A Data Set for Social Diversity Studies of GitHub Teams

Bogdan Vasilescu
University of California, Davis
vasilescu@ucdavis.edu

Alexander Serebrenik
Eindhoven University of Technology
a.serebrenik@tue.nl

Vladimir Filkov
University of California, Davis
filkov@cs.ucdavis.edu

***Abstract*—Like any other team oriented activity, the software development process is effected by social diversity in the programmer teams. The effect of team diversity can be significant, but also complex, especially in decentralized teams. Discerning the precise contribution of diversity on teams' effectiveness requires quantitative studies of large data sets.**

Here we present for the first time a large data set of social diversity attributes of programmers in GITHUB teams. Using alias resolution, location data, and gender inference techniques, we collected a team social diversity data set of 23,493 GITHUB projects. We illustrate how the data set can be used in practice with a series of case studies, and we hope its availability will foster more interest in studying diversity issues in software teams.

I. INTRODUCTION

Social diversity is an important source of creativity and adaptability in teams [1], [2]. More socially diverse teams can leverage broader information, more varied backgrounds and ideas, and enhanced problem solving skills, therefore becoming more effective. However, social diversity comes at a cost. Due to greater perceived differences in values, norms, and communication styles in more diverse teams, members become more likely to engage in stereotyping, cliquishness, and conflict [3], [4], negatively effecting the team's cohesiveness and, therefore, its performance. Team social diversity has been studied mostly in physical (offline) teams [5], [6], where most results come from controlled experiments and small sample sizes, making it difficult to effectively control for confounds.

Recently, we started exploring social diversity in teams of Open Source Software (OSS) developers on GITHUB [7], [8]. We believe OSS to be a great source of data for studies of team social diversity, since: (i) software development is inherently a collaborative and human-centric activity; (ii) OSS is as much social as it is technical [9]–[12]; (iii) the self-organized, geographically-distributed, online nature of OSS leads to teams that are quite diverse, consisting of both professionals and volunteers, with varied personalities, educational and cultural backgrounds, age, gender, and expertise; (iv) OSS teams are real-world teams that form, evolve, and dissolve organically; generate measurable artifacts (*e.g.*, source code); and leave publicly-available traces of their activities.

In this paper we present a data set we started curating during prior work [7], [8], containing longitudinal quantitative data on 23,493 active GITHUB teams, including gender, location, and tenure information as team inputs, and amount of activity as team outcomes. We made the data available online at https://github.com/bvasiles/diversity, hoping its availability will foster more interest in studying diversity issues in software teams.

II. METHODS AND DATA

Using the GHTorrent [13], [14] dump 1/2/2014, we selected projects with at least 2 committers, 10 commits, and 6 months of history, *i.e.*, we filtered out inactive and non-collaborative projects. In line with [15], we consider a *project* to be a base repository and all its forks. Then, we enhanced this data by merging multiple identities used by the same developers, and resolving location, gender, and project application domains, as described below. We further selected only those projects where we could resolve the gender *or* country information for at least 75% of their team members, on average over each quarter (90 days) in their evolution. We consider a team to comprise all contributors to a project, not just committers, in line with [8]. Finally, to ensure they are at all active, we require that projects had, on average, at least 2 team members and 1 commit each quarter. The resulting data set has 23,493 projects and 93,056 rows of quarter-level data on the composition, characteristics, and outcomes of their teams of contributors.

