COMMUNITY ENGAGEMENT AND COMMUNITY RESPONSE OF FIRST TIME CODE CONTRIBUTORS ON GITHUB

BY

MATTHEW HESTON

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science in Information Architecture
in the Graduate College of the
Illinois Institute of Technology

Approved ______________________
Advisor

Chicago, Illinois
May 2014

ACKNOWLEDGMENT

I want to thank my advisor Dr. Libby Hemphill for her support both in my completion of this thesis and also throughout my studies here at IIT. Her support and guidance have given me the desire to continue a career in research as well as the tools necessary to conduct research properly. I also want to thank my committee members Karl Stolley, who assisted me in the design of the tools I used in collecting data for this thesis as well as analyzing the data, and Aron Culotta, who provided a great deal of valuable feedback in my initial studies of GitHub data. Finally, I want to thank my parents, Robert and Susan Heston, for their continued support during my academic career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

We collect data for 13,383 first time code contributions from 45 projects on the website GitHub and analyze behavior of developers before submitting code as well as community response to code contributions. We find that most developers do not engage with the community on GitHub before attempting to submit code changes. We also find that most code submissions do not elicit much community response, and the metrics we use for community response can not predict whether or not a pull request is accepted. Our findings differ from previous research on open source software communities and social theories of learning in communities of practice. We find most users do not participate in GitHub peripheral activities before submitting code changes. We also find that community response to these submitted code changes is a poor predictor of whether or not the code is accepted.

CHAPTER 1
INTRODUCTION

GitHub is a social website that open source software developers use to host their software projects and to browse other developers' projects. It includes many features that are present on social networking sites, such as the ability to follow other users and leave comments on projects. GitHub provides a wealth of data for studying computer supported cooperative work, as it is a centralized location where many different tasks take place. For example, users can create bug reports, submit fixes, and engage in discussions about new features all on one website.

As open source software continues to grow and more users become reliant on open source technologies, it is important to understand how this software is developed. Research of open source software can also contribute to existing research in a variety of other fields, including software engineering and computer supported cooperative work. In this study, we examine the behavior of first time code contributors and the community response to their code contributions. In Section 1.2, we discuss how previous studies in open source communities provide insight for existing literature on virtual teams and distributed work. This study examines how new users join these types of communities.

We are interested in how community engagement by a developer and community response affects whether or not a first time code contribution is accepted or not. We focus on code contributions as this shows the first time a user is attempting to participate in a core act of development. Code contributions can also be accepted or rejected, so our analysis allows us to see how these social factors contribute to acceptance within the community.

The rest of this paper is organized as follows. The rest of this chapter provides an overview of terminology specific to GitHub that will be used throughout the paper, as well as a literature review on research of communities of practice and open source, establishing the importance of studying open source software development, situating it within a context of virtual work and computer mediated communication, and reviewing a theoretical basis we use to inform our empirical methods. Chapter 2 describes our data collection and data analysis methods. The results of our experiments are discussed in Chapter 3. Our results differ from previous research in this area. In Chapter 4 we situate our findings in relation to prior research and identify areas for future work.

1.1 Terminology

A software project on the website is referred to as a *repository*. Any user on GitHub can *star* a repository. Users star repositories to be able to easily navigate to them and to receieve updates on activity from the repositories. A repository can be private, meaning that it is only visible to the owner and anyone the owner grants access to, or public, meaning that anyone can view it. Our study focuses on public repositories. If a developer wants to contribute to another one of developer's repositories, he can *fork* the repository, which creates a copy of the project for him to work on. As the developer makes changes to this code, he *commits* his changes. A *commit* is a snapshot of the code at a certain point in time. When the developer is finished, he can submit a *pull request* to the owner of the project. All pull requests for a project are viewable on GitHub, and any user of the site can comment on them. A pull request can have a status of open or closed. A status of open indicates that that owner of the repository has not made a decision about whether or not to include the changes. If the owner of a repository wants to incorporate the changes the developer made, he can *merge* them into the repository. A pull request can be closed without being merged, which means that the changes the developer made were not accepted.

1.2 Related Work

1.2.1 FLOSS Research. Research in the development of free/libre open source software (FLOSS) has grown tremendously in the last several years. Crowston et al. [5] note the importance of understanding FLOSS development as it becomes a major social movement with many volunteers contributing to projects, and many FLOSS projects becoming integral parts of the infrastructure of modern society. Other studies have emphasized the role that FLOSS research can play in improving current existing research of software engineering, particularly as the importance of understanding large scale software systems in science and insustry increases [21].

