The Census of the Brazilian Open-Source Community

Gustavo Pinto[1] and Fernando Kamei[1]

Federal University of Pernambuco
Recife – PE – Brazil
{ghlp,fkk}@cin.ufpe.br

Abstract. During a long time, software engineering research has tried to better understand open-source communities and discover *who* are the contributors and *why* they contribute. Most of these researches focus on well-known OSS projects, such as Linux Kernel or Apache Tomcat. However, there is no study regarding the OSS movement in emerging countries, such as Brazil. In this paper, we attempt to fill this gap by presenting a picture of the Brazilian open-source contributor. To achieve this goal, we examined activities of more than 12,400 programmers on Github, during the period of a year. Subsequently, we correlate our findings with a survey that was answered by more than 1,000 active contributors. Our results show that exists an OSS trend in Brazil: most part of the contributors are active, performing around 30 contributions per year, and they contribute to OSS basically by altruism.

Key words: OSS, Github, Brazilian OSS Community

1 Introduction

The idea of Open Source Software (OSS) has gained more and more attention in the last years. OSS is usually developed by an internet-based community of programmers, without necessarily being paid by an institution. These programmers often rely on code hosting websites to share their contributions. Nowadays, a number of code hosting websites are available, such as SourceForge and Github. A unique characteristic of these websites is that they provide a collaborative environment with a high degree of social transparency. This makes programmers' contributions much more visible and traceable. In Github, for example, one can easily access programmers' information through a REST interface, which otherwise it might be difficult, or even impossible, to find elsewhere.

Over the last decade, some researchers have studied open-source communities and contributions [6, 15, 16], although only few of them regarding the actual scenario of the open-source community in those websites [3, 11]. Hitherto, however, there is a lack in the literature of comprehensive studies targeting the Brazilian open-source community. By the beginning of the 21th century, Brazil is South America's most influential country, an economic giant and one of the world's biggest democracies, with a population size of 200 millions. Moreover,

Brazil has also been a hotbed of open source activity in recent years. Government agencies, private industry and universities have been teaching and implementing open source solutions, creating local centers of knowledge and gain expertise around open source in the country.

Nonetheless, besides the huge open-source investments made in Brazil, few is known about the Brazilian open-source contributor. We argue that it is an important question because, for example, if there is a lack of open-source programmers in one region, the government could create better incentives for software companies in there. On the other hand, if there is a huge number of experienced contributors, software companies could open more opportunities in that location. In this paper we conducted a comprehensive study to understand *who* is this contributor, and *why* (s)he contributes to OSS. To achieve this goal, we extracted data from more than 12,000 Brazilian users on Github. In addiction, we conducted a survey comprising only the active contributors, that is: the contributor that has performed at least one commit in a year. With this data, we are able to uncover these three main research questions:

RQ1: Who is the Brazilian open-source contributor? We found out that most of the Brazilian open-source contributor are male, have between 20-30 years, are currently enrolled in an undergrad course, have between 2 to 5 years of professional software experience, and 2 to 5 years of OSS contribution. Most of them perform around 30 commits per year, but 20.35% of the contributors perform 80.32% of the contributions. Also, we observed that they are basically formed by hobbyist, instead of programmers being paid by software companies.

RQ2: Do the Brazilian contributions to open-source increase over the time? We have found out an existing open-source trend in Brazil. We noticed an increment of over 15% in the absolute number of contributions in one year. However, we also observed that these contributions are not related to more work done by the same users. In fact, the great number of contributors performing few contributions is the reason of this increment.

RQ3: Why do Brazilian programmers contribute to OSS? We found out more than forty motivational factors that motivate users to contribute to OSS. The most common were to **help the community** and to **improve the software that they use** with 19.40% and 17.75%, respectively. Moreover, the majority of the Brazilian OSS community are not motivated by self-marketing but by altruism.

2 Study Design

In this section we present our research questions and our research approach.

2.1 Research Questions

The goal of the study is to better understand the Brazilian open-source community. For this purpose, we elicited three research questions.

– **RQ1:** Who is the Brazilian open-source contributor?
– **RQ2:** Do the Brazilian open-source contributions increase over the time?
– **RQ3:** Why do Brazilian programmers contribute to OSS?

In order to address the research questions, we conducted a two-phase research, adopting a sequential mixed-method approach [7]. In Phase 1, we collected data about Brazilian users on Github (see Section 2.2). After that, on Phase 2, we performed a survey targeting only the active users (see Section 2.3). Finally, we interpret this data and answer the research questions (see Section 3).

