Relationship between Geographical Location and Evaluation of Developer Contributions in GitHub

Rastogi, Ayushi; Nagappan, Nachiappan; Gousios, Georgios; van der Hoek, André

DOI
10.1145/3239235.3240504

Publication date
2018

Document Version
Accepted author manuscript

Published in
ESEM '18

Relationship between Geographical Location and Evaluation of Developer Contributions in GitHub

Ayushi Rastogi
UCI, Irvine
ayushir@ics.uci.edu

Nachiappan Nagappan
Microsoft Research, Redmond
nachin@microsoft.com

Georgios Gousios
TU Delft, Netherlands
g.gousios@tudelft.nl

André van der Hoek
UCI, Irvine
andre@ics.uci.edu

ABSTRACT

Background Open source software projects show gender bias suggesting that other demographic characteristics of developers, like geographical location, can negatively influence evaluation of contributions too. **Aim** This study contributes to this emerging body of knowledge in software development by presenting a quantitative analysis of the relationship between the geographical location of developers and evaluation of their contributions on GitHub. **Method** We present an analysis of 70,000+ pull requests selected from 17 most actively participating countries to model the relationship between the geographical location of developers and pull request acceptance decision. **Results and Conclusion** We observed structural differences in pull request acceptance rates across 17 countries. Countries with no apparent similarities such as Switzerland and Japan had one of the highest pull request acceptance rates while countries like China and Germany had one of the lowest pull request acceptance rates. Notably, higher acceptance rates were observed for all but one country when pull requests were evaluated by developers from the same country.

CCS CONCEPTS

• **Software and its engineering** → **Open source model**; **Programming teams**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

KEYWORDS

Open source, geographical location, pull requests, GitHub.

ACM Reference Format:
Ayushi Rastogi, Nachiappan Nagappan, Georgios Gousios, and André van der Hoek. 2018. Relationship between Geographical Location and Evaluation of Developer Contributions in GitHub. In *ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '18), October 11–12, 2018, Oulu, Finland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3239235.3240504

ESEM '18, October 11–12, 2018, Oulu, Finland

ACM ISBN 978-1-4503-5823-1/18/10…$15.00
https://doi.org/10.1145/3239235.3240504

1 INTRODUCTION

Open source software (OSS) development has always been envisioned as a merit-based model [23]. This gave rise to the term 'code is king' [10][23], indicating a belief that the quality of the code being contributed should be the sole factor in determining whether or not the code is accepted to be included in the primary line of development.

Various studies, however, have shown that in addition to technical factors relating to code quality, social factors influence acceptance or rejection decisions. For instance, social closeness between submitter and integrator, as built through prior interactions, can positively influence acceptance [26]. As another example, status in the community increases acceptance [26]), while size of the contribution decreases acceptance [14].

Only recently, a demographic attribute of contributors – gender, is found to influence evaluation of contributions in OSS projects. In a study of 1.4 million GitHub user profiles, Terrell et al. found that code contributions by female developers were less often accepted than their male counterparts when their gender was identifiable [25].

This paper contributes to this emerging body of knowledge in software development with a study that focuses on a demographic attribute of developers - geographical location. We elicit the relationship between the geographical location of developers and evaluation of their contributions by modeling GitHub projects' archival data using country of residence of developers to measure geographical location and pull request acceptance decision to measure evaluation of contributions.

We present an analysis of 70,000+ pull requests originating from 17 most actively participating countries on GitHub. We control for the influence of other factors known to influence pull request acceptance decision to quantitatively analyze the relationship of the country of submitters with overall pull request acceptance decision. We also consider the case in which submitter and integrator are from the same country for its possible relationship with pull request acceptance decision.

Our findings reveal that there are statistically significant differences in the acceptance rates among the pull requests issued by developers from different countries. We, however, could not attribute the observed differences to the factors which may seem obvious otherwise. For example, we examined whether contributors from non-English speaking countries have lower pull request acceptance rate compared to English speaking countries. We did not find any pattern though.

2 BACKGROUND

This section presents a brief history on bias at work originating from individuals' demographic attribute - geographical location, and how it translates to OSS projects. This is followed by relevant background material concerning the notion of pull-based development and factors that are already known to influence pull requests.

2.1 Bias at work originating from individuals' geographical location

The role of individuals' demographic attribute - geographical location - in influencing evaluation of contributions is known for long in traditional workplaces, where people meet in-person for work and are aware of the demographic attributes of fellow contributors. Indeed, a whole body of literature has emerged, albiet dispersed over many communities. For instance, in the Olympics and other forms of international competitions, experienced judges with significant training in fairly evaluating all participants were found to nonetheless rank participants from their own nations higher than participants from other nations [22]. As another example, in academia, papers with authors from some regions receive fewer citations than papers from authors of other regions, even when the papers were of comparable quality [18].

