Perceptions of Diversity on GitHub

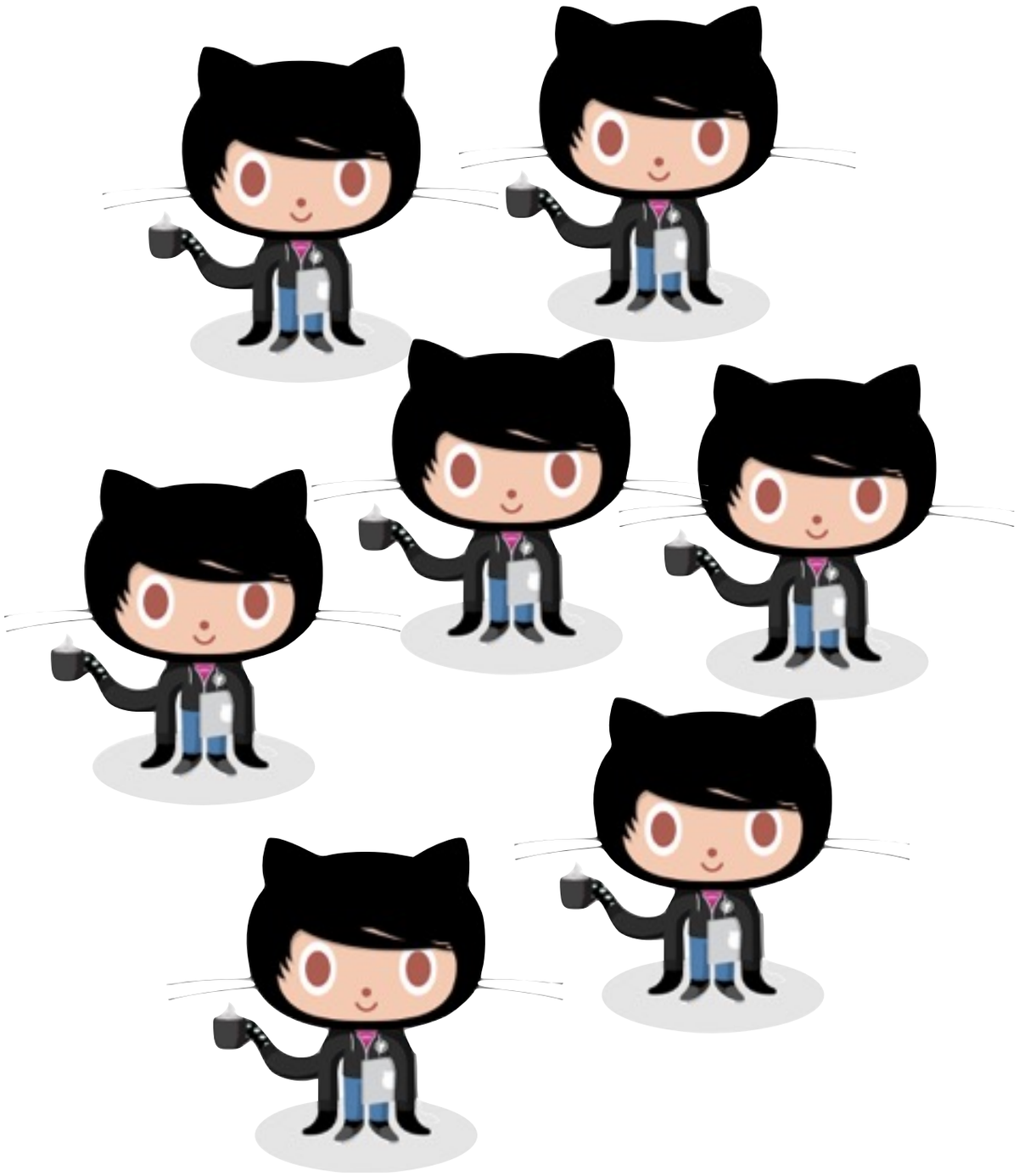
Bogdan Vasilescu, Vladimir Filkov, Alexander Serebrenik

UC DAVIS
UNIVERSITY OF CALIFORNIA

TU/e Technische Universiteit Eindhoven University of Technology

@b_vasilescu @aserebrenik @vlfilkov

Which is more effective?



Diversity 👍

Information Processing Theory
[Salancik & Pfeffer]

Diversity 👎

- Similarity attraction theory [Byrne]
- Social identity and social categorization theory [Tajfel]

CHI'15

Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu[†§*], Daryl Posnett[†], Baishakhi Ray[†], Mark G.J. van den Brand[§], Alexander Serebrenik[§], Premkumar Devanbu[†], Vladimir Filkov[†*]
[†]University of California, Davis and [§]Eindhoven University of Technology
[*]vasilescu@ucdavis.edu, filkov@cs.ucdavis.edu

ABSTRACT

Software development is usually a collaborative venture. Open Source Software (OSS) projects are no exception; indeed, by design, the OSS approach can accommodate teams that are more open, geographically distributed, and dynamic than commercial teams. This, we find, leads to OSS teams that are quite diverse. Team diversity, predominantly in offline groups, is known to correlate with team output, mostly with positive effects. How about in OSS?

Using GitHub, the largest publicly available collection of OSS projects, we studied how gender and tenure diversity relate to team productivity and turnover. Using regression modeling of GitHub data and the results of a survey, we show that both gender and tenure diversity are positive and significant predictors of productivity, together explaining a sizable fraction of the data variability. These results can inform decision making on all levels, leading to better outcomes in recruiting and performance.

Author Keywords

Open source; gender; diversity; productivity; GitHub.