2.2 Phase 1 – Mining Github

We used Github data as provided through the GHTorrent project [4], an offline mirror of the data offered through the Github API. Up to September 2013, more than 2 million users and 5 millions projects were collected. In order to identify the Brazilian open-source contributors, we have performed a query on the GHTorrent database, searching through the name of the 26 Brazilian capitals and their states. We also considered the capital of the country. Thus, we have searched users in 53 locations.

After this process, we found a total of 12,485 users. We then have used the Github API to gather more information about these users, and we observed that almost 1,000 of these users have closed their accounts. Also, we manually removed false-positive users, that is, users that location name is similar to the Brazilian ones, but actually the location is located in a different country. Therefore, the population of this study consist of 11,411 Github users. To the best of our knowledge, it is the largest population size found in open-source studies. Our data comprise the period of October 2012 to September 2013.

2.3 Phase 2 – Survey

The questionnaire used in this work was based on the recommendations of Kitchenham et al. [9], following the phases prescribed by the authors: planning, creating the questionnaire, defining the target audience, evaluating, conducting the survey, and analyzing the results. The questionnaire has 10 questions and was structured to limit responses to multiple-choice, Likert scales (responses given in a scale), and also free-forms. After defining all the questions on the questionnaire, we obtained feedback iteratively and clarified and rephrased some questions and explanations. This feedback was obtained from analysis and discussion with a group of specialists and also from one pilot of the survey. Together with the instructions of the questionnaire, we included some simple examples as an attempt to clarify our intent.

Our target population consists of programmers that have performed at least one commit to an open-source software during the analyzed period. After we identified these users, we sent them an email inviting to participate. Over a period of 20 days, we obtained 1,039 responses, resulting in a 16.68% of response

rate. This response rate is more than trice higher than the response rates found in software engineering surveys [8]. The complete list of questions and their answers are available at the companion website[1]. In the remainder of this paper, we discuss the main findings of the survey.

3 Study Results

This section presents our results organized by the research questions.

3.1 Who is the Brazilian open-source contributor?

As of September 2013, GHTorrent reports 11,411 Brazilian open-source contributors. However, not all those users are active: during the period of Oct 2012 – Sep 2013, the majority of the users (6,228 or 54.57% of them) have performed at least one commit. We consider these users as the active group. Hereafter, the following analysis of this study only encompass this group.

The Brazilian open-source community is also more active than the overall Github community (23.38% of the Github users are active). Still, our survey data show that this number of active users was not expected. On average, the respondents believe that only 12% of the population is active (SD: 23.12) and most of them (37.34%) believe that the contributors perform only 5 to 20 contributions per year. Moreover, our data show that most of the respondents are male (97.48% of them), have between 20 and 30 years (66.92%), and are currently enrolled in an undergrad course (36.36%). The participants experience in professional software development is mostly between 2 to 5 years (28.86%). The open-source experience is mostly between 2 to 5 years (36.36%), and then between 1 and 2 years (35.58%). On average, the respondents have contributed to open-source using three different programming languages (3rd Quartile = 5, max = 14). We then analyzed the correlation between the age and number of programming languages used using Pearson Correlation [10]. For Pearson $|r| < 0.3$ indicates small correlation, $0.3 > |r| < 0.5$ indicates medium correlation, and $|r| > 0.5$ indicates strong correlation. We then found out a small correlation ($r = 0.05702044$), which suggests that the age is not related to the number of programming languages used. On the other hand, we found out a strong correlation between age and professional experience ($r = 0.5636623$) and a medium correlation between age and OSS experience ($r = 0.4440954$), which is fairly similar to the results found by Morrison et al. [14].

An important property is the programming language used. Most of the contributions were done using JavaScript (15.88%), PHP (12%), C (11.58%), Java (11.04%), Python (10.43%), Ruby (9.52%), ShellScript (7.40%) and C++ (6.17%). Others languages have 16.60% of contributions. Most of the contributions are done using dynamic languages (60.04% of them) and also object

[1] http://bit.ly/Brazilian_OSS

oriented languages (45.81%) instead of functional languages (2.19%) or mixed programming languages (21.07%). One reason to this is because most of the undergrad courses in Brazil still rely on well-known programming languages, such as C and Java, in their basic courses. Also, Brazil has a number of strong open-source communities. For example, we have one of the largest Java User Group of the world, and one of the most active JavaScript, PHP and Ruby communities in Latin America. And, as reported elsewhere [2], open-source communities play an important role in the education of novice programmers.

