4-2021

How successful are open source contributions from countries with different levels of human development?

Leonardo FURTADO

Bruno CARTAXO

Christoph TREUDE
Singapore Management University, ctreude@smu.edu.sg

Gustavo PINTO

Citation

FURTADO, Leonardo; CARTAXO, Bruno; TREUDE, Christoph; and PINTO, Gustavo. How successful are open source contributions from countries with different levels of human development?. (2021). *IEEE Software*. 38, (2), 58-63.
Available at: https://ink.library.smu.edu.sg/sis_research/8822

How Successful Are Open Source Contributions From Countries with Different Levels of Human Development?

Leonardo B. Furtado
Federal University of Pará
Belém, Pará, Brazil
srleonardofurtado@gmail.com

Bruno Cartaxo
Federal Institute of Pernambuco
Paulista, Pernambuco, Brazil
email@brunocartaxo.com

Christoph Treude
University of Adelaide
Adelaide, Australia
christoph.treude@adelaide.edu.au

Gustavo Pinto
Federal University of Pará
Belém, Pará, Brazil
gpinto@ufpa.br

***Abstract*—Are Brazilian developers less likely to have a contribution accepted than their peers from, say, the United Kingdom? In this paper we studied whether the developers' location relates to the outcome of a pull request. We curated the locations of 14k contributors who performed 44k pull requests to 20 open source projects. Our results indeed suggest that developers from countries with low human development indexes (HDI) not only perform a small fraction of the overall pull requests, but they also are the ones that face rejection the most.**

***Index Terms*—Open source projects; Pull requests; Human development index**

I. Introduction

Developers in open source software (OSS) projects must make decisions on contributions made by other community members, such as whether or not to accept a pull request. Previous studies have shown that factors such as gender and community status may influence the chances of contributions being accepted [1].

In this paper we studied whether developers based in countries with low human development are less likely to succeed in contributing to OSS projects. We used the Human Development Index (HDI) to measure the human development of a country. HDI measures three dimensions of human development: health, education, and income per capita. According to the United Nations Development Programme[1], "the health dimension is assessed by life expectancy at birth, the education dimension is measured by mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The standard of living dimension is measured by Gross Domestic Product (GDP) per capita." HDI is also widely used by the United Nations (UN) [2] and many other international organizations.

To conduct this work, we analyzed 44,630 pull requests performed by 14,133 contributors into 20 well-known and well-studied OSS projects. Our investigation suggests that, indeed, developers based in low HDI locations perform fewer pull requests and, proportionally, are the ones with the highest rejection rates.

[1]http://hdr.undp.org/en/content/human-development-index-hdi

II. Method

To conduct our work, we chose OSS projects that are (1) long-lived (i.e., more than two years of historical records), (2) popular (i.e., more than 5,000 stars on GitHub), (3) well-studied (i.e., studied in other research works), (4) diverse (i.e., in terms of their domain), and (5) active (i.e., more than 1,000 pull requests submitted). We then manually selected a few OSS projects that met these criteria. They are: ATOM/ATOM, D3/D3, PHP/PHP-SRC, MICROSOFT/VSCODE, DJANGO/DJANGO, MONGODB/MONGO, IONIC-TEAM/IONIC, PYTHON/CPYTHON, FACEBOOK/REACT, MOZILLA-MOBILE/FIREFOX-IOS, APPLE/SWIFT, HOMEBREW/BREW, SCIKIT-LEARN/SCIKIT-LEARN, LARAVEL/LARAVEL, ANGULAR/ANGULAR, ZULIP/ZULIP, FACEBOOK/REACT-NATIVE, SPYDER-IDE/SPYDER, TENSORFLOW/TENSORFLOW, and VUEJS/VUE.

For each OSS project, we crawled contributors' (e.g., names, GitHub handles, location, etc.) and contributions' (e.g., pull requests performed, pull request status, etc.) data. Overall, we obtained data from 16,836 contributors and 96,592 contributions. We applied some criteria for analyzing pull requests data:

- First, we excluded pull requests that were integrated by the submitters themselves, thus excluding 22,356 pull requests.
- Second, we identified contributors with organizational email addresses and we excluded their pull requests. We did this because these developers can work for companies that support these projects and have a large stake in sending pull requests, most of which are more likely to be accepted. This excluded other 5,544 pull requests.
- Third, we excluded pull requests from contributors who are part of the project organization or who are part of some organization that funds this project. To find the names of these organizations we inspected project pages and looked for backers or funders pages. This led to the exclusion of 18,823 pull requests.