ACM Classification Keywords

H.5.3. [Information Interfaces and Presentation (e.g. HCI)]: Computer-supported cooperative work

INTRODUCTION

Because of the world-wide demand for talented and skilled labor, hiring in STEM (Science, Technology, Engineering, and Math) fields has become increasingly almost entirely meritocratic, and largely blind to demographic factors. This is certainly true for software engineering; as a result, both commercial and open source software teams can be very diverse. What are the effects of this on the project as a whole? Indeed, demographic similarity enhances mutual trust (and thus, arguably, team effectiveness), while demographic diversity may lead to stereotyping, cliquishness, and conflict [20, 43]. However, a team's social diversity seems to improve its technical performance [24].

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea.
 978-1-4503-3145-6/15/04$15.00. http://dx.doi.org/10.1145/2702123.2702549

Software development teams can be *diverse* in various ways, *e.g.*, w.r.t. gender, experience, nationality, and coding language preference; some teams can be more diverse in one attribute and less so in others. Diversity attributes may also interact (*e.g.*, in some nations, female professionals may face more obstacles), which complicates analysis and study. Team diversity has been studied in physical ("meat-space") settings; however, data is hard-won in such settings. Smaller sample sizes make it difficult to effectively control for confounds. Data requirements for such effective controls, however, increase exponentially with the number of dimensions studied (one aspect of the "curse of dimensionality" [22]). Thus, studies of effects of diversity in teams (given the ineluctable confounds) require data on a great many teams, with sufficient variance along all co-variates of concern.

GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects.[1] All commits, issues, code changes, pull-requests etc. are archived and publicly available. GitHub has become the new standard for comprehensive studies of social and technical organization and achievement [16, 37, 39, 41, 60]. Evidently, this is an attractive setting in which to study the relationship of diversity to performance. The scale of GitHub is especially relevant when considering the role of women, who are very underrepresented in programming.[2] With a large enough dataset, however, the effect of increased gender diversity becomes noticeable. Additionally, since all data in GitHub is historical (*i.e.*, archived), it is possible to study the effects of tenure, or one's length of time with a project and with GitHub. However, the reliance on *volunteers* in OSS projects complicates matters; volunteers come and go, leading to team turn-over. Team turnover can certainly influence performance, and will confound the effects of diversity. The constructs of "team" and "team turnover" clearly also depend on the observation time-scale. In a healthy project, some rate of turnover is in fact desirable, as "new blood" brings in new abilities and ideas [21]. Arguably, turnover *will* affect observed diversity in GitHub OSS teams, and must be considered carefully.

In this paper, using GitHub data, we explore several questions: How diverse are online teams with respect to gender and tenure? Does gender diversity depend on tenure? On

[1]OSS depend on distributed volunteers' efforts whereas commercial software is much more centralized, and depends more on paid groups of programmers [23]; in both, the quality can be high [8].
[2]Especially so, it seems, in OSS projects: A 2013 FLOSS Survey [49] indicates 10% females; all earlier surveys [19] agree on merely 1–5%. Industry reports slightly higher numbers, *e.g.*, Google with 17% female technology employees.

Gender & tenure diversity in **GitHub** teams

Gender diversity

Tenure diversity

@b_vasilescu @aserebrenik @vlfilkov

CHI'15

Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu[†§*], Daryl Posnett[†], Baishakhi Ray[†], Mark G.J. van den Brand[§], Alexander Serebrenik[§], Premkumar Devanbu[†], Vladimir Filkov[†*]
[†]University of California, Davis and [§]Eindhoven University of Technology
[*]vasilescu@ucdavis.edu, filkov@cs.ucdavis.edu

ABSTRACT

Software development is usually a collaborative venture. Open Source Software (OSS) projects are no exception; indeed, by design, the OSS approach can accommodate teams that are more open, geographically distributed, and dynamic than commercial teams. This, we find, leads to OSS teams that are quite diverse. Team diversity, predominantly in offline groups, is known to correlate with team output, mostly with positive effects. How about in OSS?

Using GITHUB, the largest publicly available collection of OSS projects, we studied how gender and tenure diversity relate to team productivity and turnover. Using regression modeling of GITHUB data and the results of a survey, we show that both gender and tenure diversity are positive and significant predictors of productivity, together explaining a sizable fraction of the data variability. These results can inform decision making on all levels, leading to better outcomes in recruiting and performance.

Author Keywords

Open source; gender; diversity; productivity; GitHub.

ACM Classification Keywords

H.5.3. [Information Interfaces and Presentation (e.g. HCI)]: Computer-supported cooperative work

INTRODUCTION

Because of the world-wide demand for talented and skilled labor, hiring in STEM (Science, Technology, Engineering, and Math) fields has become increasingly almost entirely meritocratic, and largely blind to demographic factors. This is certainly true for software engineering; as a result, both commercial and open source software teams can be very diverse. What are the effects of this on the project as a whole? Indeed, demographic similarity enhances mutual trust (and thus, arguably, team effectiveness), while demographic diversity may lead to stereotyping, cliquishness, and conflict [20, 43]. However, a team's social diversity seems to improve its technical performance [24].

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea.
 978-1-4503-3145-6/15/04$15.00. http://dx.doi.org/10.1145/2702123.2702549

Software development teams can be *diverse* in various ways, *e.g.*, w.r.t. gender, experience, nationality, and coding language preference; some teams can be more diverse in one attribute and less so in others. Diversity attributes may also interact (*e.g.*, in some nations, female professionals may face more obstacles), which complicates analysis and study. Team diversity has been studied in physical ("meat-space") settings; however, data is hard-won in such settings. Smaller sample sizes make it difficult to effectively control for confounds. Data requirements for such effective controls, however, increase exponentially with the number of dimensions studied (one aspect of the "curse of dimensionality" [22]). Thus, studies of effects of diversity in teams (given the ineluctable confounds) require data on a great many teams, with sufficient variance along all co-variates of concern.

