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

Presenter: Fábio Ferreira

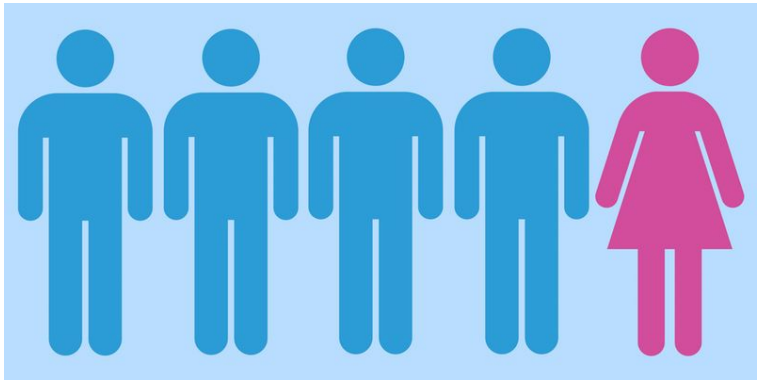
Introduction

Introduction

Gender diverse teams can be more productive than homogeneous teams.

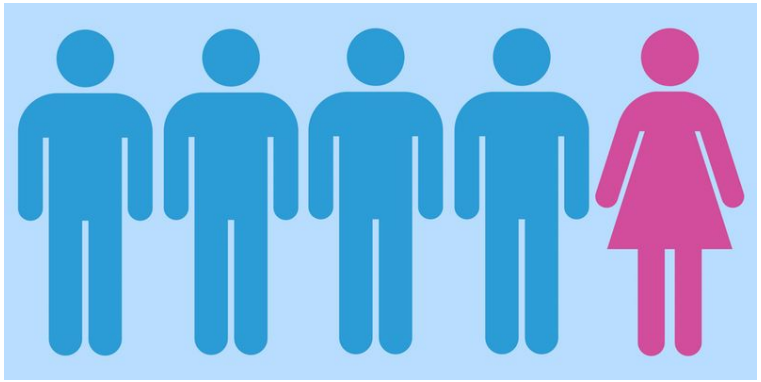
Unfortunately, increasing gender diversity in software engineering teams is challenging because so few women participate in computer science.

Introduction



Only about 18.1% of bachelor's degrees are awarded to women;

Introduction



Only about 18.1% of bachelor's degrees are awarded to women;

In an annual survey on Stack Overflow, only 7.6% of the 64,227 developers surveyed were women.


PROBLEM


PROBLEM


Gender
Bias?

The focus is to investigate **Gender Bias effects** by examining the **publicly visible behavior of men and women** on the open-source ecosystem GitHub.

Introduction

They examined the four effects of gender bias described in Williams and Dempsey's work:

Introduction

They examined the four effects of gender bias described in Williams and Dempsey's work:

- Prove-It-Again

Introduction

They examined the four effects of gender bias described in Williams and Dempsey's work:

- Prove-It-Again
- Tightrope

Introduction

They examined the four effects of gender bias described in Williams and Dempsey's work:

- Prove-It-Again
- Tightrope
- Maternal Wall

Introduction

They examined the four effects of gender bias described in Williams and Dempsey's work:

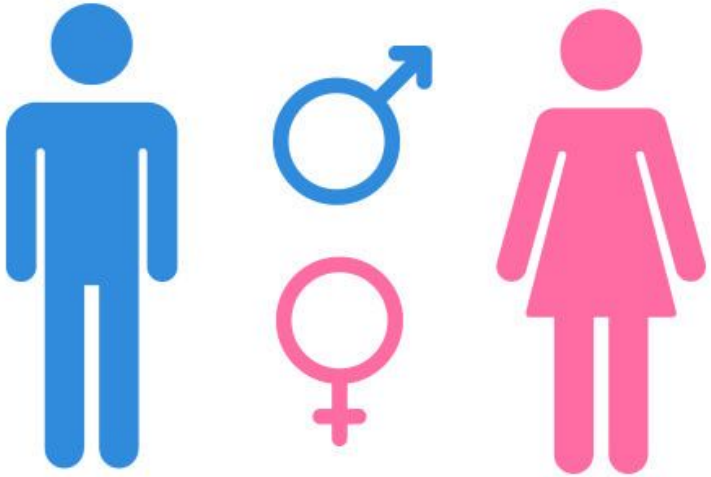
- Prove-It-Again
- Tightrope
- Maternal Wall
- Tug of War

Contributions

Contribution to software engineering;

Contributions

Contribution to software engineering;



Related Work

Related Work

Davison and colleagues' meta-analysis has shown that **sex-type of a job** affects employees' **performance ratings** [13];

Related Work

Davison and colleagues' meta-analysis has shown that **sex-type of a job** affects employees' **performance ratings** [13];

Roth and colleagues' meta-analysis showed that **women** generally **score higher in job performance ratings**, although they **still lag behind men in promotions** [36];

Related Work

Davison and colleagues' meta-analysis has shown that **sex-type of a job** affects employees' **performance ratings** [13];