In open source software development, in contrast, developers collaborate online, from same or different geographical locations and were not aware of the demographic attributes of fellow developers. In recent years, with the rise of environments such as GitHub [3], Bitbucket [1], and others (e.g., [4][5]), OSS developers today can become much more aware of the demographic attributes of their fellow developers [27], which may lead to them consciously or subconsciously changing the way they interact with and make decisions toward their fellow developers [28].

A recent study of 1.4 million GitHub user profiles reported bias in evaluation of code contributions when a demographic attribute of contributors - gender, was known to fellow contributors [25]. The study by Terrell et al. found that code contributions by female developers were less often accepted than their male counterparts when their gender was identifiable [25]. The results of this study suggests that other demographic attributes of contributors like geographical location, which are known to a significant percentage of fellow contributors [27], may influence evaluation of contributions.

2.2 Pull-based development

GitHub, the site of our study, supports two models of collaborative development: (1) the shared repository model and (2) the pull-based development model. The shared repository model grants everyone access to directly make changes to a single, shared repository of code. Every change is therefore public immediately and not subject to a further level of review first. The use of the shared repository model is prevalent with small teams and organizations collaborating on private projects.[1]

The pull-based development model contrasts the shared repository model in separating the development effort of individual contributors (submitters) from the decision whether or not to include the code they submit for consideration [11]. A separate role, often called the integrator or committer, receives what is called a pull request when a contributor submits their proposed update. The integrator then 'pulls' the code from the repository of the contributor, examines it closely, and makes their decision as to whether to push the code to the main branch. This two-phased process allows projects to be more transparent, with open discussions taking place surrounding complex pull requests, thereby making the overall process more democractic [17]. Today, nearly half of the projects on GitHub use the pull-based development model [11]. Particularly because these projects tend to be larger and public, numerous previous studies have focused on various aspects and implications of the pull-based development model (e.g., Yu et al. examined factors influencing latency in pull request evaluation [31] and Gousios et al. studied aspects of work practices and challenges in pull-based development [12]).

2.3 Factors influencing pull request acceptance

Previous studies have already begun to look at pull request acceptance and the influence that different factors may have. For instance, it has been found that a developer's technical and social reputation positively influences acceptance of the changes they submit in their pull requests [7][14][19][26]. Some of these studies further demonstrated that adhering to a project's technical and social norms increases the chances of acceptance [14][16][26], although with the caveat that, the more mature and popular a project is, the lower the chances of pull request acceptance are overall [14][26]. Focusing on the nature of a pull request itself, it has also been shown that the size of the change, its perceived quality, and the theme and objective of the pull request, among others, influence its chances of acceptance considerably [14][26][30][24].

Broadly, the factors that have been found to seemingly influence pull request acceptance can be categorized as being related to: (1) the developer themselves, (2) the project as a whole, and (3) the specific pull request. Our study adds to what is known thus far by focusing on country of developers as having a potential impact. As we will show in the below, we control for the factors that previous studies have found, thereby being able to more singularly attribute whether the country in which a developer resides has an effect on the evaluation of their contributions.

3 METHOD

We provide a step-by-step introduction to how we collected and analyzed the relationship using GitHub projects' archival data. All data and procedures used are publicly available for replication [2].

3.1 GitHub projects' archival data

To analyze the influence of the country of developers on pull request acceptance decisions, we measure a variety of factors that: (1) have been previously identified as possibly influencing pull request acceptance, and (2) can be deduced based on stored archival data regarding past activities of developers (submitters and integrators). To these factors, we add the country in which developers reside when they submit pull requests, as well as the country of the integrators. Using the combined data, we prepare two statistical models, one focusing solely on the country of the submitter in relationship

[1] https://help.github.com/articles/about-collaborative-development-models/

[2] 10.6084/m9.figshare.6865799

to pull request acceptance, and one taking into account whether the submitter and integrator are in the same country.

3.1.1 Factors influencing pull request acceptance. As discussed in Section 2.3, factors that may influence pull request acceptance can be classified into being related to the developer themself, the project overall, or the pull request. For our study, we use a combination of the factors that were identified by Tsay et al. [26] and by Gousios et al. [14]. Table 1 presents the result, as organized by project characteristics first, then developer characteristics, and finally pull request characteristics. We add a fourth category, geographical location, which naturally belongs to developer characteristics, but is separated out since it is the focus of our study.

The individual columns of the table should be read as follows. The first column lists specific factors that may influence pull request acceptance. The second and third columns list the specific measures, as mined from the respective repositories they studied, through which Tsay et al. and Gousios et al. incorporated various factors in their model (note that they did not study the exact same set of factors). The fourth column lists the specific measures we choose to use in our models, with the fifth column listing the variable name we use in the remainder of the paper for each of the measures. As one example, social closeness was estimated by Tsay et al. using two measures: whether the developer followed the integrator prior to submitting their pull request and whether the developer followed the project repository prior to submitting. Gousios et al. did not measure this factor. In our model, we choose to measure it using both measures from Tsay et al., and use the names *dev_followed_integrator* and *dev_watched_project* for these two measures.