GITHUB, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects.[1] All commits, issues, code changes, pull-requests etc. are archived and publicly available. GITHUB has become the new standard for comprehensive studies of social and technical organization and achievement [16, 37, 39, 41, 60]. Evidently, this is an attractive setting in which to study the relationship of diversity to performance. The scale of GITHUB is especially relevant when considering the role of women, who are very underrepresented in programming.[2] With a large enough dataset, however, the effect of increased gender diversity becomes noticeable. Additionally, since all data in GITHUB is historical (*i.e.*, archived), it is possible to study the effects of tenure, or one's length of time with a project and with GITHUB. However, the reliance on *volunteers* in OSS projects complicates matters; volunteers come and go, leading to team turn-over. Team turnover can certainly influence performance, and will confound the effects of diversity. The constructs of "team" and "team turnover" clearly also depend on the observation time-scale. In a healthy project, some rate of turnover is in fact desirable, as "new blood" brings in new abilities and ideas [21]. Arguably, turnover *will* affect observed diversity in GITHUB OSS teams, and must be considered carefully.

In this paper, using GITHUB data, we explore several questions: How diverse are online teams with respect to gender and tenure? Does gender diversity depend on tenure? On

[1]OSS depend on distributed volunteers' efforts whereas commercial software is much more centralized, and depends more on paid groups of programmers [23]; in both, the quality can be high [8].
[2]Especially so, it seems, in OSS projects: A 2013 FLOSS Survey [49] indicates 10% females; all earlier surveys [19] agree on merely 1–5%. Industry reports slightly higher numbers, *e.g.*, Google with 17% female technology employees.

Gender & tenure diversity in **GitHub** teams

Productivity

Gender diversity

Turnover

Tenure diversity

@b_vasilescu @aserebrenik @vlfilkov

CHI'15

Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu[†§*], Daryl Posnett[†], Baishakhi Ray[†], Mark G.J. van den Brand[§], Alexander Serebrenik[§], Premkumar Devanbu[†], Vladimir Filkov[†*]
[†]University of California, Davis and [§]Eindhoven University of Technology
[*]vasilescu@ucdavis.edu, filkov@cs.ucdavis.edu

ABSTRACT

Software development is usually a collaborative venture. Open Source Software (OSS) projects are no exception; indeed, by design, the OSS approach can accommodate teams that are more open, geographically distributed, and dynamic than commercial teams. This, we find, leads to OSS teams that are quite diverse. Team diversity, predominantly in offline groups, is known to correlate with team output, mostly with positive effects. How about in OSS?

Using GitHub, the largest publicly available collection of OSS projects, we studied how gender and tenure diversity relate to team productivity and turnover. Using regression modeling of GitHub data and the results of a survey, we show that both gender and tenure diversity are positive and significant predictors of productivity, together explaining a sizable fraction of the data variability. These results can inform decision making on all levels, leading to better outcomes in recruiting and performance.

Author Keywords

Open source; gender; diversity; productivity; GitHub.

ACM Classification Keywords

H.5.3. [Information Interfaces and Presentation (e.g. HCI)]: Computer-supported cooperative work

INTRODUCTION

Because of the world-wide demand for talented and skilled labor, hiring in STEM (Science, Technology, Engineering, and Math) fields has become increasingly almost entirely meritocratic, and largely blind to demographic factors. This is certainly true for software engineering; as a result, both commercial and open source software teams can be very diverse. What are the effects of this on the project as a whole? Indeed, demographic similarity enhances mutual trust (and thus, arguably, team effectiveness), while demographic diversity may lead to stereotyping, cliquishness, and conflict [20, 43]. However, a team's social diversity seems to improve its technical performance [24].

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea.
 978-1-4503-3145-6/15/04$15.00. http://dx.doi.org/10.1145/2702123.2702549