Existing research approaches FLOSS from many different angles, including motivation of open source developers [9, 16, 22]; governance of open souce projects [12, 19, 20]; and knowledge sharing within FLOSS communities [8, 11, 23]. Our study focuses on the behavior of first time contributors to FLOSS projects and community response to their contributions. We build on previous studies that describe the social processes of community joining [7, 13, 25]. Given the distributed nature of FLOSS development, our findings contribute to current descriptions of virtual work and distributed teams.

Previous studies have used version control histories to verify learning processes of new members in FLOSS projects [13]. GitHub, however, has not been extensively studied as it is a relatively new social platform. Dabbish et al. [6] studied how GitHub as a social application provides transparency and how that transparency affects collaboration and learning. They found that this transparency leads to inferences around commitment, work quality, community significance and personal relevance, which supports collaboration and learning. McDonald and Goggins [18] studed how different communities on GitHub measure success, finding that most developers measure success in the number of contributors and contributor growth. They also found that developers believed the GitHub interface, in particular the use of pull requests, made communities more democratic and transpartent. Choi et al. [3] studied a sample of GitHub based projects to contribute to theories of developer coordination, finding that commits tend to happen in clustered events over time. In all these cases, the social features of GitHub, e.g. the ability to follow other users and view information about them, provide new ways to study social behavior in FLOSS projects. Our study investigates members' participation in group discussions on the site. While previous studies have tried to combine data from mailing lists and version control [7], GitHub provides a centralized location to study communities in which discussion and code contribution all occur in one place. At least with regards to user support, recent research suggests that developers may be moving away from mailing lists to social Q&A sites to respond to user requests for help [24]. By focusing on GitHub data, we contribute to understanding developer behavior on this new social platform.

1.2.2 Communities of Practice. Our study focuses on the behavior of new code contributors and community response to their contributions. We use the theoretical framework of *legitimate peripheral participation* (LPP) [17] in our exploration community joining. LPP describes a process of learning in communities of practice in which newcomers join a community by participating in peripheral tasks and forming relationships to move towards the center of the community. Several studies of FLOSS development have used the LPP framework. Huang and Liu [13] mined version control history to construct developer networks and identify core and peripheral community members. Ducheneaut [7] found a pattern that resembles LPP in his study of contributors to the Python project. Ye and Kishida [26] used LPP to ground their theory of motivation in open source communities. This concept has been explored in other studies of computer mediated communication. In their study on members of Wikipedia, Bryant et al. [2] note that members initially become involved through peripheral activities. These are simple and low risk activities members can take part in to learn more about the community before trying to become major contributors. Similarly, von Krogh et. al [25] from observing open source communities generated the construct of a *joining script*, where each project has a set of tasks for new developers to go through before being accepted into the community. The data we collect in Section 2.1 is based on this theoretical framework. We consider making code changes and submitting a pull request to be a core activity, and commenting on other pull requests to be the primary peripheral activity a user can participate in on GitHub. This follows previous studies that find most users participate in community discussions prior to submitting code [7, 25].

CHAPTER 2

METHODS

2.1 Data Collection

Data was collected using the GitHub API.[1] We used a collection of node.js scripts to collect data from the API to store in a MySQL database.[2] In selecting which repositories to use for our analysis, we started with the top 100 most starred repositories on GitHub. We started with this list with the assumption that they were popular repositories that would be maintained by an active community. From these 100, we manually filtered out certain projects that we expected would follow different development patterns than a typical programming project, for example, collections of configuration files for text editors and shells, collections of icons, etc. We also excluded repositories that were used primarily for demonstration or documentation purposes, such as sample web applications to demonstrate use of a certain web framework. After filtering our intial list of 100, 45 repositories remained in our data set for analysis. We consider only pull requests with a status of closed. This resulted in approximately 44,400 pull requests. We further filtered this data by selecting only the first pull request a user submitted to a repository, leaving 13,383 pull requests. The distribution of these pull requests across repositories ranges from 10 to 1,489, with a median of 210. To find merged pull requests, we first filter all pull requests that are marked as merged by the GitHub API, meaning that the project maintainers used GitHub's merge feature to accept the pull request. In some repositories, project maintainers use a different workflow when accepting pull requests, wherein the code changes are accepted, but it is not reflected as merged on GitHub. In most of these

