Social Diversity and Growth Levels of Open Source Software Projects on GitHub

Joop Aué,[*] Michiel Haisma,[*]
Kristín Fjóla Tómasdóttir[*]
Delft University of Technology
Delft, The Netherlands
{j.aue,m.a.haisma,k.f.tomasdottir}
@student.tudelft.nl

Alberto Bacchelli
Delft University of Technology
Delft, The Netherlands
a.bacchelli@tudelft.nl

ABSTRACT

Background: Projects of all sizes and impact are leveraging the services of the social coding platform GitHub to collaborate. Since users' information and actions are recorded, GitHub has been mined for over 6 years now to investigate aspects of the collaborative open source software (OSS) development paradigm. **Aim:** In this research, we use this data to investigate the relation between project growth as a proxy for success, and social diversity. **Method:** We first categorize active OSS projects into a five-star rating using a benchmarking system we based on various project growth metrics; then we study the relation between this rating and the reported social diversities for the team members of those projects. **Results:** Our findings highlight a statistically significant relation; however, the effect is small. **Conclusions:** Our findings suggest the need for further research on this topic; moreover, the proposed benchmarking method may be used in future work to determine OSS project success on collaboration platforms such as GitHub.

CCS Concepts

•**Software and its engineering** → *Open source model;*

Keywords

social diversity; GitHub; software project growth

1. INTRODUCTION

The global nature of open source software (OSS) projects enables developers to participate and contribute code, regardless of their background, location, and social attributes. As a consequence, studying these projects and their publicly available data is a ripe opportunity for researchers interested in investigating social diversity aspects and their relationship with a number of project's characteristics. For example, social diversity is known to have a positive effect on teamwork, productivity, and quality of performance [10,15]. In this short paper we focus on this line of research and present the results of an initial investigation on the relation between *social diversity* (in terms of country and gender diversity) and *project growth* (measured through a number of proxy based metrics).

To this aim, we propose a way of rating active OSS projects to determine their success in terms of growth. In practice, grounded on previous research [6, 9, 13, 14], we use the notion of growth as a proxy for project success as it reflects the project's activity levels and developer and user interest in the project, such as the number of team members, commits, comments, and pull requests in a size-independent manner. Note that the link to success is related to the development rate of a project as opposed to usage and community success; at the same time, the activity on a project's GitHub reflects not just developers, but also (engaged) users, as anyone can raise issues and make comments.

We use a method, based on previous research by Alves *et al.* [1,2], to rate projects based on thresholds empirically from the calculated growth values. We then compare the rating of a project to its social diversity factors.

To access information about social diversity and growth metrics, we take advantage of the data prepared by Vasilescu *et al.* [16] who used the GHTorrent project [8]. The data includes a number of social features (gender, location, and tenure of the team members were inferred where possible) and is aggregated quarterly, providing a time dimension.

Our initial results highlight a statistically significant relation between project growth and both diversity metrics, but with small effect, suggesting the need for further analysis on this topic.

2. RELATED WORK

We analyse previous work investigating diversity in OSS projects and defining success metrics based on growth.

2.1 Diversity in OSS

Daniel *et al.* [6] report a positive correlation between market success in projects hosted on SourceForge [7] and cultural diversity (measured by observing used language and devel-

[*] Aué, Haisma, and Tómasdóttir contributed equally to the work and are to be considered all first authors. This work was developed as part of the Master course "Mining Software Repositories" at Delft University of Technology.

10th International Symposium on Empirical Software Engineering and Measurement '16 Ciudad Real, Spain

© 2016 ACM. ISBN 123-4567-24-567/08/06.

DOI: 10.475/123_4

oper nationality); they find that linguistic diversity negatively impacts community participation, but speculate that this may be due to the language barriers.

Vasilescu *et al.* [16] conduct a survey among GitHub contributors on how diversity is perceived and analyse the data set we use in this research. They report that gender diversity has a positive effect on productivity (measured as the number of commits in a project) on all teams, regardless of the size, while tenure diversity has a positive effect on medium and large teams.