After these procedures, we were left with 49,869 pull requests from 15,654 contributors.

Since location is not a mandatory field on GitHub, we observed that not all contributors have filled it. We discarded contributors that did not provide their location. Moreover, on GitHub, the location field is a free text form; therefore, GitHub users can fill it with any information. We created a tool that matches the textual information provided in the location field with a location database, curated by `simplemaps.com` (Simplemaps for simplicity). Simplemaps is a database that provides the name of cities, states, countries, and other geographical information. According to its website, they "built it from the ground up using authoritative sources such as the NGIA, US Geological Survey, US Census Bureau, and NASA."[2] Although Simplemaps provides a comprehensive database, some adjustments were still needed. For instance, since we perceived that some GitHub users fill their locations with well-known acronyms (e.g., developers often use NY and NYC to mean New York City), we had to enrich the database with them. Using this approach, we were able to categorize 14,133 (90%) of contributors that filled the location field. We discarded the 1,521 GitHub contributors for which we could not infer their location.

Regarding the contributions, we focused only on *closed* pull request, due to our interest in analyzing the relationship between the contributors' location and the acceptance/rejection of the pull request. Therefore, we had to rely on pull requests that have already passed through the code review process. A total of 44,630 pull requests were then selected for analysis. These pull requests were submitted between September 2010 and September 2019 (when we collected data).

Regarding the countries' population and HDI, we used the UN database[3]. We adopted the same four level HDI stratification (very high, high, medium, and low) that UN traditionally uses in their reports [2]. We considered the year of 2018, the most recent data available.

Our data and tools are available at: https://github.com/LeonardoFurtado/github-user-informations-collector.

III. Results

Contributions based on the location. Overall, developers from the United States, United Kingdom, and Germany performed the highest number of contributions (20,731 out of the 44,630 analyzed PRs), regardless if the contribution was accepted or not. In particular, contributors based in the United States are by far the most active ones in this regard, performing 14,795 PRs (33% of the total contributions). Table I summarizes the top-20 locations of the developers that contributed the most to our studied projects. If we consider the countries' HDI, only four (20%) out of the top-20 are not at the Very High HDI level (0.800–1.000), namely China, India, Brazil, and Ukraine. This lack of representativeness for lower HDI countries becomes even more evident when we consider the top-20 countries, but this time by the number of PRs per country population. All of them have Very High HDI levels. This shows that, although countries like China, India, and Brazil are in the top-20 when considering the absolute number of PRs, this is probably due to their large populations. In terms of individual work, Canada is the location that has the highest ratio of PRs per contributor (4.15 PRs/contributor), followed by France (3.77 PRs/contributor), and United States (3.72 PRs/contributor). On the other hand, Latin America-based and Africa-based developers are significantly less active than their peers from North America, Europe, and Oceania. Latin American developers performed only 1,183 (2%) of the total contributions in our dataset.

Acceptance based on the location. On average, 32% of the PRs from developers of all countries were accepted. This number rises to 41% when we consider just the top-20 countries in Table I. Japan-based developers are the ones with the highest acceptance ratio (51%) followed by United Kingdom based developers (48%), and Australia-based developers (46%). When we look at the how the PR acceptance ratio relates with countries' HDI, we can see that the higher HDI levels tend to have a higher PR acceptance ratio on average, as one can see in Table II. An exception are the countries grouped at the Low HDI level, which have a PR acceptance ratio more similar to the countries grouped under the Very High HDI level, on average. However, this discrepancy may occur due to the difference between the number of contributors in countries at the Low HDI level (58) and in countries at the Very High HDI level (11,344). This is also observable looking at the number of PRs, which is 110 summing up all Low HDI level countries, while the same number is 36,972 for the Very High HDI level. As a consequence there are countries like Syria, Rwanda, and Senegal with an acceptance rate of 50%, but with only two PRs. Developers in Africa, for instance, had 52% of their PRs accepted (although they have performed only 389 pull requests). South Africa based developers, in particular, contributed with 117 of these PRs (with 59% acceptance rate), although the country has a High HDI (0.705). South American developers faced an even smaller ratio (only 34% of their 682 contributions were accepted). Moreover, we noted that 8,794 contributors (19% of the total) performed just one single PR (the so-called drive-by-commits or casual contributors [3], that is, contributors that perform at most one contribution and leave the project). We found that casual contributors are more frequently based in the United States and United Kingdom (43% of developers based in these two countries performed just one pull request). In a manual inspection of these casual contributions, we found that a significant number of them are related to improving the documentation (e.g., pull request `18353`[4] on APPLE/SWIFT), although more complex contributions exist, such as the one from a Poland-based contributor who fixed a bug that occurred during the installation of the ATOM/ATOM project on Ubuntu linux[5].

