results of Hypothesis TR 1. Elsewhere, we found that the GHTorrent data was not always accurate: for example, some pull requests were allegedly closed before they were opened according to GHTorrent. While this error case is easily rectifiable with a common-sense approach, we might have missed other, more subtle errors in the data.

In some cases, our results are also threatened by construct validity. For example, in our examination of the Maternal Wall, we assume that if women are not posting pictures of their children, then it supports our hypothesis that it is because they know of the implicit penalty. Other hypotheses may also explain that behavior. As another example, where possible we validate our measures through random sampling and inspection, but these samples may not represent the population. A third example is our push-back measure in Hypothesis TOW 2; we aim to measure discouraging behavior, but alternatively push-back might also be measuring mentoring behavior. More generally, the strength of each hypothesis' construct validity and its evidence vary. For instance, Tightrope is relatively strong because five hypotheses are triangulated and strongly tied to the underlying theory; at the other end of the spectrum, Maternal Wall investigates a second-order and untriangulated effect. Our research philosophy is that Williams and Dempseys framework should be investigated and reported wholistically.

External validity is another threat. That is, we have taken observations of behavior in one environment – a physical work environment – and have studied how trends from there generalize to GitHub. However, we cannot claim to speak for every open source community on the Internet. For one, we can only draw conclusions for a subset of the GitHub population who have linked their social media and exposed their genders. The conclusions we draw for these users may not apply to developers who do not express their genders in this way. Furthermore, our examinations exist around the particular mechanics of GitHub's pull-based interface; other forms of collaborative websites and non-software environments may experience, or not, different challenges in interaction.

X. Conclusion

We identified and investigated cues on GitHub that are comparable to those in physical workplaces and may reflect the presence of gender bias effects. In some cases we found evidence of our hypothesized effects, whereas in other cases we found our hypotheses are not supported. We found that women did not provide more information on competence and were not generally measured at a stricter standard than men. However, we find a second-order effect, that women focused their work on fewer projects and organizations. We observed that women were less likely to express politeness and profanity than men, and were more restrictive in expressing their sentiments on the platform. While data was sparse, we found that women were no more likely to conceal their parenthood than men. We likewise did not observe a clear Tug of War effect, as we do not find women to be harsher while judging pull requests submitted by other women.

Women's underrepresentation in software development is concerning. In the light of recent studies that show gender bias in open source, it is essential to look for the possible effects of such bias and how we might address it. Our study sheds light on the effects of that bias.

XI. Acknowledgements

We thank Margaret Burnett for the suggestion to use Williams and Dempsey's framing. We also thank the anonymous reviewers for their feedback. This material is based in part upon work supported by the National Science Foundation under grant numbers 1252995 and 1451172.

References

[1] List of swear words, bad words, & curse words. https://www.noswearing.com/dictionary.

[2] Mastering issues. https://guides.github.com/features/issues. Accessed on Feb. 2, 2019.

[3] Open source guides - your code of conduct.

[4] Laura Arjona-Reina, Gregorio Robles, and S Dueñas. The floss2013 free/libre/open source survey, 2014.

[5] Felice Batlan. "if you become his second wife, you are a fool": Shifting paradigms of the roles, perceptions, and working conditions of legal secretaries in large law firms. In *Special Issue Law Firms, Legal Culture, and Legal Practice*, pages 169–210. Emerald Group Publishing Limited, 2010.

[6] Monica Biernat and Diane Kobrynowicz. Gender-and race-based standards of competence: lower minimum standards but higher ability standards for devalued groups. *Journal of personality and social psychology*, 72(3):544, 1997.

[7] Hannah Riley Bowles, Linda Babcock, and Lei Lai. Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and human decision Processes*, 103(1):84–103, 2007.

[8] Diana Burgess and Eugene Borgida. Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, public policy, and law*, 5(3):665, 1999.

[9] Fabio Calefato, Filippo Lanubile, Federico Maiorano, and Nicole Novielli. Sentiment polarity detection for software development. *Empirical Software Engineering*, pages 1–31, 2017.

[10] Shelley J Correll, Stephen Benard, and In Paik. Getting a job: Is there a motherhood penalty? *American journal of sociology*, 112(5):1297–1338, 2007.

[11] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1277–1286. ACM, 2012.

[12] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*, 2013.

[13] Heather K Davison and Michael J Burke. Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, 56(2):225–248, 2000.

[14] Rafael Xavier de Souza. Complete list of github markdown emoji markup. https://gist.github.com/rxaviers/7360908. Accessed on Feb. 2, 2019.

[15] Krystle A. Hearns Elizabeth J. Parks-Stamm, Madeline E. Heilman. Motivated to penalize: women's strategic rejection of successful women. *Personality and Social Psychology Bulletin*, 34:237–247, 2008.

[16] Rishab A Ghosh, Ruediger Glott, Bernhard Krieger, and Gregorio Robles. Free/libre and open source software: Survey and study, 2002.

