Journal of Software Engineering Research and Development

RESEARCH Open Access

How diverse is your team? Investigating gender and nationality diversity in GitHub teams

Marco Ortu[1], Giuseppe Destefanis[2*], Steve Counsell[3], Stephen Swift[3], Roberto Tonelli[1] and Michele Marchesi[1]

*Correspondence:
g.destefanis@herts.ac.uk
[2]University of Hertfordshire, Hatfield, UK
Full list of author information is available at the end of the article

Abstract

Background: Building an effective team of developers is a complex task faced by both software companies and open source communities. The problem of forming a "dream" team involves many variables, including consideration of human factors and it is not a dilemma solvable in a mathematical way. Empirical studies might provide interesting insights to explain which factors need to be taken into account in building a team of developers and which levers act to optimise productivity among developers.

Aim: In this paper, we present the results of an empirical study aimed at investigating the link between team diversity (i.e., gender, nationality) and productivity (issue fixing time).

Method: We consider issues solved from the GHTorrent dataset inferring gender and nationality of each team's members. We also evaluate the politeness of all comments involved in issue resolution.

Results: Results show that higher gender diversity is linked with a lower team average issue fixing time *(higher productivity)*, that nationality diversity is linked with lower team politeness and that gender diversity is linked with higher sentiment.

Keywords: Affective analysis, Issue report, Empirical software engineering

1 Introduction

Software development is an activity organized around team-based environments. The implementation of team structures is not a simple process and does not necessarily result in success; it is not enough just to put people together in teams and presume that everybody agrees on what to do (Allen and Hecht 2004). In such a context, any conflict is likely to affect a team's productivity. Team leaders have a vested interest in knowing how to prevent, avoid, or, in the worst case, manage conflicts which might occur and in understanding the key factors that make a team healthy and productive.

Diversity in working teams has been studied in several research fields (Horwitz and Horwitz 2007; Stahl et al. 2010) and is considered as any attribute which differentiates people (Williams and O'Reilly 1998) such as demographic attributes (e.g., age, gender, nationality), function (e.g., role, tenure, expertise), or subjective features (e.g., personality). Previous research reports contrasting evidence about the role of diversity in team work: some studies report significant positive correlations between diversity and performance (Earley and Mosakowski 2000), while others have found that diversity negatively impacts team outcomes (Watson et al. 1993). Herring (2009) used data from the 1996 to 1997 National Organizations Survey to test eight hypotheses derived from the value-in-diversity thesis. The results supported seven of these hypotheses: racial diversity was associated with increased sales revenue, more customers, greater market share and greater relative profits. Gender diversity was associated with increased sales revenue, more customers and greater relative profits.

As far as open source development is concerned, Daniel et al. (2013) studied effects of diversity on community engagement and market success in a sample of 357 SourceForge projects. Results showed that reputation and role diversity were positively correlated with market success and community engagement. Conversely, diversity of spoken languages and nationality were negatively associated with community engagement and positive impact on market success. Diversity in experience and language was studied by Chen et al. (2010) using Wikipedia Projects. The authors examined the effects of group diversity on the amount of work accomplished and on member withdrawal behaviours in the context of WikiProjects. They found that increased diversity of experience with Wikipedia increased group productivity and decreased member withdrawal up to a threshold. Beyond that threshold, group productivity remained high, but members were more likely to withdraw.

Our starting point is the fact that diversity, in the broad sense of the term, is a positive factor which brings added value to a group of people working together. Diversity stimulates changes. For example, people with different cultural backgrounds need to understand how to communicate and interact, since the concept of "good manners" in a Western country might differ from the same concept in an Eastern country.

In this paper, we investigate the impact of gender and nationality diversity on the productivity of a team. Based on evidence of previous research (Murgia et al. 2014; Mäntylä et al. 2016; Ortu et al. 2015a, c), we used the issue fixing time and politeness as proxy metrics for team productivity and level of communication, respectively. We present three logistic regression models modelling the average time required to solve an issue by development teams and the communication level measured by the politeness of a team. We exploit a dataset extracted from the 2014 dump of the GHTorrent dataset (Gousios 2013). A set of heuristics was used to infer development teams based on GitHub's issue collaboration graph, its user's gender and nationality with the final goal of building a representative diversity dataset. We decided to focus our attention on understanding the impact of gender and nationality factors, because the related information are either present or inferable from the GitHub repository, while for data about other diversity factors, e.g., religion, it is not possible to gain any insight. In addition, we think that the nationality factor might be representative of a broad cultural background (e.g., language spoken), otherwise difficult to compute. We explore the following research questions:

1.1 RQ1: Are gender or nationality diversity linked to the issue fixing time of a team?

Gender diversity in GitHub teams was linked with lower issue fixing time Our model showed that gender diversity was the most dominant metric explaining data variance and was positively associated with productivity (we observed shorter issue fixing time in teams with higher gender diversity).

1.2 RQ2: Are gender or nationality diversity linked to the overall politeness of a team?

Country diversity was linked with lower politeness. Our model showed that country diversity was the most dominant metric explaining data variance and had a negative effect on team politeness (tended to lower politeness).

1.3 RQ3: Are gender - nationality diversity linked to the overall sentiment of a team?

Gender diversity was linked with higher sentiment. Our model showed that gender diversity was the most dominant metric explaining data variance and had a positive effect on team sentiment (it had a tendency to increase politeness).

The remainder of the paper is organized as follows: in Section 2 we discuss the related work. In Section 3, we describe how we measured politeness and how issue collaboration graphs were built. Section 4 introduces the experimental design; in Section 5 we present the case study setup, while in Section 6 we present our findings followed by a discussion of the issues raised in Section 7. In Section 8 we analyse the threats to validity before drawing conclusions (Section 9).

2 Related work

2.1 Gender and diversity in software development

A range of studies have highlighted gender and diversity issues in computer science. Sheldon (2004) analysed gender stereotypes in educational software for pre-school children, finding evidence of gender inequality. The study found many more male characters in education software and that female characters presented "counter-stereotypical behaviours". A recent study by Medel and Pournaghshband (2017) in 2017 showed that gender bias is still present in computer science instructional materials. Blum et al. (2007) shifted the focus, demonstrating that the way men and women relate to computing is mainly a result of cultural and environmental conditions, highlighting the fact that gender inequality is a cultural consequence. The authors stated thaT it WAS necessary "*to look beyond gender to account for differences in the experiences and perspectives of men and women*". In a subsequent study, Cheryan et al. (2009) obtained similar results performing experiments with computer science students and concluded that a change of topics considered stereotypical (e.g., video games) towards more neutral topics was enough to increase the interest in computer science of female undergraduates to the level of their male peers.

