Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark van den Brand, Alexander Serebrenik, Prem Devanbu, Vladimir Filkov
Which is more effective?
Which is more effective?
Diversity 

Similarity attraction theory

People prefer working with others similar to them in terms of values, beliefs, and attitudes [Byrne]

Social identity and social categorization theory

People categorize themselves into specific groups. Members of own group are treated better than outsiders [Tajfel]

Due to greater perceived differences between groups than within groups, diversity can lead to confusion, stress, and conflict [Horwitz & Horwitz]
Diversity 👍

Driver of internal innovation and business growth [Forbes]

Diverse problem solvers outperform high ability problem solvers [Hong & Page]

Companies with diverse executive boards have higher earnings and returns on equity [McKinsey]

Multicultural social networks promote creativity [Harvard Business School]
Diversity 👍

Information Processing Theory

Mixture of cultural/educational backgrounds + access to different networks/broader information
=> creativity, adaptability, & problem solving skills.
[Salancik & Pfeffer]
Today: diversity in open source software (OSS) GitHub teams

Different settings
- Geographic & cultural dispersion

Different methods
- Online communities & distributed comm. channels
- Quantitative; large-scale trace data
Today: gender & tenure diversity in open source software (OSS) GitHub teams

Gender diversity = mix women/men

simplifying assumption: gender is binary

The “hacker” culture is male-dominated and unfriendly to women [Turkle]

Women are <10% in OSS [Robles et al]

Reports of active discrimination and sexism towards women [Nafus]
Today: gender & tenure diversity in open source software (OSS) GitHub teams

The “onion” structure of OSS: small (stable) core + large (loose) periphery [Ducheneaut]

Tenure diversity = mix junior/senior

High turnover [Robles & Gonzalez-Barahona]
Today: gender & tenure diversity in open source software (OSS) GitHub teams

World’s largest open source community

Trace data available @ghtorrent [Gousios et al]
Today: gender & tenure diversity in open source software (OSS) GitHub teams

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

Theoretical

Technical
Today: gender & tenure diversity in open source software (OSS) GitHub teams

Demographics are less salient in OSS [Riordan & Shore]
Today: gender & tenure diversity in open source software (OSS) GitHub teams

Theoretical

Technical

Anyone can contribute to any repository. Who’s on a team?
Today: gender & tenure diversity in open source software (OSS) GitHub teams

Gender is not explicitly recorded
Today: gender & tenure diversity in open source software (OSS) GitHub teams

People contribute under multiple aliases

Theoretical
Technical
Today: gender & tenure diversity in open source software (OSS) GitHub teams

How to analyze such large-scale longitudinal trace data?
Approach: mixed methods

Diversity survey

Welcome to our GitHub diversity survey!

This survey is aimed at developing a better understanding of the role played by gender, age, and national origin in distributed software engineering teams.

Your participation is voluntary and confidential. If you agree to

[Vasilescu et al, CHASE’15]
Survey
4,500 invitations, 816 responses

What constitutes a team?

Which differences do people recognize among team members?

Does diversity matter?

[Vasilescu et al, CHASE’15]
Survey

4,500 invitations, 816 responses

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

[Vasilescu et al, CHASE'15]
Mining

[Vasilescu et al, MSR'15]

• http://bvasiles.github.io/papers/msr_data15.pdf
• https://github.com/bvasiles/diversity

A data set for social diversity studies of GitHub teams

The data is presented in CSV format and can be directly imported in R. It contains a number of standard measures of (GitHub) activity, including number of committers, team size (committers, pull request submitters, commenters, etc.), number of commits (the most encompassing form of coding contribution to a GitHub project and a representative facet of developer productivity in open source), number of comments (on commits, pull requests, and issues; a measure of the project's social activity), number of issues opened, number of forks, and number of watchers.