Rejection based on the location. In terms of rejection, it

[2]https://simplemaps.com/data/world-cities
[3]http://hdr.undp.org/en/data
[4]https://github.com/apple/swift/pull/18353
[5]https://github.com/atom/atom/pull/3773

TABLE I: Number of pull requests performed per developers' location.

Country	# PRs	# PRs/Pop.M.	# Contr.	# PRs/Contr.	Acc.	HDI
USA	14,795	12.91	4,223	3.50	44.45%	0.920
UK	3,179	13.22	887	3.58	48.13%	0.920
Germany	2,757	11.01	915	3.01	44.98%	0.939
India	2,590	0.51	696	3.72	35.71%	0.647
Canada	2,301	14.93	554	4.15	35.77%	0.922
France	2,053	8.38	545	3.77	45.93%	0.891
China	1,860	0.53	758	2.45	39.09%	0.758
Australia	1,212	14.62	364	3.33	46.70%	0.938
Japan	1,171	2.96	377	3.11	51.24%	0.915
Netherlands	942	21.64	370	2.55	32.91%	0.933
Russia	846	2.23	325	2.60	39.01%	0.824
Brazil	780	1.62	339	2.30	32.56%	0.761
Poland	650	5.73	217	3.00	25.69%	0.872
New Zealand	576	33.40	157	3.67	39.93%	0.921
Sweden	544	19.70	197	2.76	40.07%	0.937
Switzerland	432	19.06	162	2.67	45.83%	0.946
Taiwan	362	4.77	113	3.20	44.20%	0,911
Spain	352	3.38	158	2.23	27.27%	0.893
Ukraine	338	3.44	152	2.22	33.43%	0.750
South Korea	325	2.40	123	2.64	59.69%	0.906

PRs: Pull requests; Pop.M: Country Population (in Millions); Contr: Contributors; Accept: Acceptance Ratio.

TABLE II: HDI vs. PR acceptance ratio

Human Development Index (HDI)	Acceptance Ratio		
	Median	Mean	Std. Dev.
Very High (0.800–1.000)	39.18%	35.46%	15.79%
High (0.700–0.799)	28.57%	29,02%	24,24%
Medium (0.550–0.699)	20.29%	30,17%	33,63%
Low (<0.549)	36,36%	32,62%	26,48%

seems that, regardless of their location, having a pull request rejected is commonplace. In particular, 59% of the overall pull requests were rejected. Interestingly, developers from 29 locations had 100% of rejections. These contributors, however, made very few contributions (i.e., developers from locations such as Paraguay, Ethiopia, and Burma performed at most seven pull requests). When manually inspecting these 100% rejected pull requests, we noted that some contributors may not yet master how to use Git/GitHub. For instance, pull requests `12336`[6] on SCIKIT-LEARN/SCIKIT-LEARN does not change a single line of code, and has a misleading commit message. Moreover, developers from other low HDI locations have contributed more frequently, but still face a high rejection rate. For example, developers based in Indonesia submitted 107 pull requests, with 78 of them rejected (72%). Bangladesh-based developers submitted 65 pull requests with 87% of them rejected. When taking into account only the developers' locations with more than 250 pull requests, we found that Poland-based developers were the ones that faced the most rejections (74% of their pull requests were rejected), followed by Spain-based developers (72%), and Brazil-based developers (67%).

IV. Related Work

von Engelhardt [4] employed some heuristics on SourceForge (e.g., email headers, time-zone, and IP address) to infer the location of contributors. Bird and colleagues [5] employed heuristics such as the email domain, social networks, and even commit history to determine the location of the top contributors of Firefox and Eclipse. Spinellis [6] analyzed the FreeBSD operating system by investigating the impact of geographical location on code quality. Vasilescu and colleagues [7] used the GitHub location to infer the presence of female developers on OSS projects. Bjørn and Boulus-Rødje [8] studied the role that infrastructural accessibility plays on the success of tech startup in Palestine.