The number of contributions per user is generally less than 30 per year (3rd Quartile = 50, max 4,795), with median number of 14, and standard deviation of 177. The reason of this huge standard deviation is due to the high number of outliers (11.70% of the total). The outliers are consisted by users that have performed more than 120 contributions in one year. The Figure 1 shows the distribution of the contributions per user over the year.

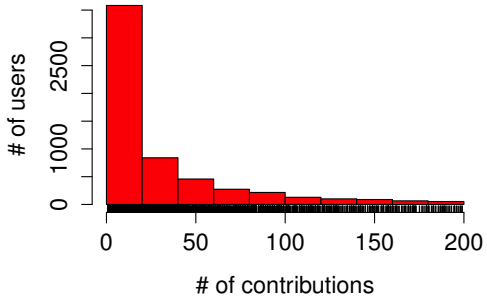

Fig. 1. The number of contributions per user over a year. We have limited the number contributions up to 200 to increase the readability of the figure.

As we can see, most of the active users are low actives (65.59%), performing less than 20 commits per year. On the other hand, we have observed that the Pareto's Laws fits perfectly in our data: 20.35% of the active users perform 80.32% of the contributions. Furthermore, the number of projects contributed per user is generally less than 5 (95% percentile: 12, 90% percentile: 6, 80% percentile: 3), with a median of 3. We have also observed that they spent on average less than 1 day per month working on open-source projects. Our data show that 8.73% of them do contributions every month, and only 2.47% of them do contributions every week. With this data, we can assume that the Brazilian open-source contributor is basically formed by hobbyist, instead of programmers being paid by software companies.

Finally, we mapped the user based on their regions. Brazil is divided into five geopolitical regions. The **North** includes seven states and it includes the Brazilian part of the Amazon rainforest. It is sparsely populated, and its economy is based on plant and mineral exploration. The **Northeast** includes nine states, an arid climate, and an economy based on agriculture, mainly sugarcane.

Region	1st Quartile	Median	3rd Quartile	S. D.	Histogram
North					
Age (years)	20 to 30	20 to 30	30 to 40	0.72	
Education (degree)	Undergrad (ongoing)	Undergrad	Specialization	1.26	
Software Experience (years)	2 to 5	5 to 8	5 to 8	1.40	
OSS Experience (years)	1 to 2	2 to 5	2 to 5	1.03	
Northeast					
Age (years)	20 to 30	20 to 30	30 to 40	0.70	
Education (degree)	Undergrad (ongoing)	Undergrad	Specialization	1.23	
Software Experience (years)	2 to 5	2 to 5	5 to 8	1.05	
OSS Experience (years)	1 to 2	2 to 5	2 to 5	1.01	
Midwest					
Age (years)	20 to 30	20 to 30	30 to 40	0.61	
Education (degree)	Undergrad (ongoing)	Undergrad	Undergrad	1.04	
Software Experience (years)	2 to 5	5 to 8	8 to 12	1.26	
OSS Experience (years)	1 to 2	2 to 5	5 to 8	1.14	
Southeast					
Age (years)	20 to 30	20 to 30	30 to 40	0.65	
Education (degree)	Undergrad (ongoing)	Undergrad	Undergrad	1.16	
Software Experience (years)	2 to 5	5 to 8	8 to 12	1.24	
OSS Experience (years)	1 to 2	2 to 5	5 to 8	1.07	
South					
Age (years)	20 to 30	20 to 30	20 to 30	0.55	
Education (degree)	Undergrad (ongoing)	Undergrad	Undergrad	1.08	
Software Experience (years)	2 to 5	5 to 8	8 to 12	1.21	
OSS Experience (years)	1 to 2	2 to 5	5 to 8	1.13	

Table 1. The Distribution of programmers in the Brazilian geopolitical regions.

The population is concentrated in a few large cities in the coast. The **Midwest** includes three states and Brasilia, the capital of Brazil. Sparsely populated, its economy is based on large farms agriculture and livestock. The **Southeast** includes four states, and it is the most developed region of Brazil. The population is distributed into very large metropolitan areas such as São Paulo and Rio de Janeiro, and mid-size cities. The economy is based on a strong and diverse industry, services, agriculture and livestock. Finally, the **South** includes three states, and its economy is based on automobile and textile, livestock and small farm agriculture. Table 1 shows the user distribution per geopolitical region by age, education, professional software experience, and open-source experience.