Roth and colleagues' meta-analysis showed that **women** generally **score higher in job performance ratings**, although they **still lag behind men in promotions** [36];

Heilman explains how gender stereotypes lead to **biased judgments against women** and impede their **career progress** [20]

Related Work

Vasilescu and colleagues' study of Stack Overflow: women disengage from the platform sooner than men. [43]

Related Work

Vasilescu and colleagues' study of Stack Overflow: women disengage from the platform sooner than men. [43]

Nafus' study points out the existence of sexist behavior in open source [16];

Related Work

Vasilescu and colleagues' study of Stack Overflow: women disengage from the platform sooner than men. [43]

Nafus' study points out the existence of sexist behavior in open source [16];

The authors studied gender differences and bias in open source by comparing the acceptance rate of pull requests from men and women on GitHub [40].

General Methodology

General Methodology

GHTorrent

General Methodology

They mined public GitHub data from June 7, 2010 to April 1, 2015:

General Methodology

They mined public GitHub data from June 7, 2010 to April 1, 2015:

- 152,534 PRs from 20,926 women;

General Methodology

They mined public GitHub data from June 7, 2010 to April 1, 2015:

- 152,534 PRs from 20,926 women;
- 3,135,384 PRs from 308,062 men;

General Methodology

They mined public GitHub data from June 7, 2010 to April 1, 2015:

- 152,534 PRs from 20,926 women;
- 3,135,384 PRs from 308,062 men;
- Woman made an average of 7.3 PRs (median = 2);

General Methodology

They mined public GitHub data from June 7, 2010 to April 1, 2015:

- 152,534 PRs from 20,926 women;
- 3,135,384 PRs from 308,062 men;
- Woman made an average of 7.3 PRs (median = 2);
- while an average man made 10.2 (median = 2).

General Methodology

Where possible, they aim to increase the **construct validity** of the measures in two ways:

- by providing **manual validation** by inspection; and
- by **triangulating** on each bias effect by using **multiple hypotheses**.

Prove-It-Again

Prove-It-Again - Hypotheses for GitHub

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

Prove-It-Again - Hypotheses for GitHub

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

- Hypothesis PIA 1: Women provide longer PR descriptions than men;

Prove-It-Again - Hypotheses for GitHub

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

- Hypothesis PIA 1: Women provide longer PR descriptions than men;
- Hypothesis PIA 2: Women's PRs generate more discussion, receive more change suggestions, and take more time to get accepted;

Prove-It-Again - Hypotheses for GitHub

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

- Hypothesis PIA 1: Women provide longer PR descriptions than men;
- Hypothesis PIA 2: Women's PRs generate more discussion, receive more change suggestions, and take more time to get accepted;
- Hypothesis PIA 3: Women put more signals of competence on their profile than men;

Prove-It-Again - Hypotheses for GitHub

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

- Hypothesis PIA 1: Women provide longer PR descriptions than men;
- Hypothesis PIA 2: Women's PRs generate more discussion, receive more change suggestions, and take more time to get accepted;
- Hypothesis PIA 3: Women put more signals of competence on their profile than men;
- Hypothesis PIA 4: Women's PRs remain more concentrated in fewer projects and fewer organizations than men's.

Prove-It-Again - Methodology

PRs submitted by outsiders with gender information yields 46,581 pull requests for women and 1,326,253 pull requests for men:

Prove-It-Again - Methodology

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: They measure length as the character count of a PR description;

Prove-It-Again - Methodology

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: They measure length as the character count of a PR description;
- For Hypothesis PIA 2: They count the review comments, review commits, and the time taken to accept if the PR is eventually merged.

Prove-It-Again - Methodology

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 3: They analyzed the proportion of men and women who have put information in their bio, company, and website URL field on their profile;

Prove-It-Again - Methodology

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 3: They analyzed the proportion of men and women who have put information in their bio, company, and website URL field on their profile;
- For Hypothesis PIA 4: They measure each user's PR concentration, across projects and organizations.

Prove-It-Again - Results

PIA 1

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

Prove-It-Again - Results

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

PIA 1

Prove-It-Again - Results

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

PIA 2

Prove-It-Again - Results

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

PIA 2

Prove-It-Again - Results

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

PIA 2

Prove-It-Again - Results

PIA 3

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%

Prove-It-Again - Results

PIA 4

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

Prove-It-Again - Results

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.