A. Preprocessing

1) Merging Identities: Since on GITHUB the name and email address of committers and authors are set locally in each developer's git client, rather than globally at GITHUB level, there is variation in these attributes across devices and time. In addition, GHTorrent may introduce artificial user accounts when encountering contributions by "unknown" users while crawling data from GITHUB's API. To link the different aliases belonging to the same GITHUB contributors as well as deal with the issue of unknown GHTorrent aliases, we performed identity merging using a series of heuristics[1] similar to those in our prior work [16], [17]. To limit the number of false positives (*i.e.*, aliases incorrectly merged [18]), we have been as conservative as possible when deciding to merge aliases. For example, if multiple aliases share the same well-formed and non-fictitious email address, then we merge them, since we consider email addresses to be individual. Otherwise, if email addresses differ, then we only merge aliases for which we have collected sufficient evidence that they belong to the same person, from their first and last names, usernames, email address prefixes/domains, and locations. Using these heuristics we found that 170,062 users (out of 2,677,443 in the GHTorrent dump we analyzed) had more than one alias (median 2; mean 2.4; maximum 14). We also linked more than half of the "unknown" users to actual GITHUB accounts.

[1]Available online at https://github.com/bvasiles/ght_unmasking_aliases

978-0-7695-5594-2/15 $31.00 © 2015 IEEE
DOI 10.1109/MSR.2015.77

2) Making Sense of Location Data: On GITHUB, location descriptions on profile pages are free-text optional entries, therefore unstructured and often noisy. Besides actual geographic data (*e.g.*, city names, latitudes/longitudes, postcodes), they also include, *e.g.*, IP-addresses, */dev/null*, country-code top-level domains, and fictitious addresses (*e.g.*, "221b Baker Street, London", the protagonist's address in Sherlock Holmes stories). In our GHTorrent dump only 14.6% of the unique users (391,012) filled in this field, for a total of 55,388 distinct location strings. We process these strings to determine countries (an essential ingredient to inferring a person gender from their name [19]; see below), by combining information obtained from the Bing Maps API with information derived using a customized set of heuristics.[2] Our heuristics recognize, among others, postcodes in different countries, states in the USA and Brazil, provinces in Canada, and country-code top-level domains, as well as consult a large list of cities. This way, we resolved countries for 339,102 (86.7%) of the users with non-empty location entries.

3) Inferring Gender: We infer gender based on personal names and, if available, countries, using our `genderComputer` tool [19], with 93% precision. This approach combines a number of transformations, diminutive resolution, and heuristics (*e.g.*, users from Russia with surnames ending in *-ova* are female), with female/male frequency name lists collected for thirty different countries. If a country is unknown or not explicitly included in the name lists, the approach seeks agreement between all country name lists available. In this way, we can infer, *e.g.*, that *(Bogdan Lalić, -)* and *(Bogdan Lalić, Croatia)* are male, despite the unknown location for the former and the missing information about Croatia for the latter, since all country lists that include this first name record it as male. We could infer gender for 873,392 GITHUB contributors (32.6% of all users, but 80% of those who disclosed their names), labeling 91% as male and 9% as female. It follows that women are under-represented on GITHUB, similarly to other OSS (*e.g.*, [20]) and Stack Overflow [19].

4) Extracting Project Application Domains: We adopt the classification of GITHUB projects into different domains of Ray *et al.* [22]. This semi-automated approach (described in detail in [22]) uses LDA, a well known topic analysis algorithm, on project descriptions and Readme files, and is able to classify projects into 11 domains: Web, Mobile, Mid tier, GUI, Application, Program analysis, DB, Development framework, Library, Educational, and Other.

B. Measures

We compute a number of standard measures of OSS (GITHUB) activity using GHTorrent, 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 OSS [23], [24]), **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 2 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, 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 [25] of team **gender** and **country diversity**, a well-established diversity measure for categorical variables [7], [26], and coefficients of variation [27] for GITHUB, commit, and project tenure, as measures of team **tenure diversity**. Table I summarizes all these attributes.

[2]Available online at https://github.com/tue-mdse/countryNameManager

TABLE I
OVERVIEW OF OUR SOCIAL DIVERSITY DATA SET OF 23,493 PROJECTS.