To the best of our knowledge, the work of Rastogi et al [9] is the closest to our work. However, their work focuses on the developers' location that contributed the most. In our study, however, we shed additional light on developers from low HDI locations, which happen to contribute the least or were rejected the most.

V. Implications

Our findings indicate that contributors from lower HDI countries might face a hard time to contribute to open source. Given this observation, open source communities might want to promote sprints, hackathons, and other onboarding programs in these locations. Similarly, companies that fund open source communities might also want to fund mentors in lower HDI locations. These local activities might contribute to foster an open source culture in other less wealthy locations.

VI. Limitations

First, our study is restricted to GitHub; although GitHub is the largest software development platform, we acknowledge that particular countries might have preferences for other platforms. Similarly, developers in some countries may have low participation in certain popular projects because they do not align with the goals of that country or software developers in that country.

[6]https://github.com/scikit-learn/scikit-learn/pull/12336

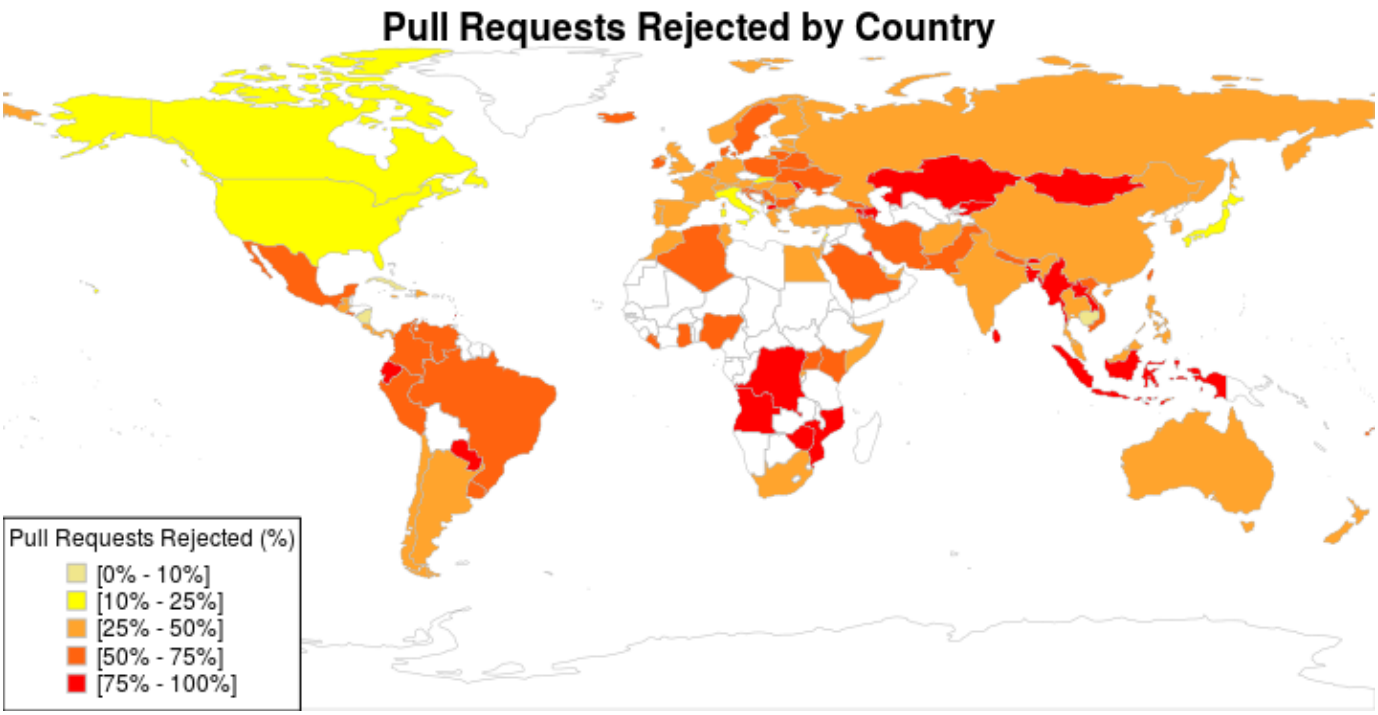

Fig. 1: Pull request rejected per developers' location

Our study is also limited to the number of pull requests studied, which clearly does not represent all possible forms of contributions available in OSS projects. Another limitation is that the location field on GitHub is a free form (i.e., it accepts any information). Although we employed some additional steps to make sure that the location exists, we still may have considered developers with inaccurate locations (e.g., outdated ones). Finally, there are many other factors that may influence the pull request decision making process. Our work focused on one factor, the location. Therefore, it is unclear how other factors correlate to ours, which we left for future work.

VII. Conclusions

In this paper we studied whether the developers' location has any correlation to the pull request decision making. We mined data from 44k pull requests performed in 20 popular OSS projects. We report three main findings: First, developers based in high HDI locations, such as United States, United Kingdom, and Germany, are the ones that contribute the most. Second, in terms of acceptance, again, developers based in high HDI locations such as Japan and United Kingdom have the highest acceptance ratios. Third, in terms of rejection, however, developers based in low HDI locations such as Ethiopia, Burma, and Paraguay never had a single contribution accepted. High rejection rates were also common in other low HDI locations.

References

[1] Jason Tsay, Laura Dabbish, and James Herbsleb. Influence of social and technical factors for evaluating contribution in github. In *ICSE*, pages 356–366, 2014.

[2] United Nations. Human development report 2019: Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century.