We can see a number of interesting findings in the above table. First, as expected, we observed that the professional software experience is related to the OSS experience (the more professional experience, the more OSS experience). We attest this finding by running a Person Correlation ($r = 0.5654273$). Second, there is no correlation between age and the geopolitical region. Most of the respondents have between 20 to 30 years with little standard deviation in all regions. Third, as not expected, education degree is not related to open-source

contribution ($r = 0.2257393$). Only in the north and in the northeast regions more than 25% of the contributors have an specialization degree. Nonetheless, the north group has only 15 samples (northeast has 137), and this sample size could represent a bias of the north population. Finally, we observed that most part of the contributions (62.38% of them) came from the southeast region. That is expected, due to its huge population size, as well as a whole range of universities and software companies located in there.

3.2 Do the Brazilian open-source contributions increase over the time?

As a side issue, we now examine the contributions on a monthly basis. During the analyzed period, we found more than 354,000 commits performed by 6,228 users in more than 98,000 projects. Figure 2 shows the commit evolution during the months.

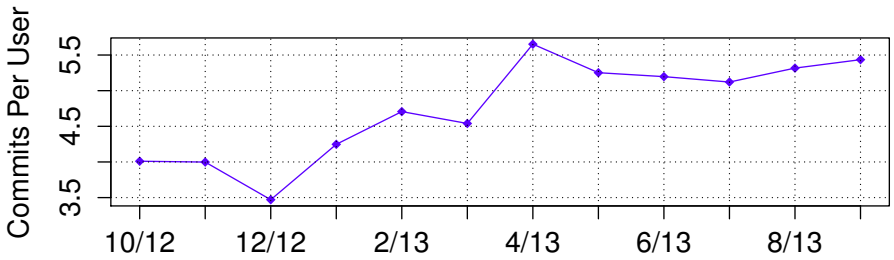

Fig. 2. The evolution of the absolute number of commits divided by the total of users.

At first impression, the above picture shows that the contributions are increasing in absolute number (about 15.08% of increment during this period of time). One might think that the increment of contributions is because the active users are working more hard, and then they are performing more contributions. Nonetheless, we have observed that the number of contributions increase because more users are willing to perform contributions in those months. Also, we noticed that the number of users willing to perform contributions per month is about 24% of the active users. Note that, it does not mean that the same user will perform contribution every month. As we stated earlier, only 8% of the active users do contributions every month. Moreover, we observed that the number of users doing contributions per month can vary greatly. For instance, in August/2013, the month the received the highest number of contributors, the total of contributor is 25% bigger than December/2013, the month the received the lowest number of contributors.

Then, the absence of contributors is reason behind the low number of contributions in December and March. In December, besides be summer break in

universities, Brazil commemorates Christmas and New Year. So, usually, companies provide one week of vacation during this period. Moreover, the carnival, a big Brazilian holiday, occurred in March/2013. And again, during this period of time, universities and companies usually provide two, three, or even more, days of vacations. With this result we can assume that exists an open-source contribution trend in Brazil, and the contributors are willing to perform few contributions during their free time, but not on holidays.

3.3 Why do Brazilian programmers contribute to OSS?

Finally, as an attempt to understand why the Brazilian programmers do contributions to OSS, we analyzed the results of our last survey question: "Why do you do contributions to OSS?". It was not a required question, but 81% of the respondents answered it. After the mining process, we found out more than 40 categories but, due to space constrains, we only describe most interesting ones.

The most common motivation factor was to **help to the community** with 19.40% of the answers. Some respondents said that "To help people that have the same problem", and "[My OSS contribution is] my 50 cents of contribution to the world", and "I use a lot of OSS projects, so I do contributions to help someone else". We observed that the OSS spirit is very strong among respondents. Some of them believe that their contributions are not only about software, but a social contribution as well. This factor is related to the social aspects of the **Open source philosophy**, which has 10.88% of the answers. As an respondent described: "Basically due to the OSS philosophy. I like the way of code sharing and I believe that it might improve my knowledge of [...]". On the other hand, some respondents do contributions because they want to **improve the software that they use** (17.75% of them). Most of these contributions are related to (i) fix simple problems or to (ii) implement new features. These findings are related to the work of Gousios [5], which found out that most of the contributions are consisted of a few lines of code. Then they advice contributors seeking to add a particular feature or fix to "keep it short".

Some respondents do contributions because they believe that they have to **give back the help received once from the community** (17.63%), as a respondent said that "Retribution, because I learned a lot from OSS communities [...]". Another motivation factor is to **improve their programming skill** (15.02%), mainly because to the code revision and discussion, and also improve their human capital by learning things other than programming. Then, the contributor might learn new concepts and good practices. However, despite the personal interest, we found out that only few respondents do contributions just to **improve their own curriculums** (6.27%) or to **gain visibility and reputation on the community** (4.49%). Due to the few number of respondents that have attested it as the main reason of their contributions, we agree that the majority of the Brazilian OSS community are not motivated by self-marketing but by altruism and the fulfilment that arises from writing programs that other persons might use.