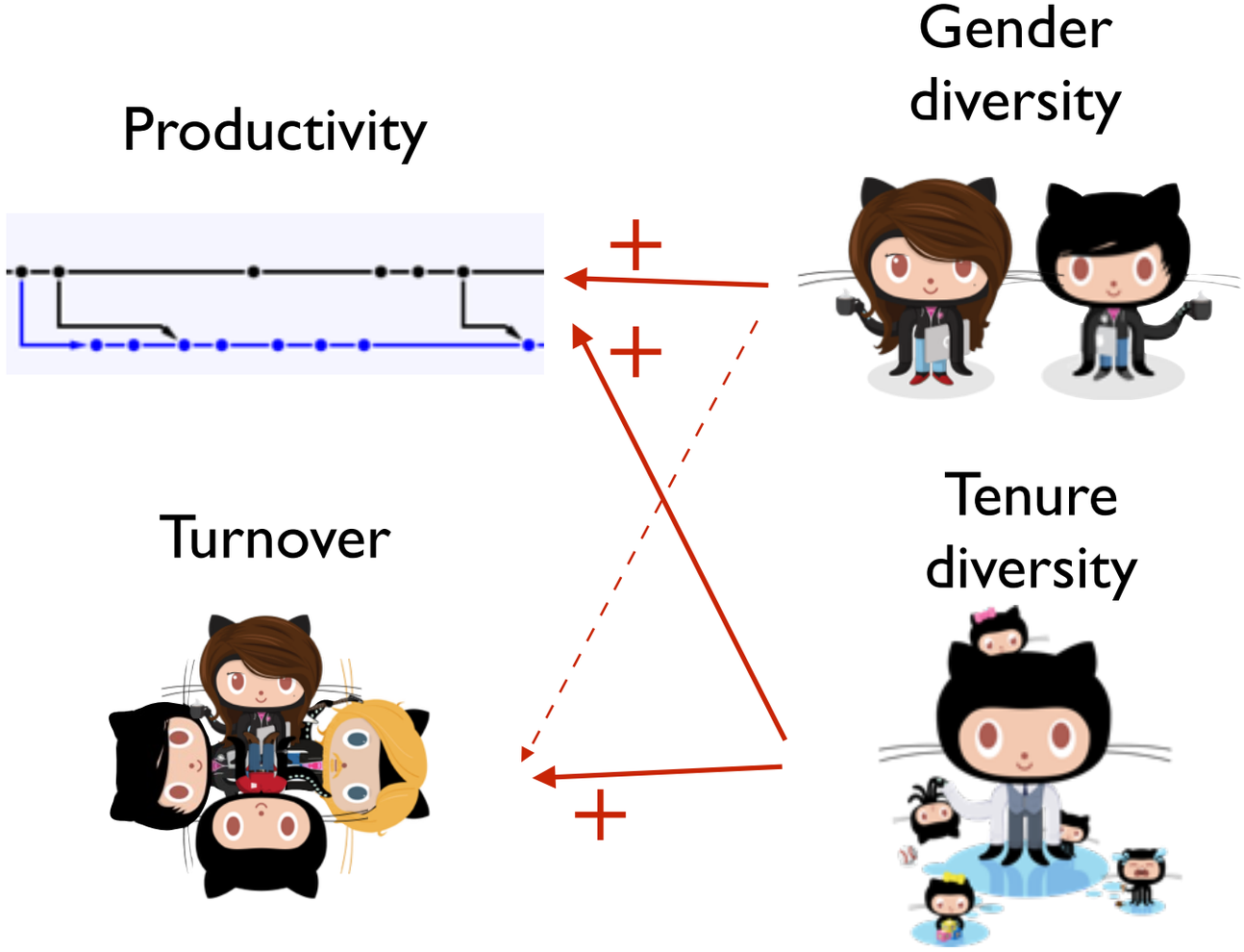
Software development teams can be *diverse* in various ways, *e.g.*, w.r.t. gender, experience, nationality, and coding language preference; some teams can be more diverse in one attribute and less so in others. Diversity attributes may also interact (*e.g.*, in some nations, female professionals may face more obstacles), which complicates analysis and study. Team diversity has been studied in physical ("meat-space") settings; however, data is hard-won in such settings. Smaller sample sizes make it difficult to effectively control for confounds. Data requirements for such effective controls, however, increase exponentially with the number of dimensions studied (one aspect of the "curse of dimensionality" [22]). Thus, studies of effects of diversity in teams (given the ineluctable confounds) require data on a great many teams, with sufficient variance along all co-variates of concern.

GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects.[1] All commits, issues, code changes, pull-requests etc. are archived and publicly available. GitHub has become the new standard for comprehensive studies of social and technical organization and achievement [16, 37, 39, 41, 60]. Evidently, this is an attractive setting in which to study the relationship of diversity to performance. The scale of GitHub is especially relevant when considering the role of women, who are very underrepresented in programming.[2] With a large enough dataset, however, the effect of increased gender diversity becomes noticeable. Additionally, since all data in GitHub is historical (*i.e.*, archived), it is possible to study the effects of tenure, or one's length of time with a project and with GitHub. However, the reliance on *volunteers* in OSS projects complicates matters; volunteers come and go, leading to team turn-over. Team turnover can certainly influence performance, and will confound the effects of diversity. The constructs of "team" and "team turnover" clearly also depend on the observation time-scale. In a healthy project, some rate of turnover is in fact desirable, as "new blood" brings in new abilities and ideas [21]. Arguably, turnover *will* affect observed diversity in GitHub OSS teams, and must be considered carefully.

In this paper, using GitHub data, we explore several questions: How diverse are online teams with respect to gender and tenure? Does gender diversity depend on tenure? On

[1]OSS depend on distributed volunteers' efforts whereas commercial software is much more centralized, and depends more on paid groups of programmers [23]; in both, the quality can be high [8].
[2]Especially so, it seems, in OSS projects: A 2013 FLOSS Survey [49] indicates 10% females; all earlier surveys [19] agree on merely 1–5%. Industry reports slightly higher numbers, *e.g.*, Google with 17% female technology employees.

Gender & tenure diversity in **GitHub** teams



@b_vasilescu @aserebrenik @vlfilkov

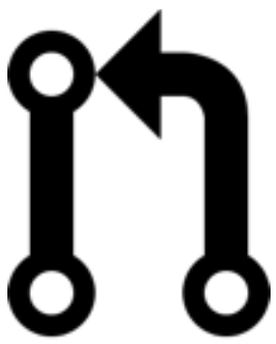
Gender & tenure diversity in **GitHub** teams

World's largest open source community

Trace data available @ghtorrent [Gousios et al]

@b_vasilescu @aserebrenik @vlfilkov

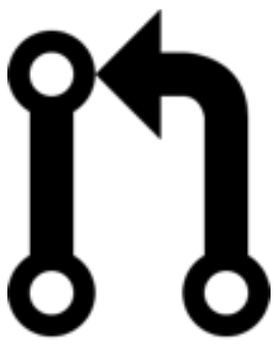
Gender & tenure diversity in **GitHub** teams



Anyone can contribute to any repository.
Who's on a team?