Where possible, we adopt the measures of Tsay et al. or Gousios et al., although we made an exception in three measures. We changed the measure of code quality from test LOC per 1,000 LOC (kloc) to test LOC per 100,000 lines of code (lloc). We do this to bring all factors with numeric values to a comparable scale of measurement. Similarly, we measure test cases per 1,000 LOC and asserts per 1,000 LOC as test cases per 100,000 LOC and asserts per 100,000 LOC respectively.

Note that our model omits a specific measure for team size, as well as two measures for code quality (*#test cases per kloc* and *#asserts per kloc*). As we discuss in the below, these three measures contain information that is included in others, and hence we leave them out of further consideration.

Note also that our model does not include gender, a factor that was recently found to correlate with pull request acceptance (specifically, pull requests by female developers were more often rejected) [25]. We were unable to find a reliable way to obtain the gender for all 70,740 pull requests we include in our study (see below). If we only included those pull requests for which we could obtain gender data, our sample would have been too small.

3.1.2 Data collection. The basis for our analysis is the data made publicly available by Gousios et al. [13], covering 1,069 projects and 370,411 pull requests in Python (357), Java (315), Ruby (359), and Scala (38). Gousios et al. selected these projects to represent the top 1% of all GitHub projects in terms of their respective counts of pull requests at the time. Our sample is therefore not representative of all projects in GitHub, but this is intentionally so in order to focus on the more active projects that are more likely to include participants from across the globe. We enriched this data set by leveraging the GHTorrent dataset made available on August 18, 2015 [2] to add the country from which each individual pull request was made by a submitter and the country from which each individual pull request was accepted or rejected by an integrator.

To decide upon a country, we needed to apply some heuristics, as developers in GitHub can choose to specify their location in free-form text in their GitHub profiles. Developers who, for instance, write 'US', 'United States', or 'XYZ Apartments, New York' all are somewhere in the United States. To infer the country of developers from these free-form textual specifications, we use the 'countryNameManager' script used in a previous study by Vasilescu et al. [28], which uses heuristics to, for instance, map the three previous examples all to the United States. We augmented the script with one additional heuristic to address situations where developers did specify an affiliation and/or domain name but did not provide a location, so we can map, for instance, someone who specified an affiliation of 'Peking University' to China. To do so, our heuristic learns from those entries where the country is specified in addition to the affiliation and/or domain name and subsequently maps it to entries where the country is missing but the affiliation and/or domain name is the same. To minimize false positives, we only apply this heuristic if the affiliation and/or domain name map to a single country across all such entries, and if more than 20 entries exist. For example, we map 'Peking University' to China only if there are at least 20 occurrences in which 'Peking University' maps to China, and no mappings of 'Peking University' to other countries.

3.1.3 Pull request selection. After we collected all the data, we applied a number of criteria – in order – to select the set of pull requests to use as the basis for our analysis. We first examined the state of the pull requests. At any time, a pull request can be in 'open', 'merged' or 'not-merged' state. 'Merged' pull requests were accepted by the integrator, 'not-merged' pull requests were rejected, and 'open' pull requests did not have a decision associated with them yet. We included in our analysis 'merged' and 'not-merged' pull requests, but excluded 'open' pull requests since we cannot predict the future of these pull requests.

Second, when we started engaging with the data, we noticed that the distribution of submitters and their countries is highly skewed (kurtosis: γ=98.2). To address this, we only kept those pull requests from countries that represent at least 1% (which equates to 1000+ pull requests) of the total number of pull requests. This ensures diversity while maintaining sufficient data points per country for analysis. As a result, our analysis includes 17 countries: United States (38%), United Kingdom (8%), Germany (6%), France (5%), Canada (4%), Japan (3%), Brazil (3%), Australia (2%), Russia (2%), Netherlands (2%), China (2%), Spain (2%), India (2%), Switzerland (1%), Sweden (1%), Italy (1%), and Belgium (1%). Together, the pull requests originating from the selected 17 countries constitute approximately 83% of the pull requests available for analysis.

As a final step, we removed any pull requests that were integrated by the submitters themselves. This occurred in a somewhat surprisingly high 37% of the cases. Given the focus of our analysis, we only include those pull requests that were integrated by developers other than the submitter.

