Perceptions of Diversity on GitHub

Bogdan Vasilescu, Vladimir Filkov, Alexander Serebrenik
Which is more effective?
Diversity 👍

Information Processing Theory
[Salancik & Pfeffer]
Diversity 👎

- Similarity attraction theory [Byrne]
- Social identity and social categorization theory [Tajfel]
Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu†, Daryl Posnett†, Raisakhil 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, and, 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 worldwide demand for talented and skilled labor, hiring in STEM (Science, Technology, Engineering, and Math) fields has become increasingly more important, 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].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Gender & tenure diversity

Gender diversity

Tenure diversity

Gender & tenure diversity in GitHub teams

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 intractable confounds) require data on a large many teams, with sufficient variance along all combinations of variety.

GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects. 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]. Evidence, 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. With a large enough dataset, however, the effect of increased gender diversity becomes noticeable. Additionally, since all data in GitHub is historical, 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 volunteered OSS projects complicates matters, volunteers come and go, leading to team-turnover. 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, it 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 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 OSS projects, we depend on the contributions of volunteers’ efforts whereas commercial software is much more centralized, and depends more on paid programmers [27]; in both, the quality can be high [5].

Especially so, it seems, in OSS projects: A 2013 FLOSS Survey [40] indicates 15% female, all our surveys [19] agree on nearly 1–5%. Industry reports slightly higher numbers, e.g., Google with 17% female technology employees.
Gender & tenure diversity in GitHub teams

Bogdan Vasilescu†, Duryl Posnett†, Raisahkhi Ray‡, Mark G.J. van den Brand§, Alexander Serebrenik, Premkumar Devanbu, Vladimir Filkov

†University of California, Davis  
‡Eindhoven University of Technology

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 relates to team productivity and turnover. Using regression modelling 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 labor-intensive, 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, disillusionment, and conflict [20, 43]. However, a team’s social diversity seems to improve its technical performance [24].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. 

Gender & tenure diversity in GitHub teams

Figure 1. Relationship of tenure and gender with productivity and turnover in GitHub teams.

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 countries, 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 intractable confoundity) require data on a great many teams, with sufficient variance along all co-variables of concern.

GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects. 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 technological organization and achievement [16, 37, 39, 43, 60].

Evidence, 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. 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 turnover. 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. Especially so, it seems, in OSS projects: A 2013 FLOSS Survey [49] indicates 10% females; all earlier surveys [19, 18, 51] have reported 10% females or less.

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 GitHub, does diversity increase volunteer’s efforts whereas commercial software is much more centralized, and depends more on paid groups of programmers [27]; in both, the quality can be high [5].

Especially so, it seems, in OSS projects: A 2013 FLOSS Survey [49] indicates 10% females; all earlier surveys [19] agree on nearly 1–5%. Industry reports slightly higher numbers, e.g., Google with 17% female technology employment.

Gender diversity

Tenure diversity

Turnover

Productivity

Gender & tenure diversity in GitHub teams

@b_vasilescu

@aserebrenik

@vlfilkov
Gender & tenure diversity in GitHub teams

Bogdan Vasilescu\textsuperscript{1,2,}\textsuperscript{†}, Daryl Poussetti\textsuperscript{3}, Raisahakhi Ray\textsuperscript{4}, Mark G.J. van den Brand\textsuperscript{1,}\textsuperscript{†}, Alexander Serebrenik\textsuperscript{5}, Premkumar Devanbu\textsuperscript{6}, Vladimir Filkov\textsuperscript{7,8,9}\textsuperscript{†}

\textsuperscript{1}University of California, Davis and \textsuperscript{2}Eindhoven University of Technology

\textsuperscript{†} vanasilceu@ucdavis.edu, filkov@cs.ucdavis.edu

Software development teams can be diverse in various ways, e.g., with 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-to-win 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") \cite{22}. Thus, studies of effects of diversity in teams (given the ineluctable conformity) require data on a great many teams, with sufficient variance along all co-variants of concern.

GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects.\textsuperscript{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 \cite{16,37,39,41,46}. 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.\textsuperscript{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 tenure” 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 \cite{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 OSS? Do gender and tenure correlate? Does OSS depend on gender diversity? How do OSS and volunteer efforts, whereas commercial software is much more centralized, and depends more on paid groups of programmers \cite{27}, in both, the quality can be high \cite{3}

\textsuperscript{1}Especially so, it seems, in OSS projects: A 2013 FLOSS Survey \cite{19} indicates 10% females; all earlier surveys \cite{21,11,10,41} with sufficient variance along all co-variants of concern.

\textsuperscript{2}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.

\textsuperscript{3}With a large enough dataset, however, the effect of increased gender diversity becomes noticeable.

\textsuperscript{4}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.

\textsuperscript{5}However, the reliance on volunteers in OSS projects complicates matters; volunteers come and go, leading to team turn-over.

\textsuperscript{6}Team turnover can certainly influence performance, and will confound the effects of diversity.

\textsuperscript{7}The constructs of “team” and “team tenure” clearly also depend on the observation time-scale.

\textsuperscript{8}In a healthy project, some rate of turnover is in fact desirable, as “new blood” brings in new abilities and ideas.

\textsuperscript{9}Arguably, turnover will affect observed diversity in GitHub OSS teams, and must be considered carefully.