In line with this work, we study diversity factors, but shift the focus from productivity to growth. Additionally we narrow down the dataset of active GitHub projects to only examine those that can be considered long-term projects.

2.2 Defining project success

Previous work on defining software projects' success based on growth, particularly in the OSS domain, highlights key measures to consider: Growth, community related aspects, and interest of developers and users. These are all factors we include in our analysis considering different angles (detailed in Section 3.1).

In detail, Crowston *et al.* [5] establish the importance of a number of the aforementioned measures through literary review and interviews, particularly individual and organizational metrics and growth; Subramaniam *et al.* [14] show that developer and non-developer interest in OSS projects and project activity levels in any time period significantly affect the project success measures in subsequent time periods, for example, one of the factors negatively impacting OSS success is a restrictive license, as this decreases developer interest, although this factor does increase user and project administration interest; Lee *et al.* [11] present a theoretical framework for performing empirical research concerning OSS success, in particular they find that information system success models can be applied to the OSS context and that the OSS success model shares similarities and differences with other contexts; McDonald and Goggins [13] report indicators that project leads and core developers of three major OSS projects value community aspects (*e.g.*, contributor growth, community involvement) more than code-related metrics when evaluating the success of their project.

3. METHOD

Previous research investigated how aspects of social diversity related to productivity, quality, or efficiency of OSS projects. Our goal is to further investigate this angle of comparison by defining a rating that encapsulates different aspects of project growth and relate it with two angles of social diversity, namely gender and geographical diversity. We structure our analysis on the following research questions:

> RQ1: How does gender diversity relate to the growth of open source projects?

> RQ2: How does geographical diversity relate to the growth of open source projects?

3.1 Growth metrics and success rating

Central to our analysis is defining a rating to evaluate growth, an indication of success. Given the high variance in projects' size (*e.g.*, number of developers and lines of code), we discard metrics related to size and, instead, we focus on change, in particular *growth.* This dimension, in fact, is less dependent on project size and McDonald *et al.* [13] found that (contributor) growth is a reasonable metric for project success. We now motivate the metrics we consider:

Team growth. The team includes all participating parties of a project (*e.g.*, committers and pull request submitters). Considering team size as a proxy for developer and user interest, it can be seen as an indicator for project success [14]. McDonald *et al.* found that contributor growth is the most commonly mentioned measure of success in OSS projects [13]. This metric is computed by counting the team members in a quarter. The other metrics are computed similarly.

Commit growth. Commits in OSS projects have commonly been used to define a project's productivity and success [6,9]: Changes done on the code reflect the project advancing.

Pull request growth. Specific to OSS projects, and, in particular, to GitHub, pull requests can serve as a proxy for a number of success measures including community activity level and developer interest. Developers interviewed by McDonald *et al.* [13] also noted that pull requests stimulate a democratic and transparent environment, which attracts other developers.

Comment growth. Comments (which can be attached to commits, pull requests and issues) indicate interest of both developers and users, thus making it useful to observe their growth. Moreover, comments serves a measure of the project's social activity and community engagement [6, 16]. The project community activity level is also considered to be positively correlated with project success [14].

After choosing these metrics, we have to establish how to compute growth in practice. We analyse the last 5 quarters of each project to calculate the growth over the last year of a project (5 data points are necessary to reflect upon 4 quarters of growth); we have found this number of quarters to be a good trade-off between having enough data to represent a project and reflecting the current state of the project. Having a fixed period makes sure that the growth metrics are comparable across projects regardless of the project age. Some projects may have already existed for many years, but this research focuses on a more current state of the projects. Moreover, this number prevents biases due to sudden quarterly increase, without adding data that is not relevant to the current state of the project. This method also excludes newer projects that do not exist long enough for 5 data points to have been collected.

We found the average quarterly growth to *overestimate* the growth (as growth spikes had a large influence on the results), thus we use the Compound Annual Growth Rate (CAGR) [3]. This metric is a measure of growth over multiple time periods and is often used to find the mean annual growth rate of investments. Based on the CAGR, we compute the quarterly growth rate of each metric as follows (p_n is the metric observed in the most recent quarter available):

$$\text{quarterly growth rate} = \sqrt[4]{\frac{p_n}{p_{n-4}}} - 1 \tag{1}$$

3.2 Project rating

After computing metrics and growth rate, we use them to determine a rating of each project relative to the others we consider.