Table 1: Factors influencing pull request acceptance

Characteristic	Measure			
	Tsay et. al [26]	Gousios et. al [14]	Our study	Variable name
Project characteristics				
Maturity	#months in existence	-	#months in existence	proj_months_existence
Team size	#contributors	#active core team members	#active core team members	-
Popularity	#watchers	#watchers	#watchers	proj_watchers
Size of code	-	#non-comment LOC	#non-comment LOC	proj_ncloc
Openness to external contributions	-	%external contributions	%external contributions	proj_external_contribs
Code quality	-	#test LOC per 1,000 LOC	#test LOC per 100,000 LOC	proj_test_loc_per_lloc
	-	#test cases per 1,000 LOC	#test cases per 100,000 LOC	-
	-	#asserts per 1,000 LOC	#asserts per 100,000 LOC	-
Developer characteristics				
Status in community	#followers	#followers	#followers	dev_followers
Status in project	direct commit access (yes=1; no=0)	-	direct commit access (yes=1; no=0)	dev_commit_access
Social closeness	followed integrator prior (yes=1; no=0)	-	followed integrator prior (yes=1; no=0)	dev_followed_integrator
	watched project prior (yes=1; no=0)	-	watched project prior (yes=1; no=0)	dev_watched_project
Experience	-	#previous pull requests on project	#previous pull requests on project	dev_prev_pull_requests
	-	previous pull request success rate	previous pull request success rate	dev_success_rate
	-	-	#months of project participation	dev_months_participation
Pull request characteristics				
Uncertainty of pull request	#comments	#comments	#comments	pr_comments
Size of change	#changed LOC	#changed LOC	#changed LOC	pr_changed_loc
	#files changed	#files changed	#files changed	pr_changed_files
Quality	inclusion of tests (yes=1; no=0)	#changed LOC in tests	inclusion of tests (yes=1; no=0)	pr_test_inclusion
Geographical location				
Country of submitter			country	geo_country
Same country submitter/integrator			same country (yes=1; no=1)	geo_same_country

Ultimately, of the 370,411 pull requests that were included in the Gousios data set, 70,740 pull requests remain as the subject of the analysis below, all of them containing full location information. Because we performed the removals described above in order, the minimum number of pull requests across all countries was 813.

3.1.4 Statistical method. Pull request acceptance is a binary classification problem. We built two logistic regression models. The first model (termed Model 1 in the remainder of the text) captures the effect of country of submitter on pull request acceptance. The second model (Model 2) captures the effect of the submitter and committer being in the same country on pull request acceptance.

Both models control for the project, developer, and pull request characteristics presented in Table 1, so we can isolate the effect of country. In both models, each pull request (with its unique project, developer, and pull request characteristics) is an independent observation. We choose the R implementation of logistic regression [15][20] and report statistical significance at a p-value <0.05. The chances of pull request acceptance are measured as log odds. If the value of log odds is 0, the chances of pull request acceptance and rejection are same. If the value of log odds is less than 0 then the probability of pull request acceptance is less than the probability of pull request rejection, and vice-versa. The impact of the various characteristics shown in Table 1 on pull request acceptance is reported as percentage of deviance [21]. The interpretation of percentage of deviance is similar to the percentage of total variance explained by least square regression [8].

During feature selection, we add the value 'one' to independent count variables and log-transform the result to stabilize the variance. We verify the variance by using the AIC and Vuong test for non-nested models [29], for both the transformed and original data. We measure the effect size using Cramer's V [9]. We computed the Variance Inflation Factor (VIF) to check for multicollinearity and eliminated highly correlated variables that caused multicollinearity. Any VIF value greater than 5 indicates multicollinearity [8]. As a result, we removed team size measured as active core team members, code quality measured as test cases per kloc, and code quality measured as asserts per kloc from further consideration (which is why Table 1 lists them as '-' for our study).

We show the effect of predictor variables on pull request acceptance via three values: coefficient, standard error, and p-value

(shown as *coefficient* (*standard error*)$^{p-value}$ in Table 2). Each coefficient in the logistic regression models reads as the effect of that variable on the log odds of pull request acceptance when other predictor variables are kept constant. For instance, in Table 2, an increase in *proj_months_existence* by a month decreases (negative coefficient) the estimated chances of pull request acceptance by exp(-0.01) = 0.99. The standard error term gives the confidence interval of the computed log odds: exp(coefficient ± 1.96 × standard error) and the p-value shows the statistical significance of the variable to predict pull request acceptance. P-values less than 0.05 are considered statistically significant.

A slight difference exists in the interpretation of the coefficients for continuous and categorical variables. The coefficient of a continuous variable signifies the estimated change in the log odds of pull request acceptance for a unit increase in the value of a continuous variable. The coefficient of a categorical variable shows the change in log odds of pull request acceptance relative to a base-level or log odds ratio. For instance, to measure the influence of country, we choose the United States, where the majority of developers reside, as the base level. The influence of being located in the United Kingdom, then, reads as the estimated change in the log odds ratio of pull request acceptance when the country is the United Kingdom versus the United States. This implies that, for every 100 pull requests accepted from the United States, approximately exp(0.13)*100=114 pull requests are accepted from the United Kingdom. Similarly, to study the influence of submitters and integrators being in the same country, we present the estimated changes in log odds ratio of pull request acceptance when the submitter and integrator are from the same country versus different countries.