Other studies have focused on team diversity, studying the impact on groups dynamic and productivity. Giuri et al. (2010) focused on how the variety and level of a developer's skills impacted project performance; both factors positively affected both survival and performances of a project. Ren et al. (2015) studied on-line self-organising groups in the context of Wikipedia projects and explained how group diversity affected the performance of the group. Members of a group can easily get tangled up around each other with expectations, misunderstandings and feelings. For an organisation or team to be effective, it is necessary to look at how people work together as well as what they do. Traditionally, teams used to be built in hierarchies with a manager at the top. With the help of new technologies (such as issue tracking systems, agile boards, etc.), we are witnessing a paradigm shift in how people organise their work. Even developers, especially in open source environments, are starting to self-organise beyond the hierarchy, working more in line with collaborative approach of flatter structures. The shift away from a hierarchy sees collections of people as similar to living systems where it is necessary to consider the whole of the individual parts and relationships together. Ren et al. studied the effects of tenure disparity and interest variety on 648 WikiProjects finding that tenure disparity had a curvilinear effect on productivity and withdrawal. Terrel et al. (2017) presented an investigation of gender bias in open source by studying how software developers respond to pull requests and proposed changes to a software project's code and documentation. Results showed that women's pull requests tend to be accepted more often than men's, yet women's acceptance rates are higher only when they are not identifiable as women. In the context of existing theories of gender in the workplace, plausible explanations include the presence of gender bias in open source, survivorship and self-selection bias and women being held to higher performance standards. Vasilescu et al. (2013, 2015) studied gender and tenure diversity in GitHub teams and found that they were significant predictors of productivity. Lin and Serebrenik (2016) evaluated the applicability of different gender guessing approaches on several datasets derived from Stack Overflow. Compared to Vasilescu et al. (2015), this study focused on diversity in gender and nationality and studied development teams based on issue collaboration rather than commit activity.

2.2 Affect, emotions, sentiment analysis and developer personality traits

A growing body of literature has investigated the importance and the influence of human and social aspects in software engineering and software development (Graziotin et al. 2014, 2015a,b, Müller and Fritz 2015, Ortu et al. 2016b) and how several factor, e.g., politeness (Destefanis et al. 2016, 2017) affect developers' productivity. Rigby and Hassan (2007) analyzed, using a psychometrically-based linguistic analysis tool the five big personality traits of software developers in the Apache httpd server mailing list. The authors found that two developers responsible for the major Apache releases had similar personalities and their personalities were different from other developers. Graziotin et al. (2014a) reported the results of an investigation into the relationship between the affective states and analytical problem-solving skills of software developers showing that "happy developers were better problem solvers in terms of their analytical abilities"; this stressed the importance of the psychological condition of software developers. In a recent study, Graziotin et al. (2017) conducted a survey for measuring developer happiness, demonstrating that software developers are not a happy population. The authors identified factors representing causes of unhappiness for developers and highlighted the importance of considering social factors (e.g., happiness) in studies related to both human and technical aspects of software engineering. In this study we study how politeness and sentiment were linked with gender and nationality diversity.

3 Background

In this section, we provide background and motivation about the framework we adopted in our analysis. Politeness, along with other affective metrics, i.e., sentiment and emotions, has been used in several studies in software engineering (Destefanis et al. 2016; Guzman et al. 2014; Murgia et al. 2014; Ortu et al. 2015a, b; 2016a; Pletea et al. 2014). In what follows, we briefly introduce politeness and the methodology used to infer GitHub development teams.

3.1 Politeness

Politeness is “the ability to make all the parties relaxed and comfortable with one another[1]”. Politeness is crucial in communication (and therefore in collaboration) as people tend to perceive linguistic markers of politeness as a form of respect and their use related to the power dynamics of social interaction. Danescu et al. (2013) performed an empirical study on Wikipedia and Stack Exchange to investigate the relationship between politeness and social dynamics. They showed how polite editors in Wikipedia were more likely to achieve higher status in the community, thus suggesting a positive association between politeness and positive social outcomes. Conversely, they showed how politeness negatively correlated with social status: once elected, the Wikipedia editors became less polite; similarly, high reputation users of Stack Exchange tended to exhibit a less polite linguistic behaviour. Burke and Kraut (2008) investigated the impact of politeness on community engagement, measured in terms of reply rates. They showed how politeness differently affected the reply rates based on the topic being discussed. In particular, rudeness appeared to be more effective in eliciting responses in political discussion, while a polite attitude attracted more contribution in technical groups where people were typically seeking help. This was consistent with the results of the study by Althoff et al. (2014) on altruistic requests in on line communities, which showed that gratitude positively correlated to success, i.e. to an increased probability of receiving help.

As far as software engineering is concerned, the adequate selection of appropriate communication devices is a crucial issue since it can significantly impact the outcome of communication (Pikkarainen et al. 2008). More specifically, recent studies have already started to investigate the role of politeness in the domain of collaborative software development (Ortu et al. 2015c). In particular, Ortu et al. (2015c) analyzed 14 open source projects on the Apache Issue Tracking in Jira from 2002 to 2013 to understand the impact of politeness on productivity and project attractiveness. They found that a higher level of politeness positively correlated with a lower fixing time, thus suggesting that politeness is positively associated to higher productivity. Furthermore, polite projects appeared more attractive for contributors. Based on this evidence, it is reasonable to assume that politeness can be used as a proxy to assess the quality/level of the communication in a team. Therefore, we decided to consider politeness as a dependent variable of our study.

3.2 Issue collaboration graph

GitHub’s issue tracking is notable because it is focuses on collaboration, is a good way of keeping track of tasks, enhancements and bugs for projects and is primarily designed for collaboration. An issue generally consists of several parts:

- A title and description indicating what the issue is about.
- Color-coded labels to help categorize and filter issues (similar to labels in email).
- A milestone which acts like a container for issues. This is useful for associating issues with specific features or project phases (e.g. Weekly Sprint 9/5-9/16 or Shipping 1.0).
- A reporter who originally opened the issue.
- An assignee responsible for working on the issue at any given time.
- Comments which allow members of the repository to provide feedback.