\textsuperscript{1}All commits, issues, code changes, pull-requests etc. are archived and publicly available.

\textsuperscript{2}GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects.

\textsuperscript{3}All commits, issues, code changes, pull-requests etc. are archived and publicly available.

\textsuperscript{4}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.

\textsuperscript{5}However, the reliance on volunteers in OSS projects complicates matters; volunteers come and go, leading to team turn-over.

\textsuperscript{6}Team turnover can certainly influence performance, and will confound the effects of diversity.

\textsuperscript{7}The constructs of “team” and “team tenure” clearly also depend on the observation time-scale.

\textsuperscript{8}In a healthy project, some rate of turnover is in fact desirable, as “new blood” brings in new abilities and ideas.

\textsuperscript{9}Arguably, turnover will affect observed diversity in GitHub OSS teams, and must be considered carefully.

\textsuperscript{1}All commits, issues, code changes, pull-requests etc. are archived and publicly available.

\textsuperscript{2}GitHub, a social coding platform, has attracted millions of developers and thousands of Open Source Software projects. 

\textsuperscript{3}All commits, issues, code changes, pull-requests etc. are archived and publicly available.

\textsuperscript{4}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.

\textsuperscript{5}However, the reliance on volunteers in OSS projects complicates matters; volunteers come and go, leading to team turn-over.

\textsuperscript{6}Team turnover can certainly influence performance, and will confound the effects of diversity.

\textsuperscript{7}The constructs of “team” and “team tenure” clearly also depend on the observation time-scale.

\textsuperscript{8}In a healthy project, some rate of turnover is in fact desirable, as “new blood” brings in new abilities and ideas.

\textsuperscript{9}Arguably, turnover will affect observed diversity in GitHub OSS teams, and must be considered carefully.
Gender & tenure diversity in GitHub teams

World’s largest open source community

Trace data available @ghtorrent
[Gousios et al]
Gender & tenure diversity in GitHub teams

Anyone can contribute to any repository.
Who’s on a team?
Gender & tenure diversity in GitHub teams

Demographics are less salient in OSS [Riordan & Shore]
Gender & tenure diversity in GitHub teams

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

med 29

Years IT/progr. experience

sig. difference M/F
med M: 9; med F: 6;
$\Delta^\wedge = 2.00$
Survey

4,500 invitations, 816 responses

<table>
<thead>
<tr>
<th>Occupation</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web developer</td>
<td>59.70</td>
</tr>
<tr>
<td>Manager / Team leader</td>
<td>21.50</td>
</tr>
<tr>
<td>Student</td>
<td>20.64</td>
</tr>
<tr>
<td>Desktop software developer</td>
<td>21.25</td>
</tr>
<tr>
<td>Mobile application developer</td>
<td>19.16</td>
</tr>
<tr>
<td>IT staff / System administrator</td>
<td>15.48</td>
</tr>
<tr>
<td>Academic</td>
<td>13.51</td>
</tr>
<tr>
<td>Other</td>
<td>13.14</td>
</tr>
<tr>
<td>Database administrator</td>
<td>9.95</td>
</tr>
<tr>
<td>Embedded application developer</td>
<td>9.46</td>
</tr>
<tr>
<td>I don’t work in tech</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Survey
4,500 invitations, 816 responses

<table>
<thead>
<tr>
<th>Occupation</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web developer</td>
<td>59.70</td>
</tr>
<tr>
<td>Manager / Team leader</td>
<td>21.50</td>
</tr>
<tr>
<td>Student</td>
<td>20.64</td>
</tr>
<tr>
<td>Desktop software developer</td>
<td>21.25</td>
</tr>
<tr>
<td>Mobile application developer</td>
<td>19.16</td>
</tr>
<tr>
<td>IT staff / System administrator</td>
<td>15.48</td>
</tr>
<tr>
<td>Academic</td>
<td>13.51</td>
</tr>
<tr>
<td>Other</td>
<td>13.14</td>
</tr>
<tr>
<td>Database administrator</td>
<td>9.95</td>
</tr>
<tr>
<td>Embedded application developer</td>
<td>9.46</td>
</tr>
<tr>
<td>I don’t work in tech</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Survey

4,500 invitations, 816 responses

<table>
<thead>
<tr>
<th>Occupation</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web developer</td>
<td>59.70</td>
</tr>
<tr>
<td>Manager / Team leader</td>
<td>21.50</td>
</tr>
<tr>
<td>Student</td>
<td>20.64</td>
</tr>
<tr>
<td>Desktop software developer</td>
<td>21.25</td>
</tr>
<tr>
<td>Mobile application developer</td>
<td>19.16</td>
</tr>
<tr>
<td>IT staff / System administrator</td>
<td>15.48</td>
</tr>
<tr>
<td><strong>Academic</strong></td>
<td>13.51</td>
</tr>
<tr>
<td>Other</td>
<td>13.14</td>
</tr>
<tr>
<td>Database administrator</td>
<td>9.95</td>
</tr>
<tr>
<td>Embedded application developer</td>
<td>9.46</td>
</tr>
<tr>
<td>I don’t work in tech</td>
<td>2.58</td>
</tr>
</tbody>
</table>

F 24%
M 75%
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

Everyone who does something in this repository

#1 (72%) - more inclusive

less inclusive
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

- 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%

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

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

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