---

[1] http://developer.github.com/

[2] These scripts are available at http://www.github.com/matthewheston/gh-collector.

cases, there is a standard way of reflecting this in the commit comments, so we use some naive heuristics for identifying these requests by searching commit comments for certain text patterns. For example, in many projects, the project maintainer will manually add the commits from the pull request, and create a new commit with a commit message that follows the pattern "Closes number" where number is the pull request ID on GitHub. Finding merged pull requests using both the status from the GitHub API as well as these text patterns results in finding 5,239, or 39.1% of first pull requests being merged.

2.2 Data Analysis

We divide our analyis into two dimensions: community engagement of developer and community response. Our research questions are presented in Table 2.1, along with a summary of the variables and methods we used.

2.2.1 Community Engagement of Developer. To explore the process of legitimate peripheral participation, we count the total number of pull requests a developer commented on before submitting his own pull request. As discussed in Section 1.2.2, we consider commenting on other pull requests the primary peripheral activity a user can participate in. Previous studies indicate that users participate in technical discussions before submitting code [25]. Pull requests allow users to comment on new features or other changes to the code base. Although there are other peripheral activities a user may engage in on the site, for example, making updates to a project wiki, not all projects use these other features offered by GitHub. Pull requests are used by all the repositories in our data set, so we examine only those comments to quantify participation by a user. This variable is shown in Figure 2.1.

2.2.2 Community Response. In addition to measuring the activity of a developer in the community before submitting a pull request, we are also interested in measuring

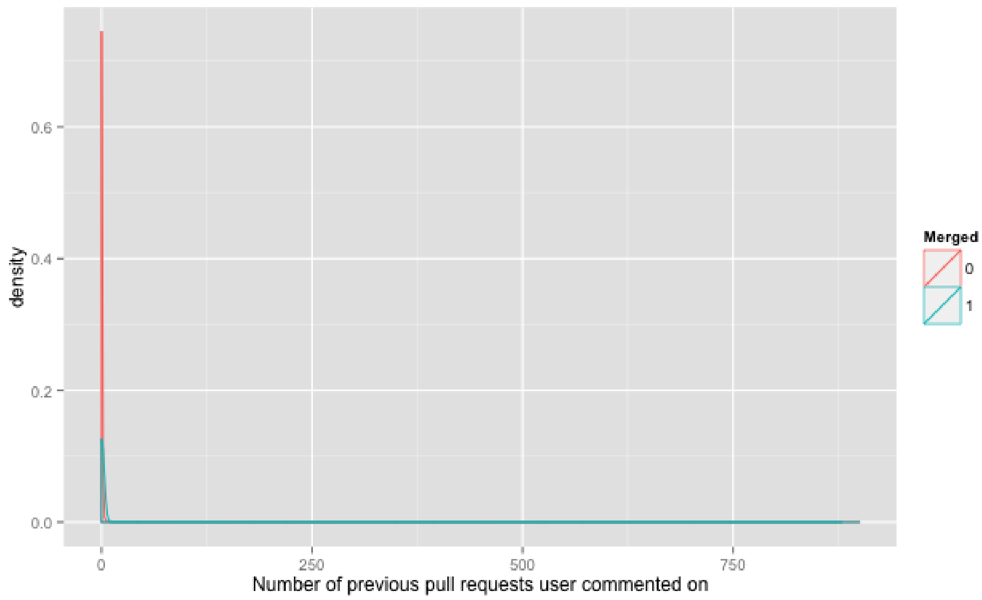

Figure 2.1. User participation density plots.