@b_vasilescu @aserebrenik @vlfilkov

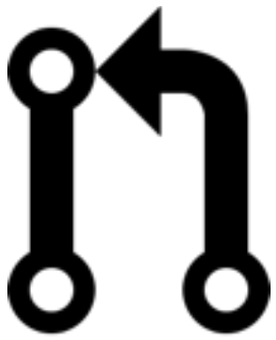
Gender & tenure diversity in **GitHub** teams



Demographics are less salient in OSS [Riordan & Shore]

@b_vasilescu @aserebrenik @vlfilkov

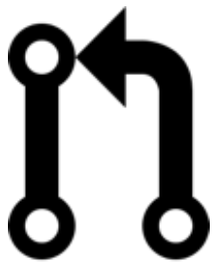
Gender & tenure diversity in **GitHub** teams



OSS as meritocracy; contribution quality as main driver of impression formation
[Dabbish et al, Marlow et al]

@b_vasilescu @aserebrenik @vlfilkov

Survey



What do people perceive constitutes a team?

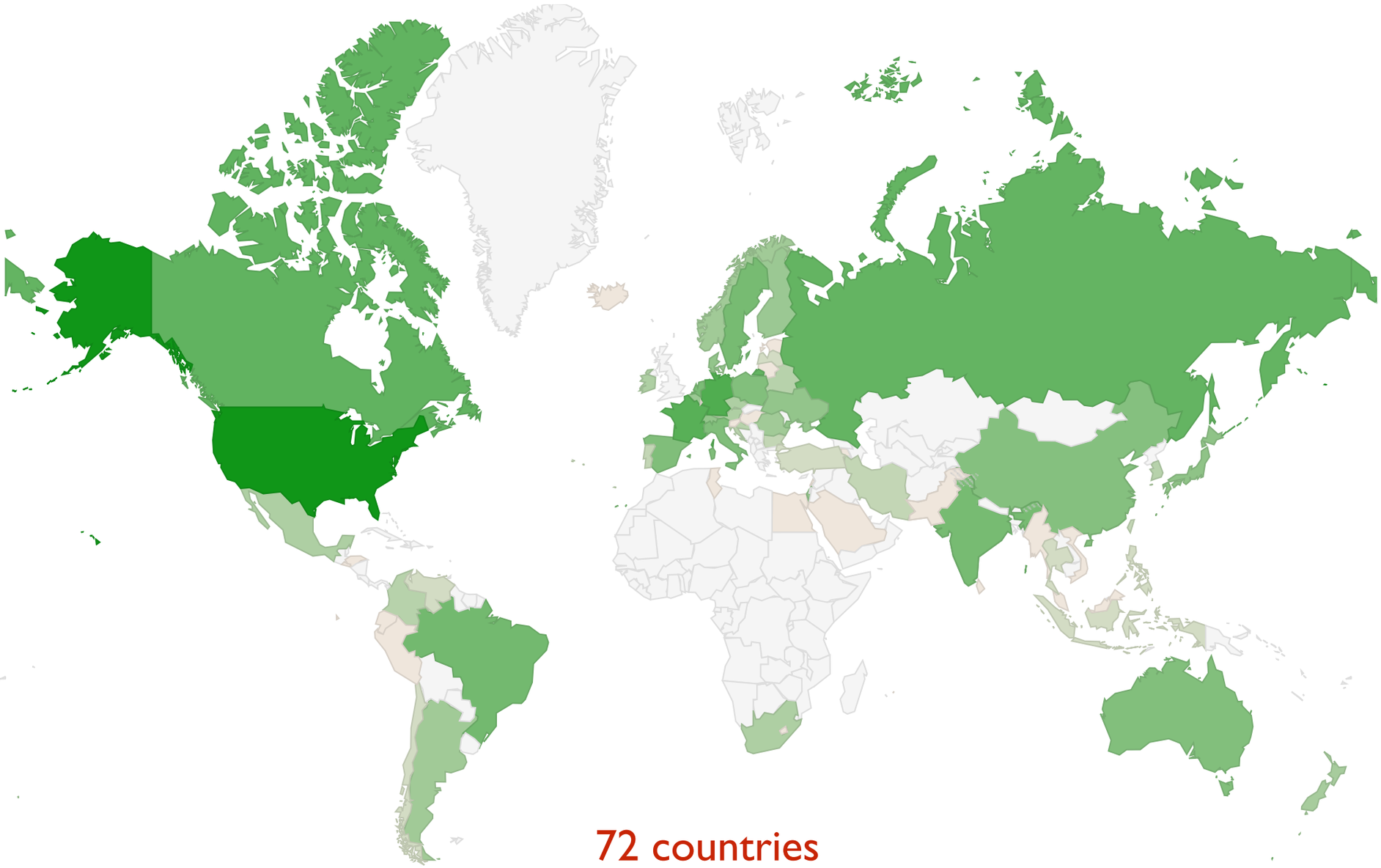
Do people recognize differences among others on their team?
Which differences are more prominent?

How is diversity perceived to influence collaboration?