We adopt the approach proposed by Alves *et al.*, who define a 1-to-5 star system (where 1 is the lowest outcome and 5 is the highest) to rate the relative success of a project [1,2]. The idea is to compute metric thresholds from a benchmark of software systems [2] and assign stars relatively to the position of analysed project with respect to these thresholds. By using such a system one can compare projects relatively on a high-level base without making a clear distinction between projects that get the same rating; moreover, the approach takes statistical properties of the metrics into account and is resilient against outliers.

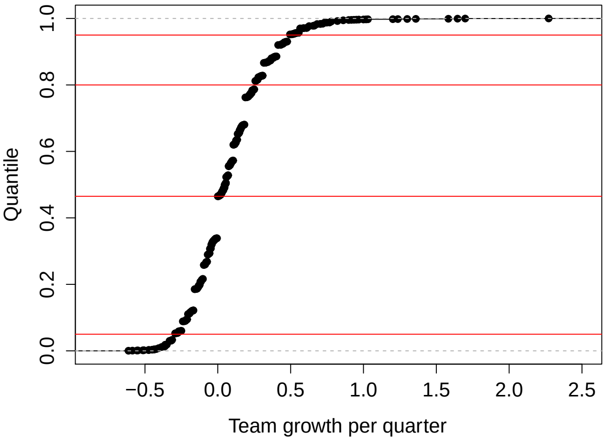

Figure 1: Team growth: Cumulative distribution with rating thresholds.

For each metric that contributes to the growth of a project, we calculate the growth as described earlier, then we plot each growth metric to manually judge where to appropriately place the thresholds for the benchmarking. Figure 1 shows the distribution of the team growth measures and the corresponding thresholds (the other growth metrics have similar cumulative distributions). Alves *et al.* [2] propose thresholds to be the 70%, 80% and 90% percentiles, which fits the left skewed distribution of their data. These thresholds do not apply correctly to the data used in our research, because they are not close to each tail. For this reason, and based on a visual assessment of our data, we place our thresholds in the 5% or 10% and 95% percentiles: The bottom threshold varied between 5% and 10% depending on the stretch of the bottom tail of the distribution; furthermore, since the measures represent growth, we place a threshold where the growth is zero; the fourth threshold is placed at the 80% percentile since it creates a category with the top 20% rated projects. The exact placement of the threshold is not critical, considering that the thresholds are used in order to categorise projects into high-level groups.

We map these thresholds using the star-rating system described in the following:

1 star: < 5% or 10% percentile

2 stars: 5% or 10% percentile - 0 growth

3 stars: 0 growth - 80% percentile

4 stars: 80% - 95% percentile

5 stars: > 95% percentile

In this system, 1- and 2-star projects show negative growth (development rate or interest in the project is declining), 3-star ones zero or positive growth, and 4- and 5-star ones the most rapid growth. The overall rating for each project is computed as the average of the ratings for the four growth metrics described earlier.

3.3 The dataset: Subject systems and social diversity metrics

To conduct our investigation, we take advantage of the dataset prepared by Vasilescu *et al.* [16]. It includes development and social information about 23,493 projects on GitHub that are both active (*i.e.*, with at least six months of history and at least one commit on average in each quarter) and collaborative (*i.e.*, with at least two team members on average in each quarter). All information about the teams and the development history is given quarterly over the project's lifetime.

To compare the projects based on the metrics that are used to define the success (team size, commits, pull requests and comments), we remove projects which did not include at least one pull request and one comment in its lifetime. We further remove projects for which there is no available data for the last 5 consecutive quarters. This filtering is performed because only long-term active OSS projects are considered. A project that has no commits during a 3 month period is not considered active enough to be relevant for this comparison. After this filtering is applied, we conduct our analysis on a total of 3,203 projects.