[17] Georgios Gousios. The ghtorent dataset and tool suite. In *Proceedings of the 10th working conference on mining software repositories*, pages 233–236. IEEE Press, 2013.

[18] Georgios Gousios, Martin Pinzger, and Arie van Deursen. An exploratory study of the pull-based software development model. In *Proceedings of the 36th International Conference on Software Engineering*, pages 345–355. ACM, 2014.

[19] Michele Grossman and Wendy Wood. Sex differences in intensity of emotional experience: a social role interpretation. *Journal of personality and social psychology*, 65(5):1010, 1993.

[20] Madeline E Heilman. Gender stereotypes and workplace bias. *Research in organizational Behavior*, 32:113–135, 2012.

[21] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th annual ACM symposium on applied computing*, pages 703–710. ACM, 2013.

[22] Sander Hoogendoorn, Hessel Oosterbeek, and Mirjam Van Praag. The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science*, 59(7):1514–1528, 2013.

[23] Nasif Imtiaz, Justin Middleton, Peter Girouard, and Emerson Murphy-Hill. Sentiment and politeness analysis tools on developer discussions are unreliable, but so are people. In *Proceedings of the Third International Workshop on Emotion Awareness in Software Engineering*, 2018.

[24] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M German, and Daniela Damian. The promises and perils of mining github. In *Proceedings of the 11th working conference on mining software repositories*, pages 92–101. ACM, 2014.

[25] Silvia Knobloch-Westerwick, Carroll J Glynn, and Michael Huge. The matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, 35(5):603–625, 2013.

[26] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*, pages 195–204. ACM, 2013.

[27] Ziva Kunda Lisa Sinclair. Motivated stereotyping of women: she's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, 26:1329–1342, 2000.

[28] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 117–128. ACM, 2013.

[29] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science facultys subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012.

[30] Dawn Nafus. 'patches don't have gender': What is not open in open source software. *New Media & Society*, 14(4):669–683, 2012.

[31] Marco Ortu, Bram Adams, Giuseppe Destefanis, Parastou Tourani, Michele Marchesi, and Roberto Tonelli. Are bullies more productive? empirical study of affectiveness vs. issue fixing time. In *Mining Software Repositories (MSR), 2015 IEEE/ACM 12th Working Conference on*, pages 303–313. IEEE, 2015.

[32] Marco Ortu, Giuseppe Destefanis, Mohamad Kassab, Steve Counsell, Michele Marchesi, and Roberto Tonelli. Would you mind fixing this issue? In *International Conference on Agile Software Development*, pages 129–140. Springer, 2015.

[33] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[34] Jeffrey Lee Rasmussen and Barbara E Moely. Impression formation as a function of the sex role appropriateness of linguistic behavior. *Sex Roles*, 14(3-4):149–161, 1986.

[35] Jennifer Rindfleish. Senior management women in australia: Diverse perspectives. *Women in Management Review*, 2000.

[36] Philip L Roth, Kristen L Purvis, and Philip Bobko. A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management*, 38(2):719–739, 2012.

[37] Laurie A Rudman and Julie E Phelan. Backlash effects for disconfirming gender stereotypes in organizations. *Research in organizational behavior*, 28:61–79, 2008.

[38] Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume 12, page 06, 2012.

[39] StackOverflow. Developer survey results. Internet. https://insights.stackoverflow.com/survey/2017.

[40] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111, 2017.

[41] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, 61(12):2544–2558, 2010.

[42] Jason Tsay, Laura Dabbish, and James Herbsleb. Influence of social and technical factors for evaluating contribution in github. In *Proceedings of the 36th international conference on Software engineering*, pages 356–366. ACM, 2014.

[43] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study. *Interacting with Computers*, 26(5):488–511, 2013.

[44] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and tenure diversity in github teams. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3789–3798. ACM, 2015.

[45] Hao Wang and Jorge A Castanon. Sentiment expression via emoticons on social media. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2404–2408. IEEE, 2015.

[46] Joan C Williams and Stephanie Bornstein. Evolution of fred: Family responsibilities discrimination and developments in the law of stereotyping and implicit bias. *Hastings LJ*, 59:1311, 2007.

[47] Joan C Williams and Rachel Dempsey. What works for women at work, 2014.

[48] Yenchun Wu, Wei-Hung Chang, and Chih-Hung Yuan. Do facebook profile pictures reflect user's personality? 51:880–889, 01 2015.

[49] Yue Yu, Huaimin Wang, Vladimir Filkov, Premkumar Devanbu, and Bogdan Vasilescu. Wait for it: Determinants of pull request evaluation latency on github. In *Mining software repositories (MSR), 2015 IEEE/ACM 12th working conference on*, pages 367–371. IEEE, 2015.

[50] Stuart Zweben and Betsy Bizot. 2017 taulbee survey. *Computing*, 29(5), 2017.