A Data Set for Social Diversity Studies of GitHub Teams

Bogdan Vasilescu     Alexander Serebrenik     Vladimir Filkov
Which is more effective?

Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu†§, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand†, Alexander Seyrek, Premkumar Devanbu, Vladimír Filipkov§

†University of California, Davis and Eindhoven University of Technology
‡vlasilescu@ucdavis.edu, filikov@eindhoven.eu

Software development teams can be diverse in various ways, e.g. (sub)gender, experience, nationality and coding language preference; some teams can be more diverse in one attribute and less in other. Diversity attributes also differ across industries (e.g. in some nations, female professionals may face more obstacles), which complicates analysis and study. Team diversity has been studied in physical (‘near-space’) settings; however, data is hard to come by for such settings. Similar sample sizes make it difficult to effectively control for confounders. Data requirements for such effective controls, however, increase exponentially with the number of dimensions studied (one aspect of the ‘curse of dimensionality’ [22]).

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 (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

Copyright held by authors. Publication rights licensed to ACM. 978-1-4503-3145-6/15/04$15.00. http://dx.doi.org/10.1145/2702123.2702549

ABSTRACT
Software development is usually a collaborative venture. Open Source Software (OSS) projects are no exception. Indeed, today, open source software has become almost entirely of the collective development efforts of volunteers. GitHub is a social coding platform, has attracted millions of developers and thousands of Open Source Software projects. All comments, issues, code changes, pull requests etc. are archived and publicly available. GitHub has become the new standard for comprehensive study of social and technical characteristics of online open source teams. This 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, '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? In a software development team can be diverse in various ways, e.g. (sub)gender, experience, nationality and coding language preference; some teams can be more diverse in one attribute and less in other. Diversity attributes also differ across industries (e.g. in some nations, female professionals may face more obstacles), which complicates analysis and study. Team diversity has been studied in physical (‘near-space’) settings; however, data is hard to come by for such settings. Similar sample sizes make it difficult to effectively control for confounders. Data requirements for such effective controls, however, increase exponentially with the number of dimensions studied (one aspect of the ‘curse of dimensionality’ [22]).

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.

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? In OSS (open-source) software, effects of horizontal occupational software is much more centralized, and depends more on equal groups of programmers (19); in the spirit of, e.g. [high (20)], we explore this.
Which is more effective?

Software development teams can be diverse in various ways, e.g., gender, experience, nationality, and coding language preference. Some teams can be more diverse in one attribute and less so in others. Diversity attributes also depend on context (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-won in such settings. Smaller sample sizes make it difficult to effectively control for confounds. Data requirements for such effective control, however, increase exponentially with the number of dimensions studied (one aspect of the “curse of dimensionality” [22]). OSS teams, and must be considered carefully.

In this paper, we analyze both OSS and traditional teams. Furthermore, we examine multi-dimensional aspects of diversity. We use multiple datasets, one of which has a large enough dataset, however, the effect of increased gender diversity becomes noticeable. Additionally, since all data is OSS in nature (i.e., archived), it is possible to study the effects of tenure, or one’s length of time on a project and with OSS. However, the reliance on volunteers in OSS projects complicates analyses. Volunteers come and go, leading to team turnover. Team turnover can certainly influence performance, and well-controlled the effects of diversity. The constructs of “team” and “team turnover” clearly also depend on the dimension of diversity. For example, OSS teams, and OSS teams, and OSS teams. In this paper, we analyze both OSS and OSS teams. Furthermore, we examine multi-dimensional aspects of diversity. We use multiple datasets, one of which has a large enough dataset, however, the effect of increased gender diversity becomes noticeable. Additionally, since all data is OSS in nature (i.e., archived), it is possible to study the effects of tenure, or one’s length of time on a project and with OSS. However, the reliance on volunteers in OSS projects complicates analyses. Volunteers come and go, leading to team turnover. Team turnover can certainly influence performance, and well-controlled the effects of diversity. The constructs of “team” and “team turnover” clearly also depend on the dimension of diversity. For example,
Social Diversity in GitHub Teams

Gender
= mix women/men

Tenure (experience)
= mix junior/senior

Culture
= mix countries
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), and 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.

https://github.com/bvasiles/diversity