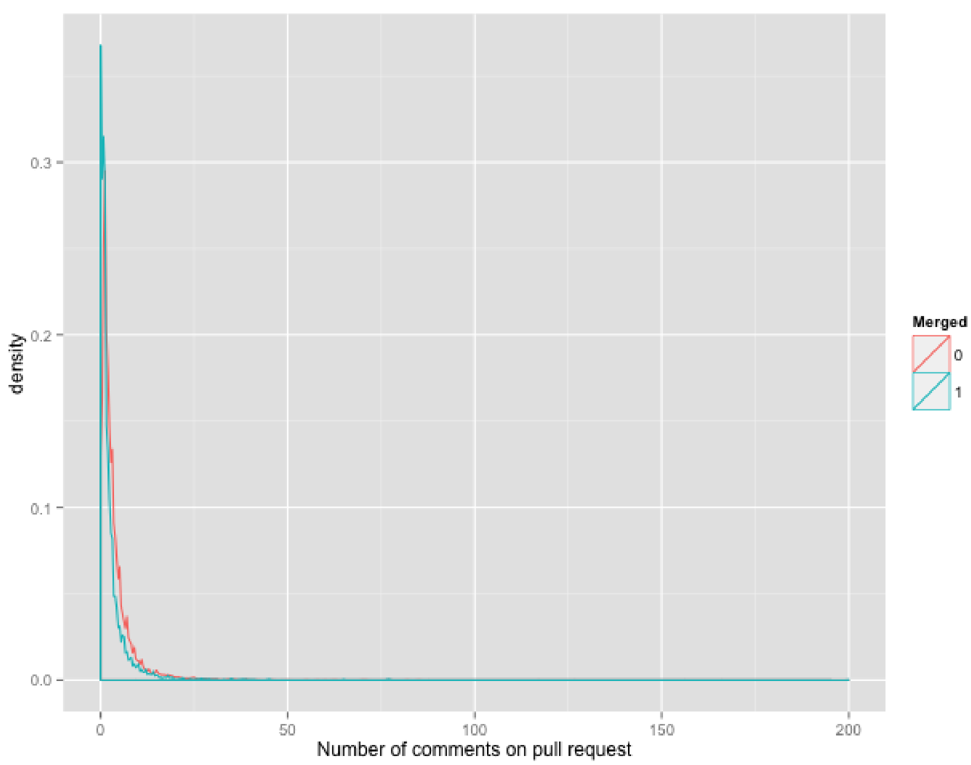

Figure 2.2. Attention pull request receives density plots.

the community response to a given pull request and how that response relates to whether or not a pull request is accepted. We measure this in two ways.

First, we simply count the number of comments on a given pull request. This is used as a basic metric of how much attention a pull request receives. This variable is shown in Figure 2.2.

Our next analysis of community response focuses on the language of the comments on a pull request. To test whether or not the content of these comments is predictive of whether or not a pull request is merged, we collect the comments for each of our first pull requests. We ignore comments made by the user who submitted the pull request, since we are interested in what other users had to say about it. We also ignore the last comment associated with a pull request, since these often will explicitly say whether or not the maintainer is merging the pull request or not. We are interested in whether the type of language used in the discussion of a pull request is predictive of whether or not it is accepted. We ignore pull requests that only have one comment associated with it. This leaves 5,674 pull requests. Of these, 3,811, approximately 67%, were not accepted. We treat the remaining comments associated with the pull request as one document, and convert them into feature vectors representing the count of each unigram and bigram in the documents, and train both a logistic regression and naive bayes classifier using this feature set. The results of testing these classifiers is shown in Table 2.2. The results shown are the result of running 10-fold cross validation.

Table 2.1. Summary of research questions and methods

Research Question	Data	Measure
How does a user engage before submitting their first pull request?	Comments on other pull requests.	We plot the total number of pull requests a user has commented in Figure **??**.
Does community response affect whether or not a pull request is merged?	Number of comments on pull request.	We plot the total number of comments for submitted pull requests in Figure **??**.
	Language in comments on pull request.	We train both a logistic regression and naive bayes classifier after converting comments on submitted pull requests to feature vectors. Results are shown in Table 2.2

Table 2.2. Classifier results

	Logistic Regression	Naive Bayes
Accuracy	69.6%	70.6%
Precision	56.0%	60.3%
Recall	36.1%	30.7%

CHAPTER 3

RESULTS

3.1 Community Engagement of Developer

We see that user participation for the majority of all first pull requests, both merged and not merged, is 0. This indicates that in general, most users are not attempting to engage in the peripheral activity of commenting on other pull requests before submitting their own. The GitHub interface makes it relatively easy for a user to fork a repository, make changes, and submit the changes for consideration. Previous studies on GitHub have shown that the number of contributions did increase for some projects that moved from other hosting options to GitHub [18]. It is possible this interface lowers the barrier of entry for a developer who wants to contribute to a project, and allows them to bypass participating in the joining script described by von Krogh et al. [25].