Survey

4,500 invitations, 816 responses

F 24%
M 75%



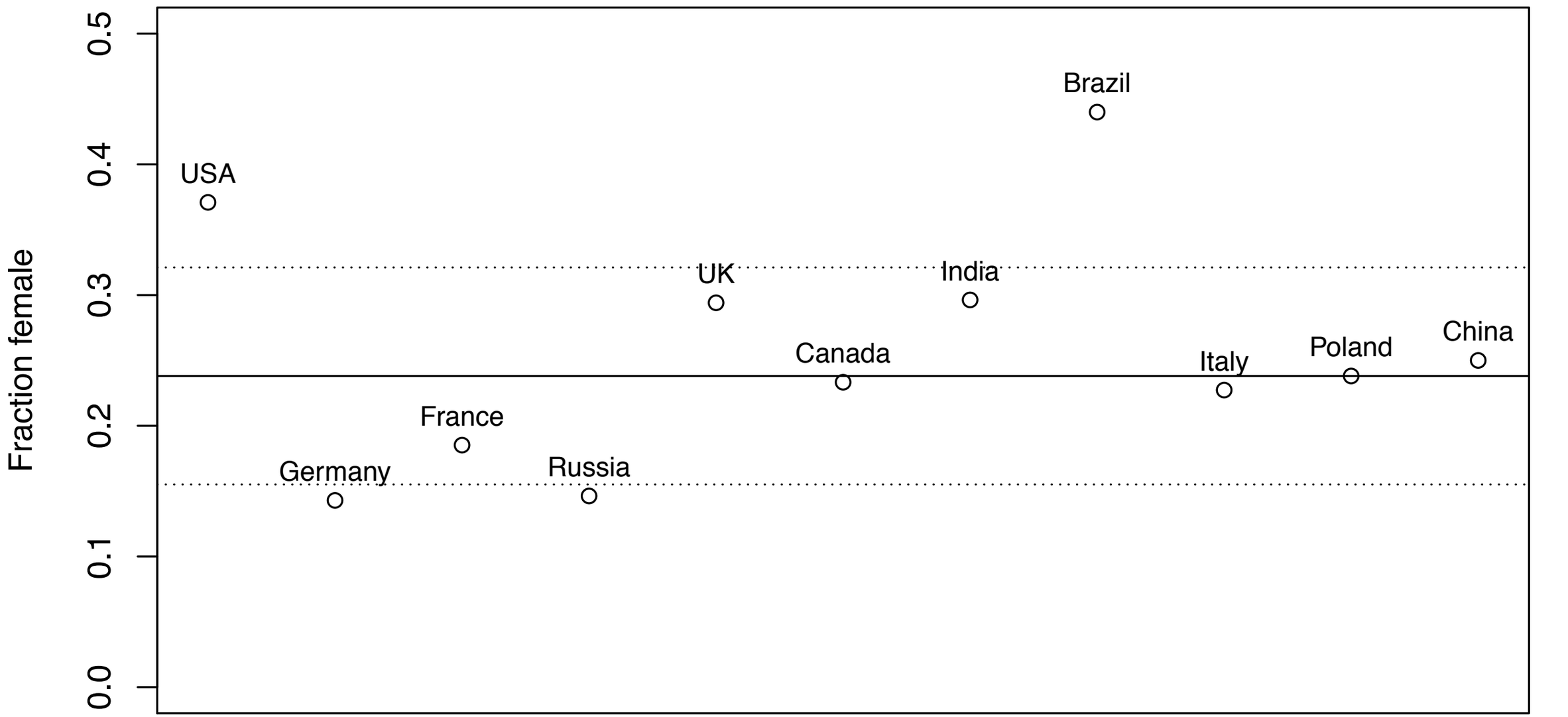
72 countries

@b_vasilescu @aserebrenik @vlfilkov

Survey

4,500 invitations, 816 responses

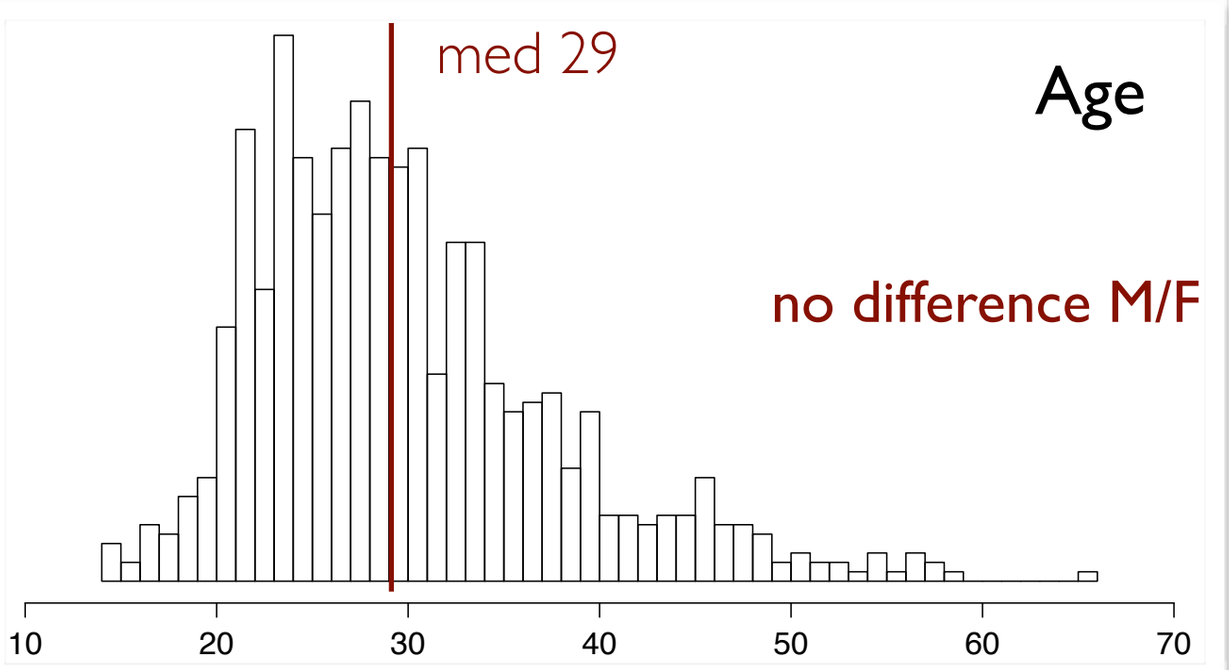
F 24%
M 75%



Survey

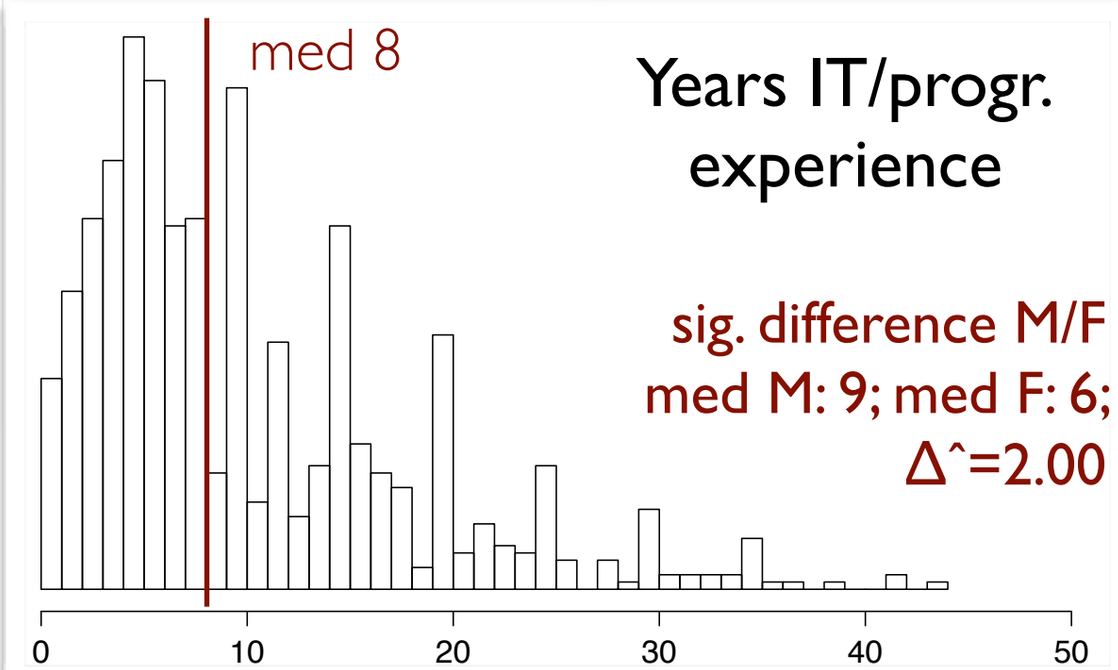
4,500 invitations, 816 responses

F 24%
M 75%



Age

no difference M/F



Years IT/progr. experience

sig. difference M/F
med M: 9; med F: 6;
Δˆ=2.00

Survey

4,500 invitations, 816 responses

F 24%
M 75%

Occupation	%
Web developer	59.70
Manager / Team leader	21.50
Student	20.64
Desktop software developer	21.25
Mobile application developer	19.16
IT staff / System administrator	15.48
Academic	13.51
Other	13.14
Database administrator	9.95
Embedded application developer	9.46
I don't work in tech	2.58

Survey

4,500 invitations, 816 responses

F 24%
M 75%

Occupation	%
Web developer	59.70
Manager / Team leader	21.50
Student	20.64
Desktop software developer	21.25
Mobile application developer	19.16
IT staff / System administrator	15.48
Academic	13.51
Other	13.14
Database administrator	9.95