4 Related Work

There is a number of researches done regarding the Brazilian IT community, such as the Computer Science scientific production [1], the adoption of agile methods [13], the opportunities for women in IT jobs and education [12], among others. However, to the best of our knowledge, there is no study regarding the Brazilian open-source community in the literature.

Nonetheless, there are several studies regarding personal aspects of open-source communities. One study [6] investigated motivation aspects of 141 kernel developers. The authors revealed various motivational forces that contribute to a person's willingness to engage to OSS, both at the community level as well as the team level. In a similar study, Roberts et al. [15] develop a theoretical model relating the motivations, participation, and performance of OSS developers. They have reported a number of findings, including that developers' motivations are not independent but rather are related in complex ways. Also, they found out that different motivations have an impact on participation in different ways. In another study, Wang et al. [16] described a set of evolution metrics for evaluating open-source software (OSS) and community (OSC). They then measure the evolution of OSS and OSC together, and they showed that the Ubuntu success is due to the growth and maturation of its community.

Our work differs from the state of the art in two key ways. First, none of the above studies try to correlate their data using more than one data source. We argue that it is important because not always what the respondent say is true. We can then minimize this problem by double checking our findings in the two data sources. Second, none of the above studies have such population size. We believe that it is an important aspect, mainly because the use of large samples can increase precision and then reduce bias.

5 Conclusion

In this work we presented an empirical study concerning the Brazilian open-source contributor. We have analyzed an entire year of software development on Github, and we also conducted a survey with the active contributors. With the results of this mixed approach, we observed that: (i) the Brazilian open-source community is active. We have found that more than 54% of the population have performed at least one commit in a year. Nonetheless, as the Pareto's Law suggests, 20.35% of the users have performed 80.32% of the contributions; also, (ii) there is an open-source trend in Brazil. We noticed an increment of over 15% in the absolute number of contributions in one year, although about 24% of the active users perform commits monthly, and only 8% of them are active every month; and finally (iii) altruism is the main motivation, instead of self-marketing aspects. As a future work, we plan to understand the OSS adoption in other countries, and then, correlate our data with them. We also intent to collect more temporal data, as well as to gather information from other

code hosting websites. Finally, we plan to investigate what are the reasons of open-source adoption in Brazil and South America as well.

References

1. D. Arruda, F. Bezerra, V. Neris, P. Toro, and J. Wainera. Brazilian computer science research: Gender and regional distributions. *Scientometrics*, 79(3):651–665, 2009.
2. R. Bagozzi and U. Dholakia. Open source software user communities: A study of participation in linux user groups. *Management Science*, 52(7), 2006.
3. L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012.
4. G. Gousios. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 233–236, 2013.
5. G. Gousios, M. Pinzger, and A. Deursen. An exploration of the pull-based software development model. sep 2013. Submitted to the ICSE'2014.
6. G. Hertel, S. Niedner, and S. Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, 32(7):1159–1177, July 2003.
7. N. V Ivankova, J. Creswell, and S. Stick. Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, 18(1):3–20, 2006.
8. B. Kitchenham and S. Pfleeger. Personal opinion surveys. In *Guide to Advanced Empirical Software Engineering*, pages 63–92. Springer London, 2008.
9. B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Trans. Softw. Eng.*, 28(8):721–734, August 2002.
10. L. Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1):255–268, March 1989.
11. N. McDonald and S. Goggins. Performance and participation in open source software on github. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 139–144. ACM, 2013.
12. C. Medeiros. From subject of change to agent of change: Women and it in brazil. In *Proceedings of the International Symposium on Women and ICT: Creating Global Transformation*, CWIT '05. ACM, 2005.
13. C. Melo, V. Santos, E. Katayama, H. Corbucci, R. Prikladnicki, A. Goldman, and F. Kon. The evolution of agile software development in brazil. *Journal of the Brazilian Computer Society*, 19(4):523–552, 2013.
14. P. Morrison and E. Murphy-Hill. Is programming knowledge related to age? an exploration of stack overflow. In *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 2013.
15. J. Roberts, I. Hann, and S. Slaughter. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Manage. Sci.*, 52(7):984–999, July 2006.
16. Y. Wang, D. Guo, and H. Shi. Measuring the evolution of open source software systems with their communities. *SIGSOFT Softw. Eng. Notes*, November 2007.