Developers can post their comments on an issue report to discuss and manage the issue resolution. To build a network graph based on issue comments, we therefore considered each developer as a node; when user A commented on an issue reported by user B, we considered it as an edge from developer A to developer B. Following this approach, we obtained a directed network graph which we called *an issue collaboration graph*. Figure 1 shows an example of issue collaboration graph, in which different colours are related to different teams of developers. The dimension of a node represents the amount of activity of those specific developers, e.g., larger nodes represent more active developers within a given team. We considered developers who performed at least one commit for the project. We then used Gephi (an interactive visualization and exploration tool (Bastian et al. 2009)) to analyze the obtained network. Since information about team members is not provided on GitHub (we know that all the developers in the network are involved in the same project, but we do not know how the workload is distributed among them), we ran a modularity algorithm, based on the algorithms developed by Blonde (2008) and Lambiotte (2008) to obtain the communities (teams in this specific case) present in the network. Blonde et al. (2008) proposed a method to extract the community structure of large networks. It is an heuristic method that is based on modularity optimization.

4 Experimental design

Figure 2 shows the general design of our experiment. We built a diversity dataset to infer gender and country of the GitHub users, extracted the issue collaboration graph (Section 3.2) and performed modularity analysis to obtain the teams shown in Fig. 1.

To maintain consistency throughout the analysis and to allow easy comparison of results in the three cases, we applied three logistic regression models using *dicotomic* variables. In two of the three cases, team Politeness and team Sentiment, the variables are binary by nature, as extracted by the commits. Team Productivity assumes a continuous range of values, but its statistical distribution is fat tail and it is possible to observe values orders of magnitudes larger than other values.

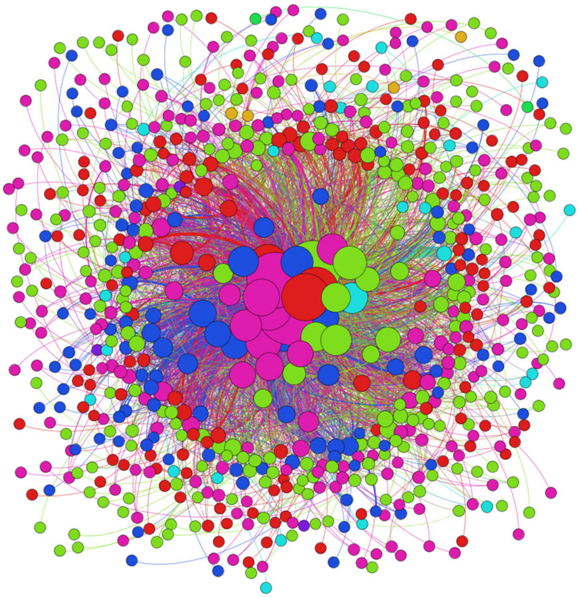

Fig. 1 Example of Issue Collaboration Graph as visualized by Gephi (Hadoop Common project)