Each model is summarized in terms of six values: Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Log Likelihood, Deviance, Number of Observations (Num. obs.), and Area Under Curve (AUC). Number of Observations reports the count of pull requests analyzed for data modeling. AIC, BIC, Log Likelihood, and Deviance assess the fit of the model: the lower their values, the stronger is the fit of the model. We also evaluated the fitness of the model by calculating the Area Under Curve (AUC). An AUC greater than 0.5 is considered acceptable.[3]

The deviance table (Table 3) presents the variance explained by each predictor variable. The NULL model is the default model without any predictor variables. Predictor variables are added to improve the model fit and are characterized by degree of freedom (Df), deviance explained (Deviance), and statistical significance (p-value<0.05) of the variable. For ease of understanding, the residual degree of freedom and residual deviance after the addition of each predictor variable are provided.

4 RESULTS

This section presents our findings from the data analysis of GitHub projects' archival data.

4.1 GitHub projects' archival data analysis

Before we discuss the effect of geographical location, we first briefly examine the control variables as shown in Table 2. We note that, in comparing Model 1 (effect of country of submitter on pull request acceptance) and Model 2 (effect of submitter and integrator being in the same country on pull request acceptance), nearly identical coefficients, standard errors, and p-values (significance indicated by stars: *** p<0.001; ** p<0.01; * p<0.05) result for the control variables. An increase in the maturity of a project, popularity of a project, openness of a project to external contributions, size of the code base, number of months a submitter has participated in a project, size of the change in code contained in the pull request, and uncertainty in the pull request all correlate with a decrease in pull request acceptance. An increase in code quality of the project, number of past pull requests on the project by the submitter, past success rate of the submitter, status in the project of the submitter, quality of the pull request, and whether the submitter followed the project prior to contribution, all correlate with an increase in pull request acceptance.

Note that the effect sizes generally are relatively small, with the exception of quality of the change (*pr_test_inclusion*), uncertainty in the pull request (*pr_comments*), and the experience of the developer (*dev_prev_pull_requests*), which have the strongest effect sizes. In this context, we note that, for variables *proj_watchers*, *proj_test_loc_per_lloc*, and *dev_months_participation*, we choose to retain the sign of the coefficients, even though the magnitude of the coefficient is less than the precision at which we report results.

Shifting focus to the effect of country, Model 1 in Table 2 shows the chances of pull request acceptance for different countries. As discussed, results are to be interpreted relative to the United States, which has the largest population of submissions. As one example, for every 100 pull requests accepted from submitters from the United States, an estimated count of exp(0.26)×100=130 pull requests are accepted from submitters from the Netherlands. This means that pull requests from Netherlands based submitters have a higher chance of being accepted than pull requests from submitters based in the United States.

We identify three groups of countries, each arranged in decreasing order of chances of pull request acceptance. The first group, consisting of Switzerland (146), Netherlands (130), Japan (128), United Kingdom (114), and Canada (113), represents countries in which submitters have a higher chance of pull request acceptance relative to the United States. The second group, consisting of Sweden (81), Germany (78), Brazil (76), Italy (73), and China (68), represents countries in which submitters have a lower chance of pull request acceptance relative to the United States. Finally, countries in the third group include Belgium (109), Spain (108), Australia (105), India (102), France (102), and Russia (94), which exhibit a pull request acceptance rate that is not distinguishable from that of submitters from the United States.

Examining this grouping, we cannot necessarily identify patterns. We might have expected, for instance, to see consistently higher or lower acceptance rates from countries in the same continent, or consistently higher acceptance rates from countries in which English is either the primary language or a major secondary language in school. This is not the case. The three countries with the highest rates (Switzerland, Netherlands, Japan) are from two different continents; the three countries with the lowest rates are from three different continents, yet two of those continents are also represented among the three countries with the higher rates. In terms of English, Sweden and Germany are known for teaching

[3] https://www.kaggle.com/wiki/AreaUnderCurve

Table 2: Logistic regression models of factors influencing pull request acceptance [AUC: 0.7]