Statistic	Mean	St. Dev.	Min	Median	Max
total_team	12.47	52.78	2	6	4,665
total_committers	9.21	32.68	2	5	3,384
total_commits	248.54	2,105.29	10	77	150,380
total_comments	46.37	448.26	0	3	37,292
total_issues	12.18	65.02	0	1	4,254
forks	21.66	139.52	0	5	5,516
watchers	45.04	290.60	0	6	9,139
num_quarters	6.23	4.76	2	5	24
prj_quarter_age	16.04	4.19	0	17	24
num_commits	57.45	294.06	1	16	22,688
num_comments	10.35	87.96	0	0	8,320
num_issues	2.83	13.94	0	0	858
num_committers	3.95	10.08	1	2	675
num_team	5.04	14.45	1	3	926
num_male	4.40	11.56	0	3	689
num_female	0.19	0.68	0	0	34
f_known_gender	0.96	0.10	0.00	1.00	1.00
f_known_country	0.56	0.36	0.00	0.53	1.00
blau_gender	0.04	0.12	0.00	0.00	0.50
blau_country	0.17	0.27	0.00	0.00	0.94
med_gh_tenure	784.10	458.14	0.00	781.50	2,123.00
med_prj_tenure	2.42	2.35	1.00	1.50	24.00
med_cmt_tenure	626.81	414.96	0.00	565.00	2,168.00
cv_gh_tenure	0.40	0.39	0.00	0.34	8.31
cv_prj_tenure	0.26	0.32	0.00	0.00	1.97
cv_cmt_tenure	0.42	0.34	0.00	0.41	3.29
turnover	0.66	0.29	0.004	0.60	1.00

III. APPLICATIONS

To illustrate the utility of this data set for studying the relationship between social diversity and technical activity in online teams, we present two novel case studies and describe, as a third case study, a full-length study published previously.

A. Case Study 1: Diversity Across Projects

Diversity in a project's team arises when team members are different from each other with respect to some attribute *e.g.*, gender, tenure, nationality, etc. Aggregate measures of diversity, such as the Blau index (for categorical variables) and

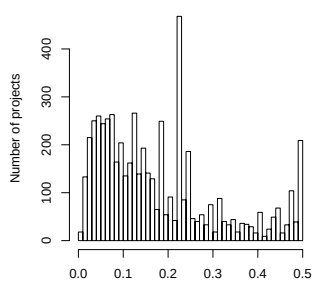

Fig. 1. Distribution of the average Blau index of gender diversity (5,624 diverse projects).

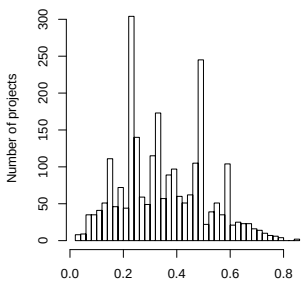

Fig. 2. Distribution of the average Blau index of country diversity (2,460 diverse projects).

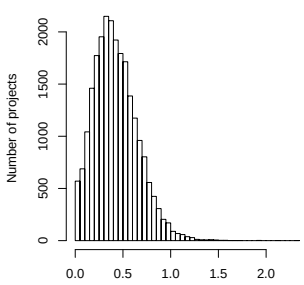

Fig. 3. Distribution of the average coefficient of variation of the commit tenure (all projects).

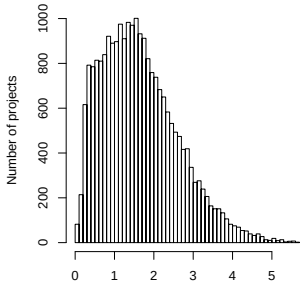

Fig. 4. Distribution of the average (over all quarters) median team commit tenure (in years; all projects).

the coefficient of variation (for numerical ones), are frequently used to capture how diverse groups are, *i.e.*, the higher the measures, the more diverse team members are with respect to a given attribute. For example, a team consisting only of same-gender members is not at all gender diverse; in contrast, a team can reach its maximal gender diversity by having equally many female and male members (assuming a simplified, binary gender), regardless of team size. Using the Blau index, gender uniformity is encoded as 0, while maximal gender diversity is encoded as 0.5 (since there are only two possible values, male and female).