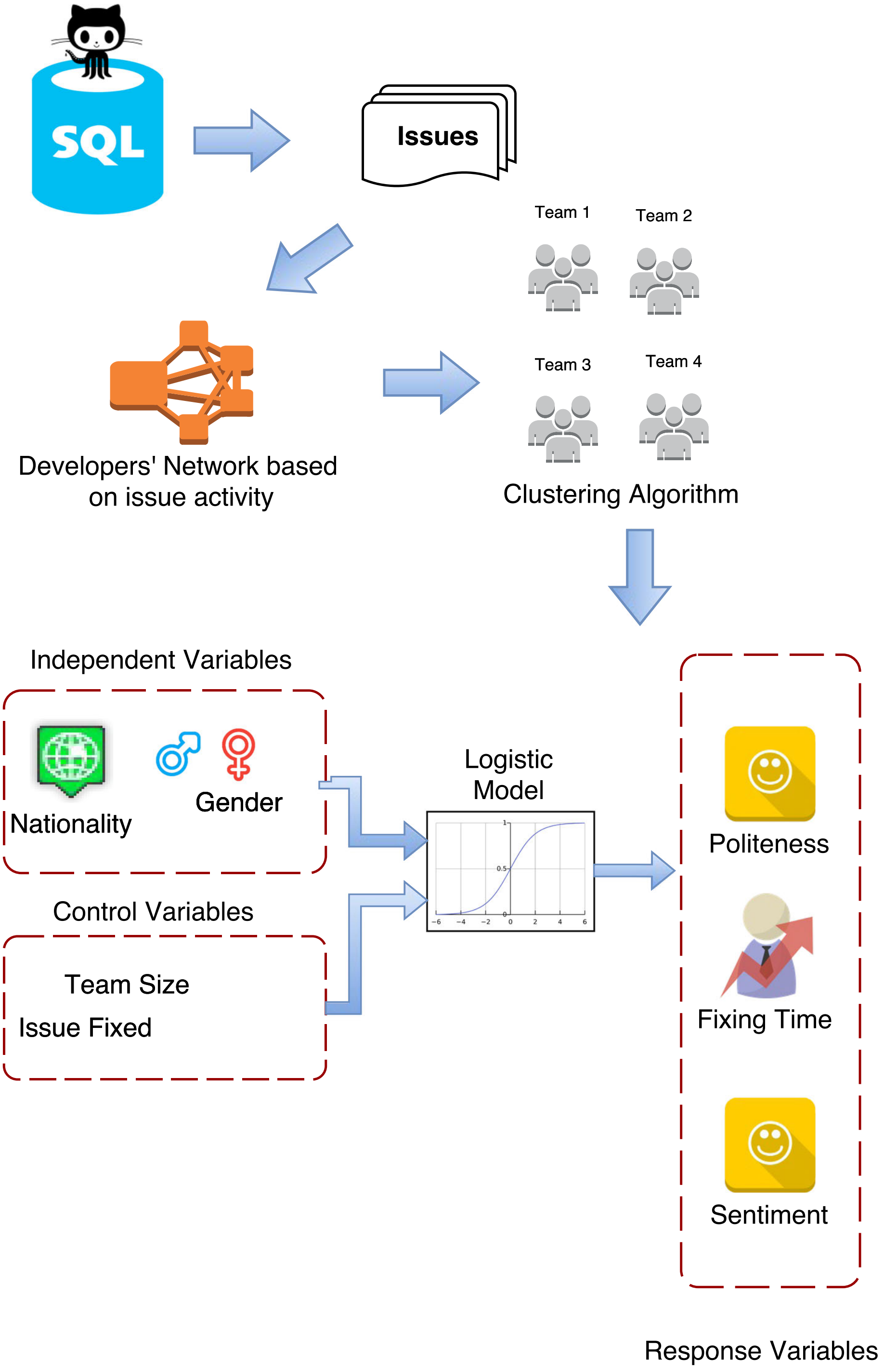

Fig. 2 Experimental design schema

For such distributions, it is difficult to select a theoretical curve best-fitting the empirical data among the various possibilities through all the range of values. Given this restriction and to maintain consistency with the other two cases, we decided to model team Productivity using the same model as that used for the other cases. We assume low team Productivity if the measured values are below the median (which is much less sensitive to outliers than the average) and high team productivity otherwise. This renders team Productivity as a binary variable. The approach has already been demonstrated successfully in Murgia et al. (2014) and Zhang et al. (2013).

4.1 Response variables

4.1.1 Team productivity

We measured team productivity considering the average Issue Fixing Time (IFT) of a team. We used a dichotomic variable for the IFT, since by analyzing the distribution of IFTs we found that in line with previous studies (Zhang et al. 2013), they are distributed as a long-tail distribution (power law). Variables with a logarithmic distribution do not fit well with linear models due to their large variations; this led us to use a logistic regression model, since it fit better with logarithmic distributions. Logistic regression models map input to a binary output. We thus considered the team average IFT to be 0 or 1, meaning that the average IFT was higher or lower than the IFT median, respectively. Furthermore, the use of a logistic regression model for IFT allows us to compare the model for productivity with that of politeness and sentiment (also logistic models). We decided to measure sentiment and politeness because they are indirect measures of the team communication level and in previous studies we found positive links among sentiment, politeness and productivity and between politeness and attractiveness of a project (Destefanis et al. 2016, Ortu et al. 2015b).

4.1.2 Team politeness

Team politeness is measured as the ratio of polite to impolite comments posted on issues resolved by a team. To obtain a discrete variable, we used a logistic regression model and considered a politeness of 1 if the ratio was greater than 1 (more polite comment) and a politeness of 0 if the polite/impolite ratio less than 1 (namely there were more impolite comments).

4.1.3 Team sentiment

Team sentiment is measured as average sentiment of comments posted on issues resolved by a team. To obtain a discrete variable, we used a logistic regression model and considered a sentiment of 1 if it was greater than the median (more positive comments) and a sentiment of 0 if the sentiment was less than the median (namely there were more positive comments).

4.2 Independent variables

4.2.1 Gender diversity

We measured gender diversity within a team using Blau's Diversity Index (Blau 1977), $1 - \sum_1^N p_i^2$, p is a categorical variable male/female. We only considered developers where we could infer gender from the evaluation of the gender diversity index.

4.2.2 Nationality diversity

Country diversity within a team is measured using Blau's Diversity Index (Blau 1977), $1 - \sum_1^N p_i^2$, p in this case is a categorical variable indicating the country (i.e. Italy, UK, etc); only developers where we could infer the country in the evaluation of the nationality diversity index were considered.

4.3 Control variables

4.3.1 Team size

Other studies have suggested that the total number of developers involved in an issue resolution is linked with longer IFTs and less polite comments (Ortu et al. 2015a; b) we thus considered the total number of developers in a team as a control variable for our experiment.

4.3.2 Team tenure

Team tenure is a measure of the number of issues resolved by a given team. It was considered as a control variable since developer experience is linked with both IFT and politeness (Ortu et al. 2015a, b). Average IFT plays a dual role in our models. It is considered as a response variable in the productivity model and as control variable in the politeness model (a longer IFT is linked with a lower politeness (Ortu et al. 2015a)).

5 Methods

5.1 Dataset

We used the last 2015 dump of the GHTorrent dataset (Gousios 2013). The dataset is a mirror of GitHub's data packed as a MySQL database. In our study, we are interested in modelling a developer network based on collaboration on issue resolution. For this purpose, only closed issues with a fixed resolution and at least 2 comments posted by different developers (including the issue reporter) were used. (We note that closed issues were considered as closed because they were the result of merged pull requests. The branch of the pull request had been merged with the main branch and, as a result, those issues were considered closed.) Finally, we considered 33,673 issues, with 71,423 comments posted by 13,872 developers belonging to 1176 different teams as shown in Table 1.

5.2 Measuring the control variables

5.2.1 Inferring gender

Following the approach and tool provided by Vasilescu et al. (2013) we inferred the GitHub developer's gender. Combining heuristics with female/male frequency name lists collected and the country the developer belong to, this tool is able to infer gender. Country data is crucial for inferring a developer's gender from their name (e.i. Andrea is a common male name in Italy, but a common female name in many other countries). The reported precision of gender is 93% (Vasilescu et al. 2013). In order to infer a developer's gender, it is necessary to know their name and country.

Table 1 Dataset statistics

Statistics	Value
Projects analyzed	8040
# of Developers	13,872
# of Issues	33,673
# of Comments	71,423
# of Communities	1176
Average of team size	118
Average # of fixed issues per team	278

5.2.2 Inferring user name and country

The simple task of retrieving the first name of a GitHub developer is challenging due to the noise and lack of information in the dataset (Kalliamvakou et al. 2014). Approximately 20% of all developers were labeled as "unknown" in the GHTorrent dump analyzed. Approximately 9% of all developers had at least two accounts. Thus, we grouped together all developer accounts with a set of heuristics proposed by Bird et al. (2006), i.e. accounts with same email were merged. We inferred developer nationality by applying a set of heuristics[2] based on the *location* available data; we were able to determine name and country of about 31% of all developers, a percentage close to that obtained in similar studies (Vasilescu et al. 2013; Vasilescu et al. 2015).