Then, for each quarter (at least 4 quarters of data per project, by construction), we compute the project age (in quarters), the number of female and male contributors, the genders and countries of team members (at least 75% resolved, by construction), their GitHub tenures (in days; capturing global GitHub presence, based on account creation date), commit tenures (in days; capturing global coding experience, based on participation in any GitHub repository), project tenures (in quarters; local project experience, not restricted to coding), the numbers of contributors leaving (i.e., active in the previous quarter but inactive now), joining (defined analogously), and staying in the team (i.e., in common between w.r.t. previous quarter), as well as the turnover ratio (i.e., the fraction of the team in a given quarter that is different with respect to previous quarter).

Finally, we compute Blau indices of team gender and country diversity, a well-established diversity measure for categorical variables, and coefficients of variation for GitHub, commit, and project tenure, as measures of team tenure diversity.
**Mining**

Infer genders (93% precision) [Vasilescu et al, IWC’13]

Sample 4K projects

Bing Maps + Heuristics
http://github.com/tue-mdse/countryNameManager

Name frequency tables for 30 countries
http://github.com/tue-mdse/genderComputer

Andrea + Italy = male
Andrea + USA = female
Response

Productivity
(#commits/quarter)

Turnover
(fraction team new w.r.t. prev. quarter)

Mining

GitHub

Mining

Sample

4K projects
Mining

Response

Productivity
(#commits/quarter)

Turnover
(fraction team new w.r.t. prev. quarter)

Independent

Gender diversity
(Blau index)

Tenure diversity
(coeff. variation)
- project
- overall coding

Sample
4K projects
Minning

Response

Productivity
(#commits/quarter)

Turnover
(fraction team new w.r.t. prev. quarter)

Independent

Gender diversity
(Blau index)

Tenure diversity
(coeff. variation)
• project
• overall coding

Controls

Team size

Time

Project age

Project activity

Sample
4K projects

Mining github social coding

@b_vasilescu
@baishakhir
@MarkvandenBrand
@aserebrenik
@devanbu
@vlfilkov
### Analysis

**Nesting: projects**

<table>
<thead>
<tr>
<th>Project</th>
<th>Created on</th>
<th>Project age</th>
<th>Total commits</th>
<th>#Forks</th>
<th>Time</th>
<th>#Commits</th>
<th>#Comments</th>
<th>Team size</th>
<th>Gender diversity</th>
<th>Commit tenure diversity</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2011-02-15</td>
<td>12</td>
<td>557</td>
<td>51</td>
<td>Q2</td>
<td>47</td>
<td>26</td>
<td>9</td>
<td>0.25</td>
<td>0.47</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q5</td>
<td>19</td>
<td>12</td>
<td>10</td>
<td>0.00</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q6</td>
<td>7</td>
<td>13</td>
<td>12</td>
<td>0.25</td>
<td>0.54</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q7</td>
<td>56</td>
<td>53</td>
<td>20</td>
<td>0.00</td>
<td>0.56</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2010-09-21</td>
<td>11</td>
<td>2075</td>
<td>578</td>
<td>Q4</td>
<td>71</td>
<td>169</td>
<td>83</td>
<td>0.03</td>
<td>0.66</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q5</td>
<td>116</td>
<td>219</td>
<td>93</td>
<td>0.05</td>
<td>0.73</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q6</td>
<td>186</td>
<td>367</td>
<td>119</td>
<td>0.06</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q7</td>
<td>129</td>
<td>453</td>
<td>114</td>
<td>0.08</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Analysis