We also examine these variables for first pull requests by users who later submit another pull request. We want to see whether or not this "no engagement" pattern continues to hold for users who will become active contributors. Our intuition here is that some users might encounter a bug they fix or desire a feature that they implement, and then submit these changes back to repository. They may not comment on other pull requests as they are not interested in becoming long term members of the community, but rather are just interested in submitting a one time patch. Users who do plan on becoming active members, however, may participate in peripheral activities more. Figure 3.1 shows a visualization of the same first pull requests, but only for users who submit at least one other pull request at a later point in our data set, and Figure 3.2 shows the data for users who submit at least 5 more times. Looking at users who submit at least one other time cuts our number of observations from 13,383 to 5,207, indicating that approximately 61% of these pull requests come from users who will not contribute any others. Looking at users who will submit at least 10 more times gives us a total of 1,155 observations.

It is clear that in all these cases, regardless of whether or not they will be continuing to submit other pull requests later, at the time of submitting their first pull request, users are generally not participating in the community. The previous graphs only consider the number of pull requests a user commented on before submitting their first pull request, so we do not capture how users who submit multiple pull request over time comment on other pull requests over time. In Figure 3.3 we plot the total number of others' pull requests that a user commented on by how many pull requests they submitted themselves, considering only users who have submitted at least two pull requests. There is not a strong correlation between these variables (Spearman's $\rho = 0.44$), indicating that users do not necessarily participate in more commenting as they continue to submit more pull requests.

It's worth noting the one extreme outlier present in our data. One user submitted pull request received 200 comments and was submitted by a user who had commented on 900 previous pull requests. This is an interesting case of a project maintainer, who has commit access and wouldn't typically need to submit a pull request to submit changes, creating a pull request for commits related to a major upgrade in the project. By creating a pull request, he was able to document all the changes associated with this change and allow community members to ask questions or comment on the changes. He has a high number of previous comments since he is in charge of accepting pull requests. Due to the nature of this pull request, there is a high number of comments on this pull requests on it, since many other developers are asking questions or voicing their opinions. This is a useful example that demonstrates the different ways pull requests may be used in different projects, and how the way they are used may change depending on the type of user submitting them.

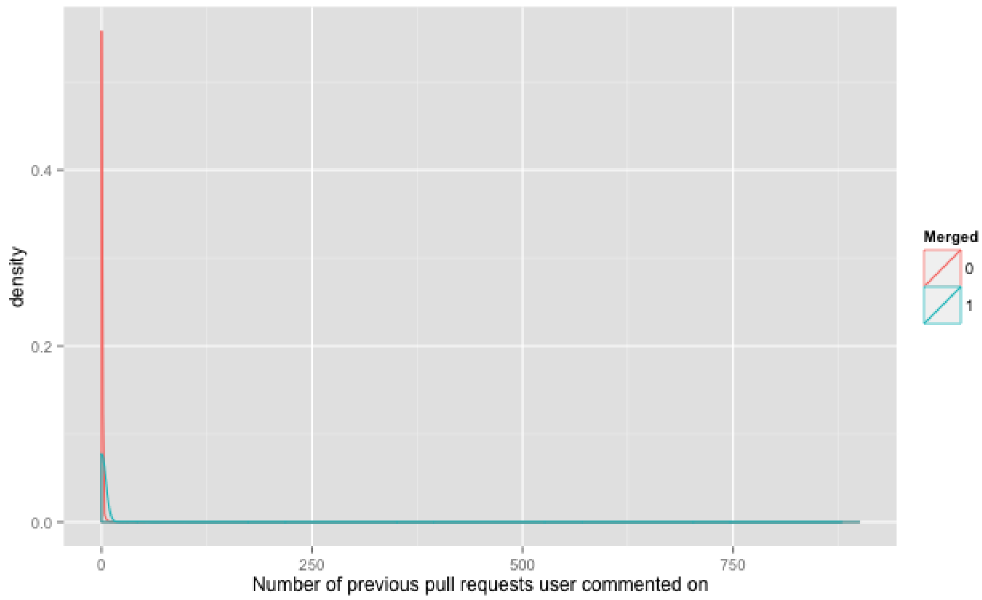

Figure 3.1. User participation density plots for users who submit at least one other pull request in our data set.