	Model 1	Model 2
(Intercept)	2.82 (0.14)***	2.61 (0.14)***
Control variables		
proj_months_existence	−0.01 (0.00)***	−0.01 (0.00)***
proj_watchers	−0.00 (0.00)***	−0.00 (0.00)***
log(proj_ncloc + 1)	−0.06 (0.01)***	−0.06 (0.01)***
proj_external_contribs	−0.01 (0.00)***	−0.01 (0.00)***
proj_test_loc_per_lloc	0.00 (0.00)***	0.00 (0.00)***
log(dev_followers + 1)	0.06 (0.01)***	0.07 (0.01)***
dev_commit_access	0.06 (0.07)	0.05 (0.07)
dev_followed_integrator1	0.11 (0.03)**	0.10 (0.03)**
dev_watched_project1	0.04 (0.03)	0.05 (0.03)
log(dev_prev_pull_requests + 1)	0.17 (0.01)***	0.17 (0.01)***
dev_success_rate	0.01 (0.00)***	0.01 (0.00)***
dev_months_participation	−0.00 (0.00)***	−0.00 (0.00)***
log(pr_comments + 1)	−0.25 (0.01)***	−0.24 (0.01)***
log(pr_changed_loc + 1)	−0.06 (0.01)***	−0.06 (0.01)***
log(pr_changed_files + 1)	0.01 (0.02)	0.01 (0.02)
pr_test_inclusion1	0.26 (0.03)***	0.26 (0.03)***
geo_country_switzerland	0.38 (0.11)***	0.46 (0.11)***
geo_country_netherlands	0.26 (0.09)**	0.36 (0.09)***
geo_country_japan	0.25 (0.08)***	0.34 (0.08)***
geo_country_united kingdom	0.13 (0.04)**	0.20 (0.04)***
geo_country_canada	0.12 (0.07)	0.22 (0.07)**
geo_country_belgium	0.09 (0.12)	0.18 (0.12)
geo_country_spain	0.08 (0.10)	0.15 (0.10)
geo_country_australia	0.05 (0.07)	0.14 (0.07)
geo_country_india	0.02 (0.07)	0.12 (0.07)
geo_country_france	0.02 (0.06)	0.11 (0.06)
geo_country_russia	−0.06 (0.07)	0.04 (0.07)
geo_country_sweden	−0.21 (0.09)*	−0.10 (0.09)
geo_country_germany	−0.25 (0.04)***	−0.16 (0.05)***
geo_country_brazil	−0.27 (0.06)***	−0.19 (0.07)**
geo_country_italy	−0.31 (0.08)***	−0.21 (0.09)*
geo_country_china	−0.39 (0.09)***	−0.27 (0.10)**
geo_same_country1		0.18 (0.03)***
AIC	49231.59	49198.20
BIC	49534.09	49509.87
Log Likelihood	−24582.80	−24565.10
Deviance	49165.59	49130.20
Num. obs.	70740	70740

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Deviance explained by factors influencing pull request acceptance

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			70739	52354.02	
proj_months_existence	1	509.46	70738	51844.56	0.0000
proj_watchers	1	175.49	70737	51669.07	0.0000
proj_external_contribs	1	321.47	70736	51347.59	0.0000
proj_test_loc_per_lloc	1	47.84	70735	51299.75	0.0000
log(proj_ncloc + 1)	1	30.85	70734	51268.90	0.0000
dev_months_participation	1	25.46	70733	51243.45	0.0000
log(dev_prev_pull_requests + 1)	1	1014.64	70732	50228.81	0.0000
dev_success_rate	1	389.62	70731	49839.19	0.0000
pr_test_inclusion	1	22.73	70730	49816.46	0.0000
log(pr_changed_loc + 1)	1	164.02	70729	49652.44	0.0000
log(pr_changed_files + 1)	1	0.11	70728	49652.34	0.7444
pr_comments	1	2.62	70727	49649.72	0.1055
log(dev_followers + 1)	1	57.46	70726	49592.26	0.0000
log(pr_comments + 1)	1	273.45	70725	49318.81	0.0000
dev_watched_project	1	2.82	70724	49315.99	0.0931
dev_followed_integrator	1	6.79	70723	49309.20	0.0092
geo_country	16	143.61	70707	49165.59	0.0000
geo_same_country	1	35.39	70706	49130.20	0.0000

students English in schools, yet are among the countries with the lowest acceptance rates. We return to this discussion in Section 6.

We note that the coefficients for the various countries are larger than the coefficients for the control variables. This does not necessarily mean that the effect is larger. We remind the reader that, while most control variables are numerical and thus interpreted relative to the intercept, countries are categorical and interpreted relative to a baseline (United States). The variability in pull request acceptance across different countries, thus, is responsible for the larger coefficients.

Compared to Model 1, Model 2 in Table 2 documents the effect of taking into account whether or not the pull request submitter is in the same country as the pull request integrator (*geo_same_country*).

First, we note that being in the same country increases the chances of pull request acceptance by a factor of exp(0.18)=1.2 as compared to when the submitter and integrator are from different countries. To understand whether this is a consistent phenomenon across all countries, or whether this is somehow an effect of one or a few countries dominating the results (for instance, based on a much higher volume of 'same country contributor and integrator' pull requests coming from those countries), we analyzed our data per country as well. We found that, for all countries except India, acceptance rates for contributions coming from submitters from the same country as integrators were higher, ranging from just 2.4% higher (United States) to 13.4% higher (China). The sole exception, India, was lower in acceptance rate by 7.4%. This implies that integrators imperceptibly have a higher preference for contributions from their own countries (except for, again, India, which we discuss later in the paper).

Second, again comparing the results from Model 2 to Model 1, we observe that, compared to the United States, Canada is included in the group of countries from which pull requests are more likely to be accepted, and Sweden is not in the group of countries for which pull request acceptance is less likely. With no other significant shifts in results, the prior observations regarding an absence of possible patterns in the groupings hold for Model 2 as well.

