A Data Set for Social Diversity Studies of GitHub Teams

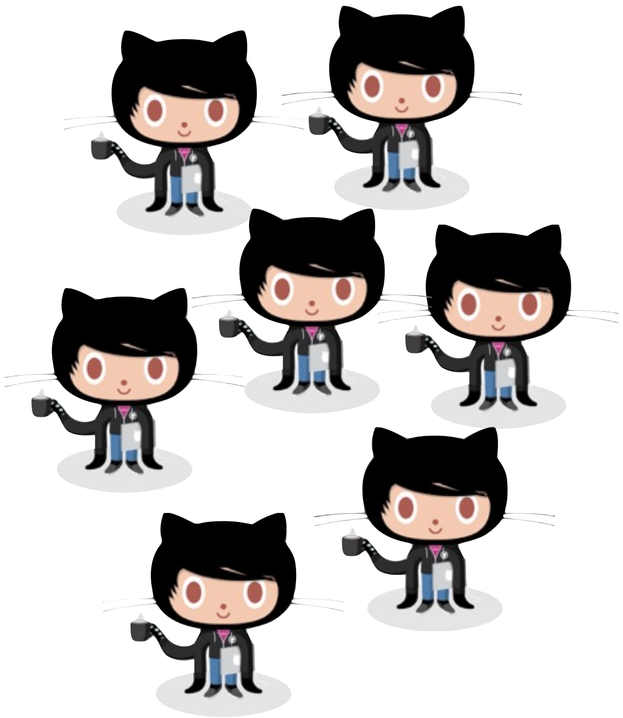
Bogdan Vasilescu Alexander Serebrenik Vladimir Filkov

UC DAVIS UNIVERSITY OF CALIFORNIA

TU/e Technische Universiteit Eindhoven University of Technology

@b_vasilescu @aserebrenik @vlfilkov

Which is more effective?



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

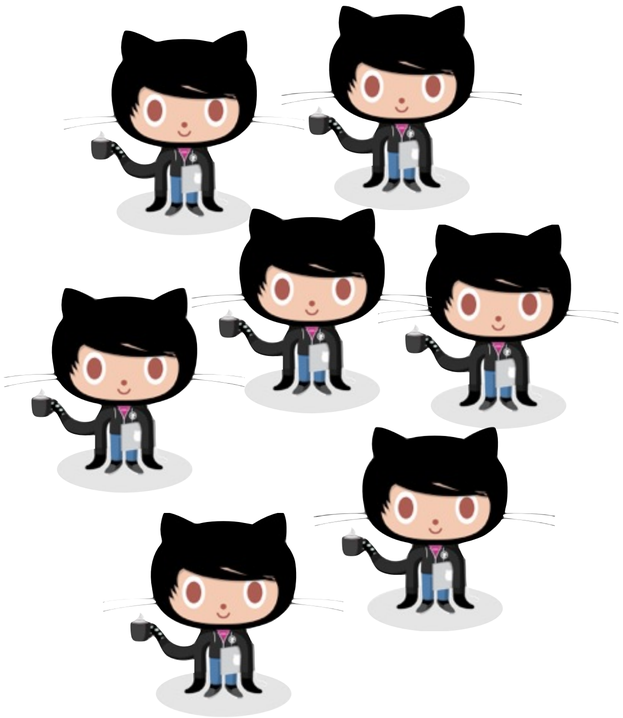
Software development teams can be *diverse* in various ways, *e.g.*, w.r.t. gender, experience, nationality, and coding language preference; some teams can be more diverse in one attribute and less so in others. Diversity attributes may also interact (*e.g.*, in some nations, female professionals may 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.

Which is more effective?



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.

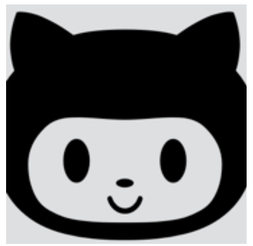
Social Diversity in GitHub Teams

Gender
= mix women/men

Tenure (experience)
= mix junior/senior

Culture
= mix countries

The Data Set



+

=

http://ghtorrent.org

bvasiles / **diversity**

Unwatch 1

A data set for social diversity studies of GitHub teams — Edit

4 commits **1** branch **0** releases **1** contributor

branch: **master** **diversity** / +

Updated to match camera-ready

bvasiles authored 21 days ago latest commit a1d6263472

LICENSE	Initial commit	2 months ago
README.md	Updated readme	2 months ago
diversity_data.csv	Updated to match camera-ready	21 days ago

README.md

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**

https://github.com/bvasiles/diversity

@b_vasilescu @aserebrenik @vlfilkov