Embedded application developer	9.46
I don't work in tech	2.58

Survey

4,500 invitations, 816 responses

F 24%
M 75%

Occupation	%
Web developer	59.70
Manager / Team leader	21.50
Student	20.64
Desktop software developer	21.25
Mobile application developer	19.16
IT staff / System administrator	15.48
Academic	13.51
Other	13.14
Database administrator	9.95
Embedded application developer	9.46
I don't work in tech	2.58

Whom do you consider part of your team?

- The repository owner and others who can push directly
- People who contribute code frequently
- People who work on my particular feature/ branch

less inclusive

more inclusive

- Everyone who does something in this repository

Whom do you consider part of your team?

- The repository owner and others who can push directly
- People who contribute code frequently
- People who work on my particular feature/ branch

less inclusive

more inclusive

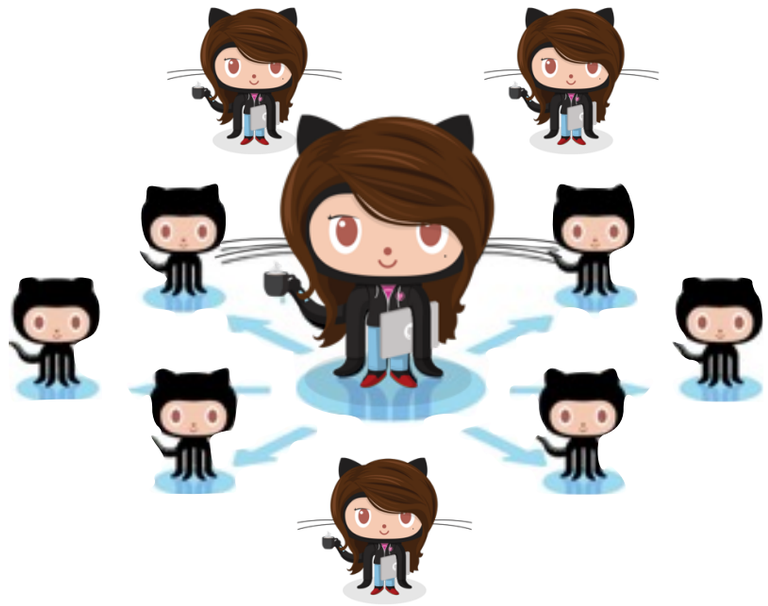
#1 (72%)

- Everyone who does something in this repository

Everyone

Which of the following characteristics of your team members are you aware of?