5.3 Measuring the dependent variables

5.3.1 Inferring issue fixing time

Figure 3 shows the typical issue timeline in GitHub:

- T_{cr} represents the time when an issue is created.
- T_{cl} represents the time when an issue is closed.
- T_a represents the time when an issue is assigned to a developer.
- T_s is the time when a developer subscribes to an issue that has been assigned to them.
- T_m represents the time when an issue is merged with the repository, namely the local commit is merged with the remote repository.

To infer the IFT, we used the approach proposed by Murgia et al. (2014). We computed the time interval between the last time an issue had been merged and the last time it had been subscribed to by an assignee (issues and pull requests are dual on Github; for each opened pull request, an issue is opened automatically (Gousios 2013)). The *closed issues* are considered as such because they are *merged pull request*. Since each pull request is an issue, and we considered the *merged* ones (the branch of the pull request has been merged with the main branch, therefore issues are closed). In the case that we could not obtain such dates, we used a conservative approach (i.e., if the subscribed date was missing we used assigned data and so on).

5.3.2 Measuring politeness

To compute the politeness of the contribution in our dataset, we adopted the library developed by (Danescu-Niculescu-Mizil et al. 2013). Given an input text, the tool calculates its overall politeness in terms of a discrete variable, i.e. *polite* or *impolite*.

The tool was trained and validated through machine learning on a gold standard of over 10,000 manually labelled requests from Wikipedia[3] and Stack Overflow[4]. The gold standard was built so as to include comments written by authors from all over the world while the annotators were selected among U.S. residents, based on a linguistic background

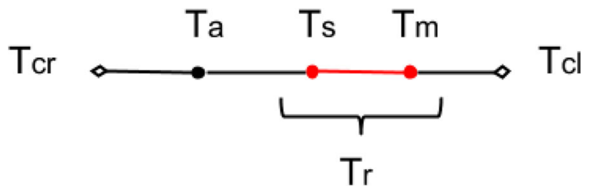

Fig. 3 Example of timeline for GitHub issue

questionnaire. The classifiers were evaluated both in an in-domain setting, with a standard leave-one-out cross validation procedure and in a cross-domain setting, where they were trained on one domain and tested on the other (Danescu-Niculescu-Mizil et al. 2013), achieving accuracy of 78.19% and 75.43%, respectively. Based on this evidence, we considered the tool by Danescu et al. (2013) robust enough to be adopted in our domain, i.e., GitHub issues, where developers post and discuss about technical aspects of issues.

5.3.3 Measuring sentiment

To compute the sentiment of the contribution in our dataset we used Sentistrength (Thelwall et al. 2012). Sentistrength is a keyword based tool, given a short text it assigns a score to each word based on the emotional content conveyed by the word. Sentistrength reports two sentiment strengths, negative which ranges from *-1 (not negative)* to *-5 (extremely negative)* and positive which ranges from *1 (not positive)* to *5 (extremely positive)*. Finally, we calculate the sentiment as the average of these two sentiment strengths.

6 Results

To conduct our analysis, we applied a logistic regression for estimating the extent to which gender and nationality diversity influenced productivity and level of communication in a team. We used logistic regression for its ease of interpretation, since it allowed us to reason about the significance of one factor given all the others. The results of the logistic regression are reported in Tables 2 and 3. The tables list the results for each of the predictors in our framework for the independent variables (namely, nationality and gender) and the control variables, namely team size and issue and the team tenure defined in our framework, grouped by actionable factor. For each predictor, logistic regression outputs three values, namely, coefficient estimate, odds ratio and statistical significance. The sign of the coefficient estimate indicates the positive/negative impact of the predictor on the success of a question. The odds ratio weighs the magnitude of this impact: the closer the value is to 1, the smaller the impact of the parameter on the chance of success. In particular, an odds ratio value lower than 1 corresponds to a negative impact of the predictor and vice versa. Finally, the statistical significance determines whether a predictor has a significant explanatory value. For each of the three research questions, using the methodology described in Section 5, we assembled a longitudinal dataset of GitHub teams based on the issue collaboration graph.

6.1 RQ1: Are gender or nationality diversity linked to the issue fixing time of a team?

6.1.1 Motivation

When building a team, one of the first goals is a high level of effectiveness. While there are many ways for measuring effectiveness of a working team, i.e, number of activities solved

Table 2 Issue fixing time model

Coefficients	Estimate	*z* value	Odds ratio	Pr(> |*z*|)
(Intercept)	0.61	0.604	1.83	0.54
Team tenure	-0.0001	-0.882	0.9	0.37
Team size	0.0007	1.195	1	0.23
Gender div	-2.72	-2.709	0.06	**
Country div	0.005	0.005	1	0.99

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 of Pr(> |*z*|)

Table 3 Politeness model

Coefficients	Estimate	*z* value	Odd Ratio	Pr(> |*z*|)
(Intercept)	4.46	0.914	0.15	0.36
Team tenure	-1.35e-3	-1.451	0.9	0.14
Team size	2.3e-3	1.261	1	0.20
Gender div	-5.57	-0.579	5e-3	0.56
Country div	-12.48	-2.074	1	*
Fixing time	5.15e-7	1.774	1	0.12

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 of Pr(> |*z*|)

in a time interval or the total time to complete all activities, we considered a measure of productivity as the average time required to resolve an issue. Knowing which factor affects this measure is crucial for a successful team.

6.1.2 *Approach*

We modelled team productivity as the average IFT and considered gender and country diversity as team characteristics controlled by the team size and number of issues solved.

6.1.3 *Findings*

Gender diversity in GitHub teams was linked with lower IFT.

In our model, we modelled average IFT as a binary variable, where 1 means a value higher than the median (considering all issues) and 0 means lower. The model showed that gender diversity was the most dominant metric explaining data variance and had a negative effect on productivity (tendency to lower issue fixing time, Table 2). The first column represent the logistic regression coefficient, the second column represents the *z*-value and the third column the odds ratio of the metric. For the odds ratio, a value below 1 means that the metric is linked with lower productivity, a value above 1 means that the metric is linked with higher productivity and a value of 1 means no effect on productivity. The final column represent the significance of the metric in the model; here, we used the R language convention where significant variables with a *p*-value lower than 0.001 are indicated with three stars *** (see the Table 2 caption). Country diversity, along with the other control metrics, were not significant. Our finding matches common sense. It is expected that diversity brings novelty, unique experience and different perspectives which can explain a rise in productivity.