Tightrope

Tightrope - Hypotheses for GitHub

- Hypothesis TR 1: Women avoid showing explicit politeness more than men;

Tightrope - Hypotheses for GitHub

- Hypothesis TR 1: Women avoid showing explicit politeness more than men;
- Hypothesis TR 2: Women avoid profane words more than men;

Tightrope - Hypotheses for GitHub

- **Hypothesis TR 1**: Women avoid showing explicit politeness more than men;
- **Hypothesis TR 2**: Women avoid profane words more than men;
- **Hypothesis TR 3**: Women are more neutral in showing sentiment than men on GitHub;

Tightrope - Hypotheses for GitHub

- Hypothesis TR 1: Women avoid showing explicit politeness more than men;
- Hypothesis TR 2: Women avoid profane words more than men;
- Hypothesis TR 3: Women are more neutral in showing sentiment than men on GitHub;
- Hypothesis TR 4: Compared to men, women use more sentiment-neutral ideograms;

Tightrope - Hypotheses for GitHub

- **Hypothesis TR 1**: Women avoid showing explicit politeness more than men;
- **Hypothesis TR 2**: Women avoid profane words more than men;
- **Hypothesis TR 3**: Women are more neutral in showing sentiment than men on GitHub;
- **Hypothesis TR 4**: Compared to men, women use more sentiment-neutral ideograms;
- **Hypothesis TR 5**: Women are more likely than men to avoid showing stereotypical masculine and stereotypical feminine traits.

Tightrope - Methodology

They get **1581 women users** who have **653,031 comments** and an equal number of **randomly selected men** users who have **958,062 comments**.

Tightrope - Methodology

They get **1581 women users** who have **653,031 comments** and an equal number of **randomly selected men** users who have **958,062 comments**.

1. **Politeness Analysis:** They used a tool developed by Danescu-Niculescu-Mizil and colleagues

Tightrope - Methodology

They get **1581 women users** who have **653,031 comments** and an equal number of **randomly selected men** users who have **958,062 comments**.

1. **Politeness Analysis:** They used a tool developed by Danescu-Niculescu-Mizil and colleagues

2. **Profane Words Usage Analysis::** They automatically analyzed comments for profane words;

Tightrope - Methodology

They get **1581 women users** who have **653,031 comments** and an equal number of **randomly selected men** users who have **958,062 comments**.

1. **Politeness Analysis:** They used a tool developed by Danescu-Niculescu-Mizil and colleagues

2. **Profane Words Usage Analysis::** They automatically analyzed comments for profane words;

3. **Sentiment Analysis**: Senti4SD

Tightrope - Methodology

They get **1581 women users** who have **653,031 comments** and an equal number of **randomly selected men** users who have **958,062 comments**.

1. **Politeness Analysis:** They used a tool developed by Danescu-Niculescu-Mizil and colleagues

2. **Profane Words Usage Analysis::** They automatically analyzed comments for profane words;
3. **Sentiment Analysis**: Senti4SD
4. **Ideograms Usage Analysis**: They manually rated these 228 ideograms based on sentiment.

Tightrope - Methodology

They get **1581 women users** who have **653,031 comments** and an equal number of **randomly selected men** users who have **958,062 comments**.

1. **Politeness Analysis:** They used a tool developed by Danescu-Niculescu-Mizil and colleagues

2. **Profane Words Usage Analysis::** They automatically analyzed comments for profane words;
3. **Sentiment Analysis**: Senti4SD
4. **Ideograms Usage Analysis**: They manually rated these 228 ideograms based on sentiment.
5. **Stereotypical Gendered Behavior Analysis**: 550 comments for each;

Tightrope - Results

TABLE IV
TIGHTROPE ANALYSES FOR WOMEN AND MEN

	Women	Men
TR 1 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%

Tightrope - Results

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%

TR 2

Tightrope - Results

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%
TR 3 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%

Tightrope - Results

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%

TR 4

Tightrope - Results

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%

TR 5

Maternal Wall

Maternal Wall - Hypotheses for GitHub

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

Maternal Wall - Methodology

From their data of GitHub contributors with Google+ accounts:

- They randomly selected 1,547 men's accounts and 1,582 women's accounts which had a user-uploaded Google+ profile image;
  -
- One author manually inspected every corresponding Google+ and GitHub profile image;

Maternal Wall - Results

MW 1

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

Maternal Wall - Results

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.

Tug of War

Tug of War - Hypotheses for GitHub

They would expect that women may be especially harsh on other women.

Hypothesis TOW1: Women are less likely to accept the PRs created by other women, compared to PRs created by men.

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

Tug of War - Methodology

They analyze PRs coming from identifiable men and women, for which the PRs were reviewed by a woman:

- For Hypothesis TOW 1: They compare the PR acceptance rate when men are the creator versus when women are the creator.

- For Hypothesis TOW 2: They follow similar methodology to Hypothesis PIA 2;

Tug of War - Results

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%

TOW 1

Tug of War - Results

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

TOW 1

Tug of War - Results

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.

Threats to Validity

Internal Validity

Is whether the gender data is accurate?

Some of the underlying data might have changed while we were carrying out the collection;

The number of child pictures might have changed;

Communication on GitHub represents only a part of the overall project discussions;

Come pull requests are merged off-site;

Inaccuracies in the tools we used might yield inaccurate results.

Construct Validity

If women are not posting pictures of their children, then it supports their hypothesis that it is because they know of the implicit penalty.

The samples may not represent the population..

External validity

They cannot claim to speak for every open source community on the Internet;

Their examinations exist around the particular mechanics of GitHub's pull-based interface;

Discussion and Conclusion

Discussion

*"...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."*

"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."

"Interventions like adopting a "code of conduct" [3], may help mitigate Tightrope effects..."