[3] Gustavo Pinto, Igor Steinmacher, and Marco Aurélio Gerosa. More common than you think: An in-depth study of casual contributors. In *SANER*, pages 112–123, 2016.

[4] Sebastian von Engelhardt, Andreas Freytag, and Christoph Schulz. On the geographic allocation of open source software activities. *International Journal of Innovation in the Digital Economy (IJIDE)*, 4(2):25–39, 2013.

[5] Christian Bird and Nachiappan Nagappan. Who? where? what?: Examining distributed development in two large open source projects. In *MSR*, pages 237–246, 2012.

[6] Diomidis Spinellis. Global software development in the freebsd project. In *Proceedings of the 2006 International Workshop on Global Software Development for the Practitioner*, GSD '06, pages 73–79, 2006.

[7] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G. J. van den Brand, Alexander Serebrenik, Premkumar T. Devanbu, and Vladimir Filkov. Gender and tenure diversity in github teams. In *CHI*, pages 3789–3798, 2015.

[8] Pernille Bjørn and Nina Boulus-Rødje. Infrastructural inaccessibility: Tech entrepreneurs in occupied palestine. *ACM Trans. Comput.-Hum. Interact.*, 25(5), October 2018.

[9] Ayushi Rastogi, Nachiappan Nagappan, Georgios Gousios, and André van der Hoek. Relationship between geographical location and evaluation of developer contributions in github. In *ESEM*, pages 22:1–22:8, 2018.

Leonardo B. Furtado Is an undergraduate student at the Federal University of Pará, Brazil. He does research on empirical software engineering.

Bruno Cartaxo is an associate professor at the Federal Institute for Education, Science, and Technology of Pernamuco (IFPE), Brazil. He received his Ph.D. and a M.Sc. degree in Computer Science from the Center of Informatics (CIn) at Federal University of Pernambuco (UFPE). He conducts pure and applied research in the broad area of Software Engineering and Technology Transfer.

Christoph Treude is an ARC DECRA Fellow and a Senior Lecturer in the School of Computer Science at the University of Adelaide, Australia. He received his Ph.D. in computer science from the University of Victoria, Canada in 2012. The goal of his research is to advance collaborative software engineering through empirical studies and the innovation of tools and processes that explicitly take the wide variety of artefacts available in a software repository into account. He currently serves on the editorial board of the Empirical Software Engineering journal and was general co-chair for the IEEE International Conference on Software Maintenance and Evolution 2020.

Gustavo Pinto is an assistant professor of computer science at the Federal University of Pará, Brazil. He received his PhD from Federal University of Pernambuco, Brazil in 2015. His research focuses on the interactions between people and code, spanning the areas of software engineering and programming languages. He currently serves as the co-Editor-in-Chief of the Journal of Software Engineering Research and Development (JSERD). Contact him at gpinto@ufpa.br.