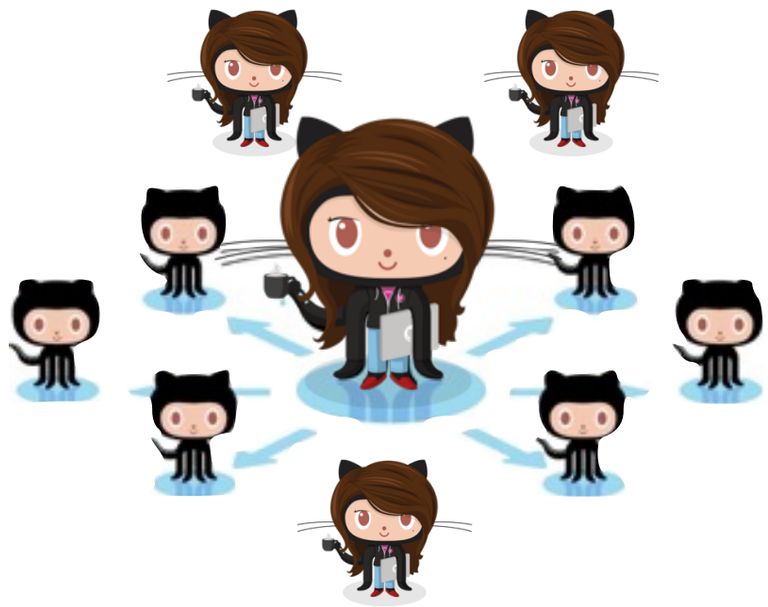
… for (none other / few other / most other) team members



- Programming skills
- Social skills
- Gender
- Ethnicity
- Overall GitHub experience
- Reputation as programmer
- Country of residence
- Personality
- Age
- Educational level
- Real name
- Hobbies
- Employment
- Political views

Which of the following characteristics of your team members are you aware of?

… for (none other / few other / **most other**) team members



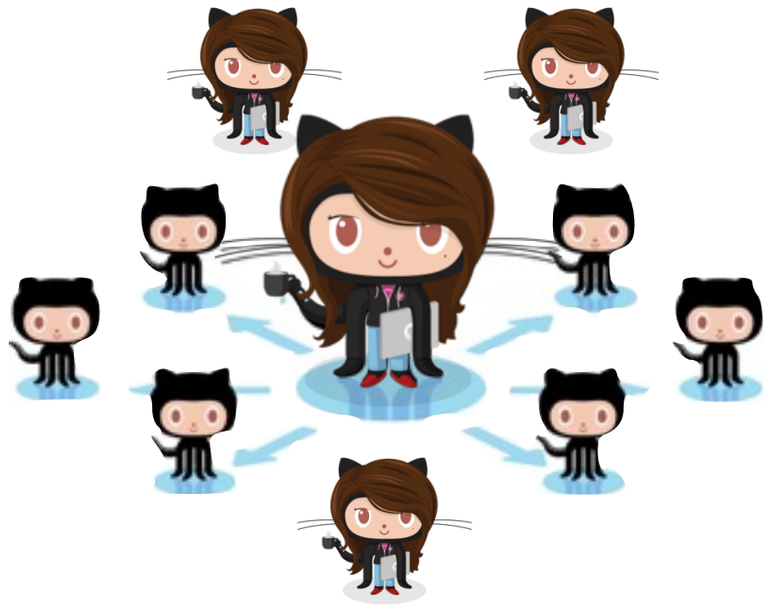
<—> Demographics not salient is OSS [Riordan & Shore]

- Programming skills 74%
- Gender 48%
- Real name 45%
- Social skills 42%
- Country of residence 40%
- Personality 39%
- Reputation as programmer 31%
- Ethnicity 30%
- Employment 30%
- GitHub experience 28%
- Educational level 26%
- Age 23%
- Hobbies 11%
- Political views 4%

Developers are aware of each other's gender

Which of the following characteristics of your team members are you aware of?

… for (none other / few other / **most other**) team members



- Programming skills 74%
- Gender 48%
- Real name 45%
- Social skills 42%
- Country of residence 40%
- Personality 39%
- Reputation as programmer 31%
- Ethnicity 30%
- Employment 30%
- GitHub experience 28%
- Educational level 26%
- Age 23%
- Hobbies 11%
- Political views 4%

Experiences working in a diverse team

"code sees no color or gender"

"any **demographic identity is irrelevant**"

"**more about the contributions to the code** than the 'characteristics' of the person"

Meritocracy; no effects of diversity

Experiences working in a diverse team

"diverse viewpoints often lead to **lively discussions and new ideas**"

"in general it is always **enriching** to communicate with someone different"

"diversity in the body of folks willing to interact and contribute works to **strengthen the usability** of the library"

Positive effects of diversity

Experiences working in a diverse team

Gender related

"I have used a fake GitHub handle (my normal GitHub handle is my first name, which is a distinctly female name) **so that people would assume I was male**"

"interactions are usually positive too, with occasional sexism, but **nothing more then one encounters in the rest of life**"

"… caused me to **leave a project**"

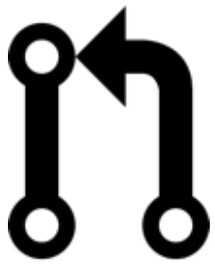
Negative effects of diversity

Perceptions of Diversity on GitHub

Bogdan Vasilescu, Vladimir Filkov, Alexander Serebrenik

UC DAVIS UNIVERSITY OF CALIFORNIA

TU/e Technische Universiteit Eindhoven University of Technology



What constitutes a team?

The team is everyone

Which differences do people recognize among team members?

Gender is surprisingly salient

Does diversity matter?

Positive/negative/no effects of diversity

@b_vasilescu @aserebrenik @vlfilkov