Finally, we observe that the coefficients of all countries increase in Model 2 as compared to Model 1. We caution against over-interpretation of the difference here. The variable *geo_same_country* clearly is related with the various *geo_country* variables. Given that its addition to the model is the only addition, it is multicollinearity that we believe causes the amplification of the coefficients for the individual countries. Interpreting the increase in other ways is likely to draw false conclusions. The reason we did not remove the individual countries from Model 2 is that, given our results in

Model 1, they serve as a control for the independent variable of *geo_same_country*.

Table 3 shows the impact on pull request acceptance in terms of deviance. Relative to the control variables, the country of submitters (*geo_country*) and the same country of submitters and integrators (*geo_same_country*) each explain some of the effect present. While the effect of being in the same location is relatively small (not surprising given that it is multicollinear and residual after the effect of individual countries is already accounted for), the effect of *geo_country* shows actually a moderate effect. This implies, once again, that integrators are somehow differentiating pull requests from submitters from different countries.

5 THREATS TO VALIDITY

5.1 Construct validity

Pull request acceptance. A pull request can be reopened multiple times after it is first merged or rejected. Because we chose to use the first decision regarding acceptance and not later choices, our results may not reflect eventual decisions if later decisions exhibit a different pattern. This choice was made in previous studies (e.g., [26][31]), but future work should consider extending our study (and previous studies) to assess the entire life cycle of a pull request.

Other confounding factors. It is possible that other confounding factors (e.g., academic versus industry affilitation, programming language, gender) may impact our findings.

5.2 Internal validity

Data accuracy. The accuracy of the results of a study depends on the accuracy of the data on which it is built. To mitigate collecting our own data and possibly introducing errors, we used GHTorrent data, which has been extensively used in several prior studies (e.g., [25][28]). The only additional data collection we performed concerned obtaining the geographic location of submitters and integrators. For this, we used an existing script used in prior research and added a heuristic to improve on its results.

5.3 External validity

Generalisability. The quantitative analysis presented in this study is performed on a subset of projects found on GitHub. Our sample is not representative of all software projects (see Section 3.1.2 as to why).

6 DISCUSSION

Is there a difference in pull request acceptance rates among different countries? Our analysis of GitHub pull requests indeed shows differences to exist, with the differences significant for a number of countries as compared to the United States as a baseline. Five countries (i.e., Switzerland, the Netherlands, Japan, United Kingdom, and Canada) have significantly higher acceptance rates; five other countries (i.e., Sweden, Germany, Brazil, Italy, and China) have significantly lower acceptance rates.

Despite these wide differences in acceptance rates across the 17 countries analyzed, all countries except India, have a higher pull request acceptance rate when both the submitters and integrators are from the same country.

What factors cause difference in pull request acceptance? It is known that many different factors can influence pull request acceptance, with several previous studies (discussed throughout the paper) identifying various factors and showing how they play a role. Our study adds two additional factors: (1) country of the submitter who issues pull request, and (2) co-location of a submitter and integrator in the same country. Controlling for the factors identified in prior studies, we built two models through which we showed that both the country of the submitter who issues a pull request and the co-location of a submitter and integrator in the same country have an effect on pull request acceptance. Our study, thus, adds to the set of factors known.

What could explain the effect of country on pull request acceptance? This is perhaps the most difficult question to answer. Our analysis already eliminates, through an extensive set of control variables, typical reasons for differences in pull request acceptance rates. Moreover, as part of our analysis we dove deeper into our results to make sure that hidden effects within the data were not inadvertently leading us to draw certain conclusions. We carefully set minimum thresholds (to avoid over-interpretation of small data) and examined whether certain skews in the data could be a cause (such as, for instance, one country dominating all others). These kinds of skews turned out not to be present. This means we need to look elsewhere for possible explanations as to why country might have something to do with pull request acceptance.

One particular reason we looked at was language,... Another reason could be demographics: could one country mostly have, for instance, college students contributing whereas another has mostly professionals? Our data does not capture this information, though we have no good reason to believe that such strong differences in open source participant demographics exist along country lines.

As can been seen from the above discussions, it is evident that there is a lot more work required in this space to understand the reasons for our observed distribution. We hope to partner with experts in the empirical community with experience in qualitative analysis to help investigate these factors more. Additionally, we look forward to collaborating with other empirical researchers in the analyzed countries to understand better if there are specific cultural or external factors which can help in understanding the results. That said, this empirical study to the best of our knowledge is the first step towards understanding this problem. In empirical studies it is important to contextualize the environment in which the results are obtained to generalize results across studies. In general, practitioners become more confident in a theory when similar findings emerge in different contexts [6]. Towards this end, we hope that other researchers replicate our study in different context and environments to build an empirical body of knowledge. Overall, at present, we do not have a strong reason as to why the differences are as they are. We can only conclude that they exist.

What is next? Our study represents a first step in highlighting the issue of differences in pull request acceptance across submitters from different countries. The issue must be understood in more detail: why it seems to happen, how it happens, what triggers it happening, and in what ways might it be preventable? We advocate a multi-faceted research agenda moving forward: replication of

our study on other platforms and in commercial settings, investigation of potential factors explaining differences in acceptance rate through field and laboratory studies.