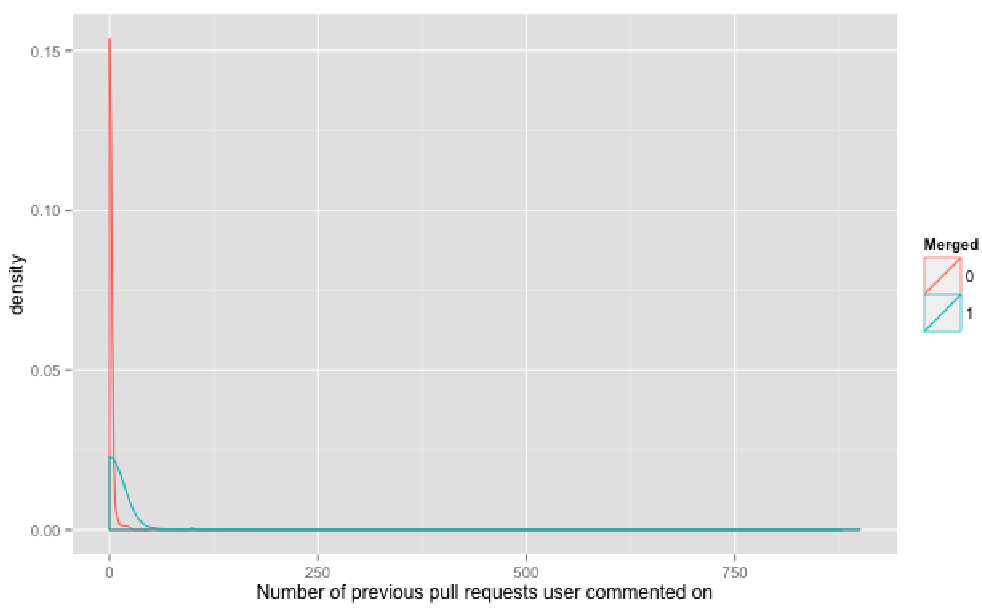

Figure 3.2. User participation and attention a pull request recieves variables for users who submit at least 10 other pull requests in our data set.

3.2 Community Response In Figure 2.1, we see more variance in the number of comments on first pull requests than we did with the number of pull requests users commented on before submitting. However, this variable does not seem to be a good predictor of whether or not a pull request is merged, since both the merged and not merged distributions follow a similar pattern. It seems just viewing the amount of activity a pull request receives is not enough to explain whether or not it gets merged.

Training classifiers using the comment text may help address this problem, as this can capture the valence of the comments, rather than just the raw number. However, the low recall rates we see in Table 2.2 indicate that the text data is not sufficient to distinguish positive cases. We list the top five features for each class in Table 3.1. Despite leaving out the last comment from each comment thread to avoid words that explicitly describe the action being taken, we still see these in our top features. Both "land" and "landed" are used in some repositories when a commit is merged, but not through the GitHub interface to indicate the git commit hash where the pull request commits were merged. We also see "closing" in the negative class, which shows explicit action being taken. Some of our top features do seem fitting. The phrase "lgtm" which stands for "looks good to me" seems indicative of positive feedback from the community. Other examples, however, seem to indicate overfitting, which we see in the top feature for the negative class, the word "the."

Our sample size of 5,674 is relatively small, but it is interesting to note that only 42% of the first pull requests in our data set have more than 1 comment associated with them. Our research question that drove these experiments was how community response affects whether or not a pull request is accepted. However, we see that the majority of pull requests don't attract any community attention.

Table 3.1. Classifier top features

Merged		Not merged	
land	0.817	the	-0.734
landed	0.767	bootstrap	-0.711
fine	0.717	good thanks	-0.682
lgtm	0.709	in	-0.682
it will	0.656	closing	-0.605