The dataset includes social measures on demographic diversity, namely the geographic location, gender, and experience of the team members in each project. In this investigation, we focus on inferred gender and country diversity of the project's team members. The diversity of both gender and country is computed for each quarter as the Blau index, a well-established diversity measure for categorical variables [4]. This index computes the probability that entities taken at random from the dataset of interest (with replacement) belong to a different category. *E.g.*, given a group of people, in which both genders are evenly represented, the computed Blau index is 0.5. With more categories the highest possible index increases up to the limit of 1. A single diversity value for each project was computed as the average Blau index over the last 4 quarters of its lifetime.

4. RESULTS

Initially, the ratings are calculated for each individual growth metric. As an indication, the country diversity for the comment growth metric is shown in Figure 2. This shows the result for a single metric.

Next, we computed the overall rating, the average of the four growth metric ratings, for the $3,203$ subject projects. Table 1 shows the resulting distribution of the projects in the different rating. Note that, because of using the average, we consider continuous values for the overall rating. Projects rate most frequently in the $[2,3)$ stars range, while score the least frequently in the $[4,5]$ star interval.

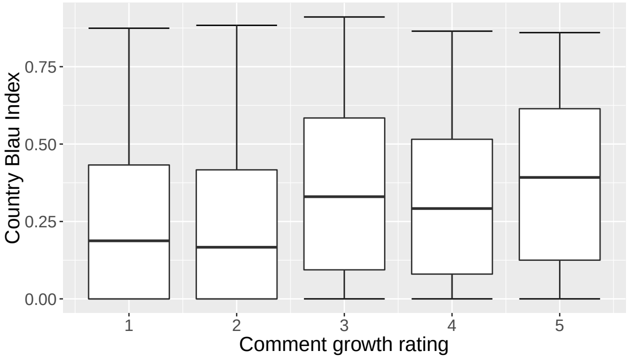

Figure 2: Median country diversity (Blau index), by comment growth rating.

Table 1: Overall rating of projects

Stars	Number of projects
$[1,2)$	491
$[2,3)$	1,547
$[3,4)$	923
$[4,5)$	230
5	12
Total	**3,203**

4.1 Comparing growth and social diversity

We aim to use a single value per rating to differentiate between 1 to 5 star projects. By analysing the distribution of both gender and country diversity per rating, we found that in both cases the data is right-skewed, especially the gender index having a large number of zero (0) values. In the first instance, we thus consider the median to obtain a single value per rating for both gender and country diversity. Figure 3 and Figure 4 are the resulting box plots (for readability reason, these plots do not follow the custom layout of plotting the independent variable vertically).

4.2 Gender diversity

Considering Figure 3, we do not see a strong relation between gender diversity and overall project rating. Also by splitting the rating into the different dimensions, we found no strong noticeable pattern. We also computed Spearman's correlation between the two variables, but the value is less than 0.065, thus showing no correlation.

With more analysis, we exposed a statistically significant relation, but with minor effect. First, we built a ordered logistic regression model [12] to describe overall rating through gender and country diversity. The low number of cases for many pairs of overall rating and gender/country Blau index may limit the reliability of the model. For this reason, we consider that there is a large amount of zero (0) values for the gender diversity (*i.e.*, a project has either all male or all female team members) and we split it in two bins (Blau index equals to 0 and higher than 0, respectively). Similarly, we split the country diversity in three bins (0, less than the median, more than the median).

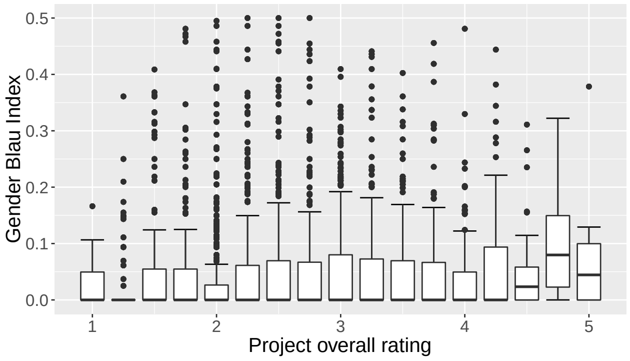

Figure 3: Median gender diversity (Blau index), by overall average rating.