6.2 RQ2: Are gender or nationality diversity linked to the overall politeness of a team?

6.2.1 *Motivation*

Communication within a team is as important as productivity, since a poor level of communication can lead to unproductive and unsuccessful teams. Measuring the overall sentiment and politeness expressed by a development team while discussing an issue resolution, we have been able to obtain useful insights into which factors might affect the overall team communication.

6.2.2 *Approach*

We considered gender and country diversity as team characteristics controlled by the team size, number of issues solved and average IFT.

6.2.3 Findings

Country diversity is linked with lower politeness.

We modelled team politeness as a binary variable. A value of 1 means that the ratio polite/impolite comments is greater than 1 (more polite comments) and 0 means less than 1 (more impolite comments). Table 3 shows the results for politeness. This table is presented in the same way as explained in 6.1. The model shows that country diversity is the most dominant metric explaining data variance and has a negative effect on team politeness (it has a tendency towards lower politeness). Gender diversity, along with the other control metrics, is not significant in our model. Lower politeness within a team of "diverse" developers in term of nationality could be related to habits and customs co-existing in the same environment. We analysed teams of developers of open source projects; hence, people were often working remotely and communicating through emails, chat, short messages exchanged on issue tracking systems. The open source paradigm brings constraints due to lack of shared physical spaces and personal bonds.Different language backgrounds may be a prolific source of misunderstanding and misinterpretation of written requests or comments, leading to a general lower level of politeness.

6.3 RQ3: Are gender - nationality diversity linked to the overall sentiment of a team?

6.3.1 Approach

We considered gender and country diversity as team characteristics controlled by the team size, number of issues solved and average IFT.

In this case, we considered the overall sentiment of an issue as described in 5.3.3. SentiStrength given a text, returns two strenghts: positive from 1 to 5 and negative from -1 to -5. We then sum up those two strengths for each issue's comment and average them to obtain the issue's overall sentiment.

6.3.2 Findings

Gender diversity is linked with higher sentiment.

We modelled team sentiment as a binary variable. A value of 1 means that the overall sentiment of comments is greater than the median. Table 3 shows the results for politeness (presented in the same way explained in 6.1 and 6.2). Gender diversity is the most dominant metric explaining data variance and has a positive effect on team sentiment (it has a tendency to increase sentiment). Nationality diversity, along with the other control metrics, is not significant in our model. Higher sentiment within a team of "diverse" developers in term of gender could be related to interactions in the same environment. We analysed teams of developers of open source projects, hence, people working were often remotely and communicating through emails, chat, short messages exchanged on issue tracking systems.

7 Discussion

In this paper, we presented an analysis about the links between a team's diversity in terms of nationality and gender and the overall team's politeness, sentiment and productivity. Politeness is a factor which certainly helps in diminishing conflict and friction between people as well as Sentiment. The findings in this study contribute in highlighting the importance and the impact of two factors that needs to be considered when building a team: gender and nationality diversity.

Firstly, we studied the influence of nationality and gender diversity on productivity and showed that gender diversity was the most dominant metric explaining data variance and had a positive (i.e., downward) effect on productivity (tendency to lower issue fixing time). Country diversity, along with the other control metrics, were not significant. Although more sophisticated models are needed to confirm our findings, they match common sense. It is expected that diversity brings novelty, unique experience and different perspectives which may explain a rise in terms of productivity. This result related specifically to gender suggests that a successful team is one where there is a combination of genders and skills, complementing each other towards a common goal. In many countries, it is a challenge to recruit female students to Computing-related courses. If this situation persists, then clearly the implication is that productivity will be lower than it could (and should) have been.The research sends out a clear message to development companies, but more so to Universities and Colleges that more perhaps needs to be done for initiatives on Women in IT.

Secondly, we studied the influence of nationality and gender diversity on politeness. Results show that country diversity is the most dominant metric explaining data variance and has a negative effect on team politeness (it has a tendency to lower politeness). Gender diversity, along with the other control metrics, is not significant in our model. Lower politeness within a team of “diverse” developers in term of nationality could be related to habits and customs co-existing in the same environment. The open source paradigm brings constraints due to lack of shared physical spaces and personal bonds. Different language backgrounds may be a prolific source of misunderstanding and misinterpretation of written requests or comments, leading to a general lower level of politeness. This result emphasises the need for all development staff to appreciate the customs of different cultures and languages and that effective communication (through all different mediums) is an essential, critical part of a team’s effectiveness and well-being.

Thirdly, we studied the influence of nationality and gender diversity on sentiment and found that gender diversity is the most dominant metric explaining data variance with a positive effect on team sentiment (it has a tendency to increase sentiment). Nationality diversity, along with the other control metrics, is not significant in our model. Higher sentiment within a team of “diverse” developers in term of gender could be related to interactions in the same environment.

All of the authors have worked in the IT industry either in the past or are working in industry in spinout industry-academic companies currently. The issue of gender in the work-place is a highly important, emotive and relevant topic which is significantly under-researched, but which as our study has shown, can demonstrate significant benefits in terms of our knowledge. Discrimination in the work-place and eliminating it can be helped, albeit in a small way, through concrete empirical evidence such as that presented in the study. (Part motivation for this study was to uncover in a positive way the value of various team dynamics and composition.) The same is true of nationality issues. Being respectful, polite and understanding the opinions, culture and feelings of other people is a crucial part of a) working in the IT industry (or indeed any industry), b) completing project tasks, c) communicating properly and d) creating a harmonious and productive work environment for everyone.We see this as just one motivation for our choice of those variables in this study. The results presented in this study can be also helpful when defining a team of developers.

Knowing the profile (e.g., "average politeness/sentiment" of a developer) of the developers, their nationality and gender could be useful in providing advice for creating balanced teams. A GitHub plug-in which is able to match team members based on heuristics from general politeness, sentiment along with other characteristics such as gender and diversity could be useful to developers and managers alike in building better teams for a company or for a project (in the case of open source collaboration paradigm). The results have one final, but important implication. This is that computing as a discipline has as much to gain through study of social science (e.g., psychology, sociology etc) as it does from the mathematical sciences. This is often forgotten; the study presented shines a light on that lesson.

8 Threats to validity

Threats to *internal validity* concern confounding factors that can influence the obtained results. We assume a causal relationship between a developer's emotional state and what they write in issue report comments, based on empirical evidence (in another domain). We built an explanatory model to understand the characteristics of development teams considering productivity and collaboration.