Next, we illustrate harnessing this data set to characterize the distributions of gender, country, and tenure diversity.

1) Gender: We restrict the discussion here to those projects where we could resolve the gender for at least 75% of their team members, on average over all their quarters: 22,788 out of 23,493 projects (97%).

As expected, gender diversity has a very unbalanced distribution across projects. 17,164 (75.3%) projects in our data set had no team gender diversity at all, in any quarter. Among these, 171 (1%) projects are all-female projects, and the rest are all-male projects. Overall, average team size and average gender diversity are Spearman correlated at $\rho = -0.41$ ($p < 2.20 \times 10^{-16}$); and average team size and average number of commits at $\rho = 0.55$ ($p < 2.20 \times 10^{-16}$). As an example, we test whether there is any difference in the average team size and average number of commits per quarter between diverse and non-diverse projects. Since the groups are unbalanced, we subsample the non-diverse group 100 times. We find that gender diverse teams tend to be slightly larger on average, (median diverse: 4; median non-diverse: 3; WMW $p < 2.20 \times 10^{-16}$; Hodges-Lehmann point estimate for median difference $\hat{\Delta} = 1.00$), and responsible for more commits (median diverse: 30; median non-diverse: 18; WMW $p < 2.20 \times 10^{-16}$; $\hat{\Delta} = 9.05$). Figure 1 depicts the distribution of the average Blau index of gender diversity per project (over all quarters) among the 5,624 gender-diverse projects. We observe a broad spectrum of diverse teams, including 208 projects which achieved perfect team gender balance (Blau$= 0.5$) throughout their history,

2) Country: Here we restrict the discussion to only those projects where we could resolve the countries for at least 75% of their team members, on average over all their quarters: 3,922 out of 23,493 projects (17%). In this sample, in contrast to gender diversity, team country diversity is not a rare occurrence, as 2,460 (62.7%) projects did experience diversity during at least one of their quarters. The average Blau index of country diversity for these projects has the distribution given in Figure 2. Performing a similar analysis as above reveals statistically significant differences only in average team size (median for diverse = 2.1; median for non-diverse = 2.7; WMW $p < 2.20 \times 10^{-16}$), but negligible effects ($\hat{\Delta} = -0.42$).

3) Tenure: We illustrate tenure diversity using commit tenure, measured for each team member with prior GITHUB commit experience, per quarter, as the number of days since their first GITHUB commit until the end of the given quarter. Figure 3 depicts the distribution of the average measure of commit tenure diversity (computed over the different quarters in a project's history). The distribution is right skewed (skewness = 0.71) which indicates that most projects have low tenure diversity. We found, similarly, that the distribution of tenures is likewise right skewed (Figure 4). Together, these two results provide evidence that tenure diversity and team tenures are on average small, which makes it likely that on average teams with less experienced developers have less tenure diversity.

B. Case Study 2: Diversity Over Time

This data set can also be used in evolutionary settings, *e.g.*, to discover outliers and interesting trends. We illustrate this scenario in Figure 5, which plots the evolution of the Blau index of gender diversity in `hibernate-search` during six years (24 quarters) of history. One can observe how the project's team of contributors was gender balanced in the first two years (Blau index values above 0.4). However, with time, as the contributor team grew (team size shown dotted for comparison), it also became male-dominated. Recent years are characterized by stabilization, with at least one female contributor active each quarter (Blau index values around 0.3 for teams of size around 10).