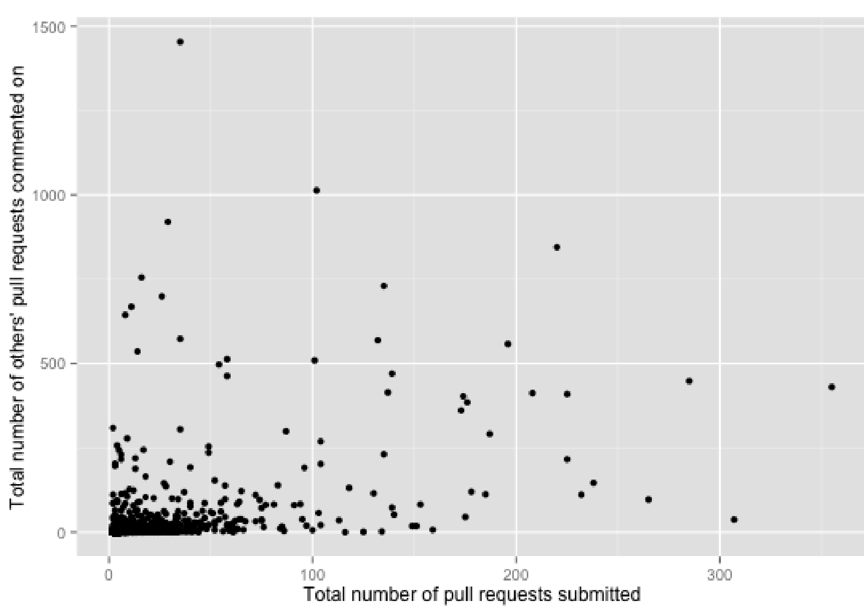

Figure 3.3. Total number of pull requests commented on and total number of pull requests submitted for each user.

CHAPTER 4

CONCLUSION

In this study, we analyzed community engagement and community response of first time code contributors on GitHub. We found that most developers do not engage by participating in discussions on GitHub before submitting code changes. We also fou dthat most submitted pull requests do not attract much community response, and our attempts at measuring community response did not provide good predictors for whether or not a pull request is accepted. Our findings have implications for researchers, open source contributors, and open source project maintainers.

We found that most users do not engage with the community in the way we expected from previous FLOSS and LPP literature. Some reasons for this are discussed below in Section 4.2. Since GitHub is a new platform that encourages new types of social interactions, it can be used as a new source of data to study open source communities. Our study focused on a relatively small number of repositories pulled from the most starred repositories on GitHub. Further work should be done on a larger sample of different types of repositories. It's possible that new types of social websites like GitHub will require new theoretical ways of thinking about distributed work and virtual teams.

The majority of first pull requests in our data set were not accepted, but community engagement is not a good predictor or whether or not they are accepted, and most developers do not engage before submitting a pull request, even if they end up becoming active code contributors. The implication for developers who want to become involved in an open source project on GitHub is that they do not necessarily need to participate in peripheral activities before submitting pull requests. Although our study did not identify which factors do contribute to acceptance of pull request, it's possible that developers should focus more on things like finding relevant issues to fix or features to work on rather than on social interactions.

Many open source projects fail due to insufficient volunteer participation [4], [15]. It is therefore important for project maintainers to continue to attract volunteer developers to keep a project alive. We found that 61% of first pull requests in our data set come from users who will not submit any other pull requests. Project owners should consider ways to convert these one time contributers to regular active contributors.

Our major finding in this study is that, despite previous FLOSS research which did indicate social patterns that follow legitimate peripheral participation framework [7, 13, 26], we generally do not see this pattern in our GitHub data set. Although some developers do leave comments on other pull requests before submitting their own, the majority of developers do not, including those that will continue to submit code changes. There are many reasons this might be the case.

As mentioned in Section 3.1, GitHub's interface makes it fairly easy to submit changes. Users only need to click a button to create a copy of the repository that they can make commits on, and then click another button to submit those commits as a pull request when they are finished. This process may lower barriers to entry for new developers. While previous studies of FLOSS communities have focused on mailing lists and centralized version control repositories, the distributed nature of git may be altering the social patterns that take place in developer networks. This study may suggest we need to alter existing social theory as new interfaces change social interactions.

4.1 Limitations of the Study

This study only focused on data from GitHub, and our notion of community engagement by developers was limited to activity by developers on GitHub. While we have shown that this data is not sufficient to predict whether or not a pull request is merged, further studies should create joint data sets merging GitHub data with data from other sources, such as mailing lists, forums, or chat rooms to test whether or not developers engage with the community using these other platforms, and how those social interactions affect their acceptance in the community. One difficulty in creating these types of data sets is identify merging, the process of matching different logins from different services to the same physical person, but a number of techniques have been proposed to assist in this process [1, 10, 14]. We also focused only used pull request comments to measure community engagement by a developer. Other GitHub features, such as issues and wikis, provide other ways to participate in a repository that we did not cover.