Threats to *construct validity* focus on how accurately the observations describe the phenomena of interest. We used state-of-the-art tools (Danescu-Niculescu-Mizil et al. 2013) provided by Danescu et al. to measure politeness, in addition to heuristics for evaluating issue fixing time.

Threats to *external validity* correspond to the generalisability of our experimental results. We consider issues as a representative sample of the universe of open source software projects, with different development teams and satisfying different customers' needs. We accept that our definition of an issue does not take into consideration the possibility that not all issues have commits; some issues do not have commits, since they are not pull requests. However, it is our belief that this only represents a small part of the dataset used. Other threats concern the validity of the models used for our analysis. In particular, the models presented are probably not complex enough to properly represent the measured phenomena (as suggested by the low significance of *control variables* in both experiments). Replications of this work on other open source systems and on commercial projects are needed to confirm our findings.

9 Conclusions

In this paper, we presented three logistic regression models representing the average time required to solve an issue by development teams and the overall communication level measured by the overall politeness of a team. Results showed that gender diversity is linked with lower average issue fixing time and nationality diversity is linked with lower team politeness. We used a set of heuristics to infer development teams based on GitHub's issues collaboration graph, GitHub's user gender and nationality with the final goal of building a representative diversity dataset. Our results also indicated that both gender and nationality diversity played a significant role when considering the productivity and collaboration within a team.

The results have several implications for the way that developers of OS and project managers in closed-source projects should view team make-up. Firstly, for the analysis presented, if a team brings together a complementary set of skills and behaviour, putting people's welfare at the heart of the team ethos, then it stands a good chance of being successful. This implies being aware of the pressures on each team member and understanding difficulties and conflicts when they arise. Secondly, it is becoming increasingly apparent that while a single study does not solve all the potential ills of team co-ordination, there is clearly a case for social science disciplines to inform more on how modern day IT projects perform.Perhaps even, rather controversially, social scientists and anthropologists should be be part of a project team, advising on the decisions made and processes followed. Finally, we are seeing that, through technology, the world is becoming a smaller workplace at a rapid rate. Software development across continents will only continue to grow in scale. Understanding different cultures and customs and accommodating those into the process will become essential. The study presented informs that understanding.

Further research on different datasets is needed to validate and extend our findings. For example, considering other factors such as developers experience, technical background and the level of acceptance of newcomers are all promising avenues. We believe that these results are a good starting point for stimulating more research activities toward this direction and could help managers and team directors in taking better decisions during the crucial starting phase when building a team of developers.

Endnotes

[1] http://en.wikipedia.org/wiki/Politeness

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

[3] https://en.wikipedia.org/wiki/Main_Page

[4] http://stackoverflow.com

Funding
The research presented in this paper was partly funded by the Engineering and Physical Sciences Research Council (EPSRC) of the UK under grant ref: EP/M024083/1.

Availability of data and materials
We used a dataset extracted from the 2014 dump of the GHTorrent dataset (Gousios 2013).

Authors' contributions
MO, GD and SC conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and tables, performed the computation work, reviewed drafts of the paper. SS performed the experiments, analyzed the data, reviewed drafts of the paper. MM and RT reviewed drafts of the paper. All authors read and approved the final manuscript.

Competing interests
There are no competing interests for any of the authors.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]DIEE, University of Cagliari, Cagliari, Italy. [2]University of Hertfordshire, Hatfield, UK. [3]Brunel University, Uxbridge, UK.

Received: 19 May 2017 Accepted: 9 November 2017
Published online: 20 December 2017

References

Allen NJ, Hecht TD (2004) The romance of teams: Toward an understanding of its psychological underpinnings and implications. J Occup Organ Psychol 77(4):439–461

Althoff T, Danescu-Niculescu-Mizil C, Jurafsky D (2014) How to ask for a favor: A case study on the success of altruistic requests. In: Proceedings of ICWSM. ICWSM '14. https://web.stanford.edu/~jurafsky/pubs/icwsm2014_pizza.pdf

Bastian M, Heymann S, Jacomy M, et al. (2009) Gephi: an open source software for exploring and manipulating networks. ICWSM 8:361–362

Bird C, Gourley A, Devanbu P, Gertz M, Swaminathan A (2006) Mining email social networks. In: Proceedings of the 2006 international workshop on Mining software repositories. ACM. pp 137–143. https://dl.acm.org/citation.cfm?id=1138016

Blau PM (1977) Inequality and heterogeneity: A primitive theory of social structure, vol 7. Free Press New York. https://books.google.com.ph/books/about/Inequality_and_Heterogeneity.html?id=jvq2AAAAIAAJ&redir_esc=y

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008

Blum L, Frieze C, Hazzan O, Dias D (2007) A Cultural Perspective on Gender Diversity in Computing. In: Burger CJ, Creamer EG, Meszaros PS (eds). Reconfiguring the Firewall: Recruiting Women to Information Technology across Cultures and Continents. AK Peters, Ltd.

Burke M, Kraut R (2008) Mind your ps and qs: The impact of politeness and rudeness in online communities. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW '08. ACM, New York. pp 281–284

Chen J, Ren Y, Riedl J (2010) The effects of diversity on group productivity and member withdrawal in online volunteer groups. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM. pp 821–830

Cheryan S, Plaut VC, Davies PG, Steele CM (2009) Ambient belonging: how stereotypical cues impact gender participation in computer science. J Personal Soc Psychol 97(6):1045

Danescu-Niculescu-Mizil C, Sudhof M, Jurafsky D, Leskovec J, Potts C (2013) A computational approach to politeness with application to social factors. In: Proceedings of ACL

Daniel S, Agarwal R, Stewart KJ (2013) The effects of diversity in global, distributed collectives: A study of open source project success. Inf Syst Res 24(2):312–333

Destefanis G, Ortu M, Counsell S, Swift S, Marchesi M, Tonelli R (2016) Software development: do good manners matter? PeerJ Comput Sci 2:e73

Destefanis G, Ortu M, Counsell S, Swift S, Tonelli R, Marchesi M (2017) On the randomness and seasonality of affective metrics for software development. In: Proceedings of the Symposium on Applied Computing. ACM. pp 1266–1271