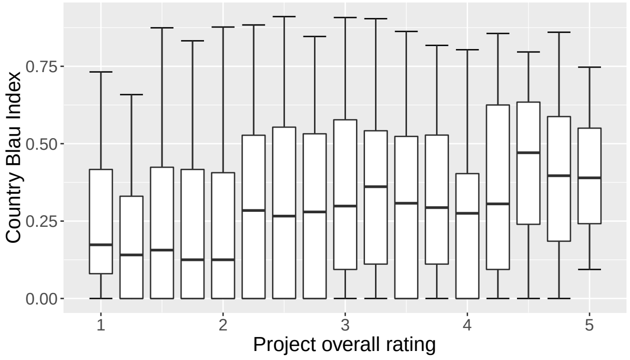

Figure 4: Median country diversity (Blau index), by overall average rating.

The resulting model reports gender diversity to be statistically significant (associated p-value < 0.001) and its parameter estimate to be 0.24. This means that for a unit increase in gender diversity (i.e., going from 0 to 1), we expect a 0.24 increase in the ordered log odds of being in a higher level of overall rating, given all of the other variables in the model are held constant. Additionally, we plot the means of the resulting overall ratings. The left hand side of Figure 5 shows the results for gender: We see a statistically significant difference, but both results are still within the (2.6,2.8) range.

4.3 Country diversity

In Figure 4, we see that the medians in country diversity are growing as the rating grows, so does variability. The Spearman's correlation between the two variables is 0.08, thus negligible. Among the different dimensions composing the overall rating, we found the highest Spearman's correlation to be with the team growth rating (*i.e.*, 0.10), yet not relevant.

Similarly to the gender diversity, with further analysis we exposed a statistically significant relation, but an effect that is even smaller than that of gender diversity. The aforementioned ordered logistic regression model reported country

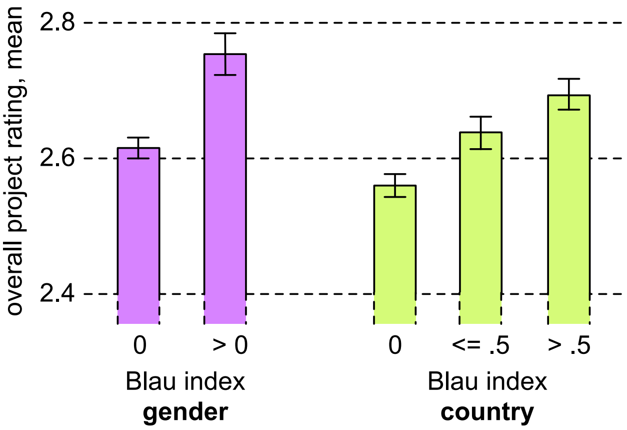

Figure 5: Mean overall project rating, by discretised diversities

diversity to be statistically significant (associated p-value < 0.001), with a parameter estimate of 0.15. The right hand side of Figure 5 shows the values for country diversity: We see a statistically significant difference, but both results are still within the (2.5,2.7) range.

5. THREATS TO VALIDITY

Aspects of our research may pose a threat to its validity. The data is filtered to consider 3,203 projects; by applying this filtering (in addition to the original filtering), the resulting dataset may not be representative of all projects on GitHub. Also, the growth of a project can be calculated in multiple ways and over different time periods; we found it to be appropriate to use the last year of a project's history, but changing that might yield different results. For projects that have had largely fluctuating growth, we might observe a period where our start of measuring coincides with the project's start of growth; this also may occur for young projects that have exactly 5 quarters of history where starting with a low number of team members and commits might result in a favourable representation of growth. On the other hand, projects that have been successful in the past but are now only being maintained might be unfavourably represented as they are no longer growing.

6. DISCUSSION

Our results report a statistically significant relation between project rating and the considered diversity metrics, yet the reported effect is minor. We found similar results when considering the growth metric ratings separately.