C. Case Study 3: Diversity and Productivity

Regression studies on this data set can be used to assess the precise relationship between process outcomes and team diversity. As an illustration, here we describe our recent study [7] linking gender and tenure diversity to team productivity in GITHUB projects. There, we sought to model the outcome of

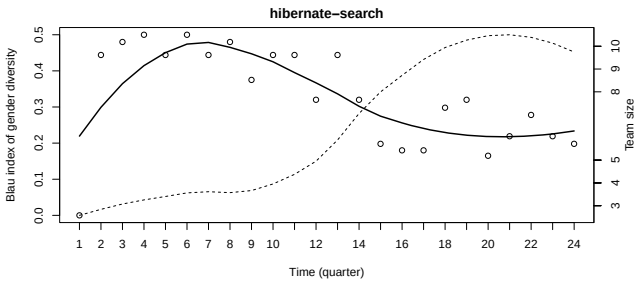

Fig. 5. Evolution of the gender diversity in `hibernate-search` during 6 years (24 quarters) of history. Team size shown dotted. The solid and dotted trend lines are Loess curves with 0.75 span.

productivity, captured through the number of commits by a project team per quarter, as a function of gender and tenure diversity. We used a number of control variables, including total number of commits, project age, size of the team, and number of committers. Our results show that there is a small (1-2.5%), but significant, positive effect on productivity by gender and tenure diversity for teams larger than 10 people. We also modeled the changes in team composition over time. Our results show that tenure has a large negative effect on turnover, while tenure diversity has a small, but positive effect on turnover. Gender diversity had no appreciable effect on turnover.

IV. Conclusions

The social and technical value of diverse teams is increasingly recognized in task oriented settings. Here we presented a comprehensive data set of diversity in GitHub teams with respect to gender and tenure. This is to our knowledge the first data set that resolves developer gender, tenure, and technical contributions, and on a systemic scale. The mashup of those attributes can prove valuable in the study of social diversity and its effectors in online teams. Moreover, studying the effects of diversity at large-scale can bring the benefits of increased resolution to quantitative studies. With that, smaller effects can be teased out, which while important, get swept under the "noise rug" in smaller studies.

Still, this data set has several limitations: a relatively small scale compared to GitHub (due to the generally noisy nature of GitHub data, and our constraint to have resolved gender or country for most team members); potential false positives (*e.g.*, aliases incorrectly merged, developers mislabeled with the opposite gender) and false negatives (*e.g.*, unidentified aliases), and missing temporal resolution on location data (which is a snapshot from the time of collection by GHTorrent).

V. Acknowledgments

We wholeheartedly acknowledge the help in assembling parts of this data set from members of our lab: Dr. Daryl Posnett and Dr. Baishakhi Ray. We also thank Prof. Prem Devanbu for critical comments.

References

[1] S. E. Jackson and A. Joshi, "Diversity in social context: a multi-attribute, multilevel analysis of team diversity and sales performance," *J. Organ. Behav.*, vol. 25, no. 6, pp. 675–702, 2004.

[2] W. E. Watson, K. Kumar, and L. K. Michaelsen, "Cultural diversity's impact on interaction process and performance: Comparing homogeneous and diverse task groups," *Acad. Manag. J.*, vol. 36, no. 3, pp. 590–602, 1993.

[3] D. de Gilder and H. A. M. Wilke, "Expectation states theory and the motivational determinants of social influence," *European Review of Social Psychology*, vol. 5, no. 1, pp. 243–269, 1994.

[4] E. Molleman and J. Slomp, "The impact of team and work characteristics on team functioning," *Hum. Factors Ergon. Manuf.*, vol. 16, no. 1, pp. 1–15, 2006.

[5] S. K. Horwitz and I. B. Horwitz, "The effects of team diversity on team outcomes: A meta-analytic review of team demography," *J. Manag.*, vol. 33, no. 6, pp. 987–1015, 2007.

[6] G. K. Stahl, M. L. Maznevski, A. Voigt, and K. Jonsen, "Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups," *J. Int. Bus. Stud.*, vol. 41, no. 4, pp. 690–709, 2010.