**Nesting:** projects  
**Cross-classification:** quarters

<table>
<thead>
<tr>
<th>Project</th>
<th>Created on</th>
<th>Project age</th>
<th>Total #commits</th>
<th>#Forks</th>
<th>Time</th>
<th>#Commits</th>
<th>#Comments</th>
<th>Team size</th>
<th>Gender diversity</th>
<th>Commit tenure diversity</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2011-02-15</td>
<td>12</td>
<td>557</td>
<td>51</td>
<td>Q2</td>
<td>47</td>
<td>26</td>
<td>9</td>
<td>0.25</td>
<td>0.47</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Q5</td>
<td>19</td>
<td>12</td>
<td>10</td>
<td></td>
<td>19</td>
<td>12</td>
<td>10</td>
<td>0.00</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Q6</td>
<td>7</td>
<td>13</td>
<td>12</td>
<td></td>
<td>7</td>
<td>13</td>
<td>12</td>
<td>0.25</td>
<td>0.54</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Q7</td>
<td>56</td>
<td>53</td>
<td>20</td>
<td></td>
<td>56</td>
<td>53</td>
<td>20</td>
<td>0.00</td>
<td>0.56</td>
<td>0.87</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2010-09-21</td>
<td>11</td>
<td>2075</td>
<td>578</td>
<td>Q4</td>
<td>71</td>
<td>169</td>
<td>83</td>
<td>0.03</td>
<td>0.66</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Q5</td>
<td>116</td>
<td>219</td>
<td>93</td>
<td></td>
<td>116</td>
<td>219</td>
<td>93</td>
<td>0.05</td>
<td>0.73</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Q6</td>
<td>186</td>
<td>367</td>
<td>119</td>
<td></td>
<td>186</td>
<td>367</td>
<td>119</td>
<td>0.06</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Q7</td>
<td>129</td>
<td>453</td>
<td>114</td>
<td></td>
<td>129</td>
<td>453</td>
<td>114</td>
<td>0.08</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Analysis

**Nesting:** projects  
**Cross-classification:** quarters

<table>
<thead>
<tr>
<th>Project</th>
<th>Created on</th>
<th>Project age</th>
<th>Total #commits</th>
<th>#Forks</th>
<th>Time</th>
<th>#Commits</th>
<th>#Comments</th>
<th>Team size</th>
<th>Gender diversity</th>
<th>Commit tenure diversity</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2011-02-15</td>
<td>12</td>
<td>557</td>
<td>51</td>
<td>Q2</td>
<td>47</td>
<td>26</td>
<td>9</td>
<td>0.25</td>
<td>0.47</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q5</td>
<td>19</td>
<td>12</td>
<td>10</td>
<td>0.00</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q6</td>
<td>7</td>
<td>13</td>
<td>12</td>
<td>0.25</td>
<td>0.54</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q7</td>
<td>56</td>
<td>53</td>
<td>20</td>
<td>0.00</td>
<td>0.56</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2010-09-21</td>
<td>11</td>
<td>2075</td>
<td>578</td>
<td>Q4</td>
<td>71</td>
<td>169</td>
<td>83</td>
<td>0.03</td>
<td>0.66</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q5</td>
<td>116</td>
<td>219</td>
<td>93</td>
<td>0.05</td>
<td>0.73</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q6</td>
<td>186</td>
<td>367</td>
<td>119</td>
<td>0.06</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q7</td>
<td>129</td>
<td>453</td>
<td>114</td>
<td>0.08</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Nesting:** projects  
**Cross-classification:** quarters

**Linear mixed-effects (hierarchical) models**
Results

Productivity
(#commits/quarter)

Team size
Overall project activity
Forks
Project age

Gender diversity
Tenure diversity

+  +  -  -
Results

Productivity (#commits/quarter)

Team size
Overall project activity
Forks
Project age

+ + + all team sizes
+ mid-size & large teams

Gender diversity
Tenure diversity

Results
Results

Productivity
(#commits/quarter)

Gender diversity

Tenure diversity
Results

Productivity  
(#commits/quarter)

Turnover  (fraction team new w.r.t. prev. quarter)

Gender diversity

Tenure diversity

@b_vasilescu  @baishakhir  @MarkvandenBrand  @aserebrenik  @devanbu  @vlfilkov
The takeaway

Today: diversity in open source software (OSS) GitHub teams

Different settings
- Geographic & cultural dispersion

Different methods
- Online communities & distributed comm. channels
- Quantitative; large-scale trace data
The takeaway

Which is more effective?
The takeaway

Today: diversity in open source software (OSS) GitHub teams

Different settings
- Geographic & cultural dispersion
- Online communities & distributed comm. channels

Different methods
- Quantitative; large-scale trace data

Which is more effective?

Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark van den Brand, Alexander Serebrenik, Prem Devanbu, Vladimir Filkov