Regarding the gender diversity, our initial findings complement those by Vasilescu *et al.* [15] who use the original dataset that we constructed our sample from. They found gender diversity to have a very significant, positive effect on productivity (measured as the number of commits to a project). Directly comparing our results to that of Vasilescu *et al.* is not possible for several reasons, including the following ones: (1) By filtering the dataset from 23,493 projects down to 3,203, we analyse a different sample that may not be representative for all cases; (2) Vasilescu *et al.* consider three models: small teams, medium-sized teams and large teams, while the success metrics we use are team size independent; (3) we compare the gender diversity based on growth as a proxy for success while Vasilescu *et al.* look at productivity.

Further work could be done by investigating how projects change over time regarding growth or success, and social diversity. We computed growth over time and compared to the average social diversity over the same time period; one could look at the change over a project's lifetime and see whether there is a trend in the change of success and diversity.

Also we identified 12 projects with a maximum 5 star rating. Their reasons for success are interesting to investigate in future work.

7. CONCLUSION

In this study, we investigated the relation between long-term active OSS project growth, which we used as a proxy for success, and both gender and country diversity metrics in GitHub projects. Using a GHTorrent based dataset curated by Vasilescu *et al.* [16], we calculated a number of growth metrics from active projects and used them to define a 5-star rating system for project success. We found a statistically significant, but minor relation between project success and both diversity metrics.

It is our hope that these results will trigger further analysis on these aspects to more generally determine factors related to diversity. We also hope that our method for objectively rating OSS projects may be used in similar research that requires an objective multi-dimensional rating OSS growth.

8. REFERENCES

[1] T. L. Alves, J. P. Correia, and J. Visser. Benchmark-based aggregation of metrics to ratings. In *Software Measurement, 2011 Joint Conference of the 21st Int'l Workshop on and 6th Int'l Conference on Software Process and Product Measurement (IWSM-MENSURA)*, pages 20–29. IEEE, 2011.

[2] T. L. Alves, C. Ypma, and J. Visser. Deriving metric thresholds from benchmark data. In *Software Maintenance (ICSM), 2010 IEEE International Conference on*, pages 1–10. IEEE, 2010.

[3] M. J. Anson, F. J. Fabozzi, and F. J. Jones. *The handbook of traditional and alternative investment vehicles: investment characteristics and strategies*, volume 194. John Wiley & Sons, 2010.

[4] P. M. Blau. *Inequality and heterogeneity: A primitive theory of social structure*, volume 7. 1977.

[5] K. Crowston, H. Annabi, and J. Howison. Defining open source software project success. *ICIS 2003 Proceedings*, page 28, 2003.

[6] S. Daniel, R. Agarwal, and K. J. Stewart. The effects of diversity in global, distributed collectives: A study of open source project success. *Information Systems Research*, 24(2):312–333, 2013.

[7] GitHub. http://sourceforge.net/. Accessed 2016/01/14.

[8] G. Gousios. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13, pages 233–236, Piscataway, NJ, USA, 2013. IEEE Press.

[9] R. Grewal, G. L. Lilien, and G. Mallapragada. Location, location, location: How network

embeddedness affects project success in open source systems. *Management Science*, 52(7):1043–1056, 2006.

[10] S. K. Horwitz and I. B. Horwitz. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management*, 33(6):987–1015, 2007.

[11] S.-Y. T. Lee, H.-W. Kim, and S. Gupta. Measuring open source software success. *Omega*, 37(2):426 – 438, 2009.

[12] P. McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.

[13] N. McDonald and S. Goggins. Performance and participation in open source software on github. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 139–144, New York, NY, USA, 2013. ACM.

[14] C. Subramaniam, R. Sen, and M. L. Nelson. Determinants of open source software project success: A longitudinal study. *Decision Support Systems*, 46(2):576–585, 2009.

[15] B. Vasilescu, D. Posnett, B. Ray, M. G. van den Brand, A. Serebrenik, P. Devanbu, and V. Filkov. Gender and tenure diversity in github teams. In *CHI. ACM*, 2015.

[16] B. Vasilescu, A. Serebrenik, and V. Filkov. A data set for social diversity studies of github teams. In *Mining Software Repositories (MSR), 2015 IEEE/ACM 12th Working Conference on*, pages 514–517. IEEE, 2015.