Earley CP, Mosakowski E (2000) Creating hybrid team cultures: An empirical test of transnational team functioning. Acad Manag J 43(1):26–49

Giuri P, Ploner M, Rullani F, Torrisi S (2010) Skills, division of labor and performance in collective inventions: Evidence from open source software. Int J Ind Organ 28(1):54–68

Gousios G (2013) The ghtorrent dataset and tool suite. In: Proceedings of the 10th Working Conference on Mining Software Repositories, MSR '13. IEEE Press, Piscataway. pp 233–236

Graziotin D, Fagerholm F, Wang X, Abrahamsson P (2017) On the unhappiness of software developers. In: Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering. ACM. pp 324–333

Graziotin D, Wang X, Abrahamsson P (2014) Happy software developers solve problems better: psychological measurements in empirical software engineering. PeerJ 2:e289

Graziotin D, Wang X, Abrahamsson P (2014) Software developers, moods, emotions, and performance. IEEE software 31(4):24–27

Graziotin, D, Wang X, Abrahamsson P (2015) Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering. J Softw Evol Process 27(7):467–487

Graziotin D, Wang X, Abrahamsson P (2015) How do you feel, developer? An explanatory theory of the impact of affects on programming performance. PeerJ Comput Sci 1:e18

Guzman E, Azócar D, Li Y (2014) Sentiment analysis of commit comments in github: an empirical study. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM. pp 352–355

Herring C (2009) Does diversity pay?: Race, gender, and the business case for diversity. Am Sociol Rev 74(2):208–224

Horwitz SK, Horwitz IB (2007) The effects of team diversity on team outcomes: A meta-analytic review of team demography. J Manag 33(6):987–1015

Kalliamvakou E, Gousios G, Blincoe K, Singer L, German DM, Damian D (2014) The promises and perils of mining github. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM. pp 92–101

Lambiotte R, Delvenne J-C, Barahona M (2008) Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770

Lin B, Serebrenik A (2016) Recognizing gender of stack overflow users. In: Proceedings of the 13th International Workshop on Mining Software Repositories. ACM. pp 425–429

Mäntylä M, Adams B, Destefanis G, Graziotin D, Ortu M (2016) Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity? In: Proceedings of the 13th International Workshop on Mining Software Repositories. ACM. pp 247–258

Medel P, Pournaghshband V (2017) Eliminating gender bias in computer science education materials. In: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education. ACM. pp 411–416

Müller SC, Fritz T (2015) Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In: Software Engineering (ICSE) 2015 IEEE/ACM 37th IEEE International Conference on, vol 1. IEEE. pp 688–699

Murgia A, Concas G, Tonelli R, Ortu M, Demeyer S, Marchesi M (2014) On the influence of maintenance activity types on the issue resolution time. In: Proceedings of the 10th International Conference on Predictive Models in Software Engineering. ACM. pp 12–21

Murgia A, Tourani P, Adams B, Ortu M (2014) Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM. pp 262–271

Ortu M, Adams B, Destefanis G, Tourani P, Marchesi M, Tonelli R (2015a) Are bullies more productive? Empirical study of affectiveness vs. issue fixing time. In: Proceedings of the 12th Working Conference on Mining Software Repositories. MSR 2015

Ortu M, Destefanis G, Kassab M, Counsell S, Marchesi M, Tonelli R (2015b) Would you mind fixing this issue? an empirical analysis of politeness and attractiveness in software developed using agile boards. In: Agile Processes, in Software Engineering, and Extreme Programming. Springer. pp 129–140

Ortu M, Destefanis G, Kassab M, Marchesi M (2015c) Measuring and understanding the effectiveness of jira developers communities. In: Proceedings of the Sixth International Workshop on Emerging Trends in Software Metrics. IEEE Press. pp 3–10

Ortu M, Murgia A, Destefanis G, Tourani P, Tonelli R, Marchesi M, Adams B (2016a) The emotional side of software developers in JIRA. In: Mining Software Repositories (MSR), 2016 IEEE/ACM 13th Working Conference on. IEEE. pp 480–483

Ortu M, Murgia A, Destefanis G, Tourani P, Tonelli R, Marchesi M, Adams B (2016b) The emotional side of software developers in jira

Pikkarainen M, Haikara J, Salo O, Abrahamsson P, Still J (2008) The impact of agile practices on communication in software development. Empirical Softw Engg 13(3):303–337

Pletea D, Vasilescu B, Serebrenik A (2014) Security and emotion: sentiment analysis of security discussions on github. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM. pp 348–351

Ren Y, Chen J, Riedl J (2015) The impact and evolution of group diversity in online open collaboration. Manag Sci 62(6):1668–1686

Rigby PC, Hassan AE (2007) What can OSS mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list. In: Proceedings of the Fourth International Workshop on Mining Software Repositories. IEEE Computer Society. p 23

Sheldon JP (2004) Gender stereotypes in educational software for young children. Sex Roles 51(7-8):433–444

Stahl GK, Maznevski ML, Voigt A, Jonsen K (2010) Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. J Int Bus Stud 41(4):690–709

Terrell J, Kofink A, Middleton J, Rainear C, Murphy-Hill E, Parnin C, Stallings J (2017) Gender differences and bias in open source: pull request acceptance of women versus men. PeerJ Comput Sci 3:e111

Thelwall M, Buckley K, Paltoglou G (2012) Sentiment strength detection for the social web. J Am Soc Inf Sci Technol 63(1):163–173

Vasilescu B, Capiluppi A, Serebrenik A (2013) Gender, representation and online participation: A quantitative study. Interact Comput:iwt047

Vasilescu B, Posnett D, Ray B, van den Brand MG, Serebrenik A, Devanbu P, Filkov V (2015) Gender and tenure diversity in github teams. In: CHI. ACM

Watson WE, Kumar K, Michaelsen LK (1993) Cultural diversity’s impact on interaction process and performance: Comparing homogeneous and diverse task groups. Acad Manag J 36(3):590–602

Williams KY, O’Reilly CA (1998) Demography and diversity in organizations: A review of 40 years of research. Res Organ Behav 20:77–140

Zhang H, Gong L, Versteeg S (2013) Predicting bug-fixing time: an empirical study of commercial software projects. In: Proceedings of the 2013 international conference on software engineering. IEEE Press. pp 1042–1051