REFERENCES

[1] 2017. Bitbucket. https://bitbucket.org/
[2] 2017. GHTorrent. http://ghtorrent.org/downloads.html.
[3] 2017. GitHub, Inc. https://github.com/
[4] 2017. Stack Exchange. https://stackexchange.com/
[5] 2017. Stack Overflow. https://stackoverflow.com/
[6] Victor R Basili, Forrest Shull, and Filippo Lanubile. 1999. Building knowledge through families of experiments. *IEEE Transactions on Software Engineering* 25, 4 (1999), 456–473.
[7] Christian Bird, Alex Gourley, Prem Devanbu, Anand Swaminathan, and Greta Hsu. 2007. Open borders? immigration in open source projects. In *Mining Software Repositories, 2007. ICSE Workshops MSR'07. Fourth International Workshop on.* IEEE, 6–6.
[8] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences.* Routledge.
[9] Harald Cramér. 1999. *Mathematical methods of statistics.* Vol. 9. Princeton university press.
[10] Shane Curcuru. 2015. 3 key elements that define every open source. https://opensource.com/life/15/2/3-key-elements-every-open-source-project.
[11] Georgios Gousios, Martin Pinzger, and Arie van Deursen. 2014. An exploratory study of the pull-based software development model. In *Proceedings of the 36th International Conference on Software Engineering.* ACM, 345–355.
[12] Georgios Gousios, Margaret-Anne Storey, and Alberto Bacchelli. 2016. Work practices and challenges in pull-based development: The contributor's perspective. In *Software Engineering (ICSE), 2016 IEEE/ACM 38th International Conference on.* IEEE, 285–296.
[13] Georgios Gousios, Margaret-Anne Storey, and Alberto Bacchelli. 2016. Work Practices and Challenges in Pull-Based Development: The Contributor's Perspective. In *Proceedings of the 38th International Conference on Software Engineering (ICSE).* https://doi.org/10.1145/2884781.2884826
[14] Georgios Gousios and Andy Zaidman. 2014. A Dataset for Pull-based Development Research. In *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014).* ACM, New York, NY, USA, 368–371. https://doi.org/10.1145/2597073.2597122
[15] Marek Hlavac. 2015. *stargazer: Well-Formatted Regression and Summary Statistics Tables.* http://CRAN.R-project.org/package=stargazer R package version 5.2.
[16] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on Computer supported cooperative work.* ACM, 117–128.
[17] Nora McDonald and Sean Goggins. 2013. Performance and participation in open source software on github. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems.* ACM, 139–144.
[18] Gianmarco Paris, Giulio De Leo, Paolo Menozzi, and Marino Gatto. 1998. Region-based citation bias in science. *Nature* 396, 6708 (1998), 210.
[19] Raphael Pham, Leif Singer, Olga Liskin, Fernando Figueira Filho, and Klaus Schneider. 2013. Creating a shared understanding of testing culture on a social coding site. In *Software Engineering (ICSE), 2013 35th International Conference on.* IEEE, 112–121.
[20] R Core Team. 2015. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[21] Baishakhi Ray, Daryl Posnett, Vladimir Filkov, and Premkumar Devanbu. 2014. A large scale study of programming languages and code quality in github. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering.* ACM, 155–165.
[22] Anna Sandberg. 2015. Competing biases: Effects of gender and nationality in sports judging. (2015).
[23] Walt Scacchi. 2007. Free/open source software development: Recent research results and methods. *Advances in Computers* 69 (2007), 243–295.
[24] Daricélio Moreira Soares, Manoel Limeira de Lima Júnior, Leonardo Murta, and Alexandre Plastino. 2015. Acceptance factors of pull requests in open-source projects. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing.* ACM, 1541–1546.
[25] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, and Chris Parnin. 2016. *Gender bias in open source: Pull request acceptance of women versus men.* Technical Report. PeerJ PrePrints.
[26] Jason Tsay, Laura Dabbish, and James Herbsleb. 2014. Influence of social and technical factors for evaluating contribution in GitHub. In *Proceedings of the 36th international conference on Software engineering.* ACM, 356–366.
[27] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2015. Perceptions of diversity on GitHub: A user survey. *CHASE. IEEE* (2015).
[28] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and tenure diversity in GitHub teams. In *CHI. ACM.*
[29] Quang H Vuong. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society* (1989), 307–333.
[30] Peter Weißgerber, Daniel Neu, and Stephan Diehl. 2008. Small patches get in!. In *Proceedings of the 2008 international working conference on Mining software repositories.* ACM, 67–76.
[31] Y. Yu, H. Wang, V. Filkov, P. Devanbu, and B. Vasilescu. 2015. Wait for It: Determinants of Pull Request Evaluation Latency on GitHub. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories.* 367–371. https://doi.org/10.1109/MSR.2015.42