[7] B. Vasilescu, D. Posnett, B. Ray, M. G. J. van den Brand, A. Serebrenik, P. Devanbu, and V. Filkov, "Gender and tenure diversity in GitHub teams," in *CHI*. ACM, 2015, to appear.

[8] B. Vasilescu, V. Filkov, and A. Serebrenik, "Perceptions of diversity on GitHub: A user survey," in *CHASE*. IEEE, 2015, to appear.

[9] M. Gharehyazie, D. Posnett, B. Vasilescu, and V. Filkov, "Developer initiation and social interactions in OSS: A case study of the Apache Software Foundation," *Emp. Softw. Eng.*, pp. 1–36, 2014.

[10] N. Bettenburg and A. E. Hassan, "Studying the impact of social structures on software quality," in *ICPC*. IEEE, 2010, pp. 124–133.

[11] N. Nagappan, B. Murphy, and V. Basili, "The influence of organizational structure on software quality: an empirical case study," in *ICSE*. ACM, 2008, pp. 521–530.

[12] J. T. Tsay, L. Dabbish, and J. D. Herbsleb, "Influence of social and technical factors for evaluating contribution in GitHub," in *ICSE*. ACM, 2014, pp. 356–366.

[13] G. Gousios, "The GHTorent dataset and tool suite," in *MSR*. IEEE, 2013, pp. 233–236.

[14] G. Gousios, B. Vasilescu, A. Serebrenik, and A. Zaidman, "Lean GHTorrent: GitHub data on demand," in *MSR*. ACM, 2014, pp. 384–387.

[15] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, "The promises and perils of mining GitHub," in *MSR*. ACM, 2014, pp. 92–101.

[16] B. Vasilescu, A. Serebrenik, M. Goeminne, and T. Mens, "On the variation and specialisation of workload–A case study of the Gnome ecosystem community," *Emp. Softw. Eng.*, vol. 19, no. 4, pp. 955–1008, 2014.

[17] E. Kouters, B. Vasilescu, A. Serebrenik, and M. G. J. van den Brand, "Who's who in Gnome: Using LSA to merge software repository identities," in *ICSM*, 2012, pp. 592–595.

[18] M. Goeminne and T. Mens, "A comparison of identity merge algorithms for software repositories," *Science of Computer Programming*, vol. 78, no. 8, pp. 971–986, 2013.

[19] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study," *Interacting with Computers*, vol. 26, no. 5, pp. 488–511, 2014.

[20] G. Robles, L. Arjona-Reina, B. Vasilescu, A. Serebrenik, and J. M. Gonzalez-Barahona, "FLOSS 2013: A survey dataset about free software contributors: challenges for curating, sharing, and combining," in *MSR*. ACM, 2014, pp. 396–399.

[21] D. M. Blei, "Probabilistic topic models," *CACM*, vol. 55, no. 4, pp. 77–84, 2012.

[22] B. Ray, D. Posnett, V. Filkov, and P. T. Devanbu, "A large scale study of programming languages and code quality in GitHub," in *FSE*. ACM, 2014, pp. 155–165.

[23] P. J. Adams, A. Capiluppi, and C. Boldyreff, "Coordination and productivity issues in free software: The role of Brooks' law," in *ICSM*. IEEE, 2009, pp. 319–328.

[24] S. Daniel, R. Agarwal, and K. J. Stewart, "The effects of diversity in global, distributed collectives: A study of open source project success," *Inform. Syst. Res.*, vol. 24, no. 2, pp. 312–333, 2013.

[25] P. M. Blau, *Inequality and heterogeneity: A primitive theory of social structure*. Free Press New York, 1977, vol. 7.

[26] J. Chen, Y. Ren, and J. Riedl, "The effects of diversity on group productivity and member withdrawal in online volunteer groups," in *CHI*. ACM, 2010, pp. 821–830.

[27] P. D. Allison, "Measures of inequality," *American Sociological Review*, pp. 865–880, 1978.