It is also important to note that the GitHub platform is used in different ways across projects. As described in Section 2.1, there was some work required to identify merged pull requests due to the different ways a project maintainer accepts the pull request, e.g. through the GitHub interface or not. In Section 3.1, we discuss an outlier in our data where a user who did not typically contribute code through the pull request mechanism did use this feature in order to allow community feedback and questions. There are many other examples in differences in use of the GitHub platform. Some projects, for example, may require all developers to submit pull requests for the purposes of code review, while others may grant commit access to certain developers that allows them to bypass the pull request mechanism. Although we did account for differences in accepting pull requests, there may be other nuances across projects that affect the data we collected that are harder to detect when working with data at scale.

4.2 Future Work

In examining community engagement and community response, our study was primarily concerned with social factors within open source software development communities. It's possible that looking at only social factors is not enough to understand what contributes to acceptance on GitHub. Static analysis tools may be used in addition to the metrics we used in this study to further explore how the code itself affects whether or not a pull request is accepted.

In studying how communities respond to pull requests, we found in most cases, there was little community response. As mentioned above, attraction of new developers and retention of developers is important to project success. Future work should how community response can affect developer retention in GitHub projects.

In examining the top features for the classifiers we trained in studying community response, we found some issues. Our goal in filtering the text collected was to avoid words that explicitly described action being taken. We still found examples of these types of words in both our positive and negative classes. In the case of the positive (merged) class, the top feature was "land." This feature affects classifier performance since this word is used often in positive cases, but only in a few repositories. Future work studying linguistic features in this way should account for these types of project specific vocabularies.

Finally, to better understand the way that the GitHub interface affects developer behavior will require further qualitative study. While there have been surveys of GitHub developers previously [18], there remains a lot of work left to do in this area. Future quantitative studies should collect larger samples of data and explore new ways of measuring these types of social variables.

4.3 Conclusion

In this study, we measured community engagement and community response of first time code contributors to GitHub. Using data from the pull request feature of GitHub, we studied how developers engage open source communities before contributing code, as well as how community response to these code contributions predicts whether or not they are accepted. We found that most users do not participate in other social GitHub features before submitting code. In attempting to measure community response, we found that most code contributions do not receive much communtiy response, and that both the raw amount of comments on a GitHub pull request as well as the language in those comments were not good predictors of whether or not the pull request is accepted.

BIBLIOGRAPHY

[1] C. Bird, A. Gourley, P. Devanbu, A. Swaminathan, and G. Hsu. Open borders? immigration in open source projects. In *Fourth International Workshop on Mining Software Repositories, 2007. ICSE Workshops MSR '07*, pages 6–6, May 2007.

[2] Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, page 110, New York, NY, USA, 2005. ACM.

[3] Joohee Choi, Junghong Choi, Jae Yun Moon, Jungpil Hahn, and Jinwoo Kim. Herding in open source software development: an exploratory study. In *Proceedings of the 2013 conference on Computer supported cooperative work companion*, CSCW '13, page 129134, New York, NY, USA, 2013. ACM.

[4] Kevin Crowston, Hala Annabi, and James Howison. Defining open source software project success. 2003.

[5] Kevin Crowston, Kangning Wei, James Howison, and Andrea Wiggins. Free/Libre open-source software development: What we know and what we do not know. *ACM Comput. Surv.*, 44(2):7:17:35, March 2008.

[6] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim 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*, CSCW '12, page 12771286, New York, NY, USA, 2012. ACM.

[7] Nicolas Ducheneaut. Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work (CSCW)*, 14(4):323–368, August 2005.

[8] Megan Lee Endres, Steven P. Endres, Sanjib K. Chowdhury, and Intakhab Alam. Tacit knowledge sharing, self-efficacy theory, and application to the open source community. *Journal of Knowledge Management*, 11(3):92–103, June 2007.

[9] Yulin Fang and Derrick Neufeld. Understanding sustained participation in open source software projects. *Journal of Management Information Systems*, 25(4):9–50, April 2009.

[10] Mathieu Goeminne and Tom Mens. A comparison of identity merge algorithms for software repositories. *Science of Computer Programming*, 78(8):971–986, August 2013.

[11] Andrea Hemetsberger and Christian Reinhardt. Collective development in open-source communities: An activity theoretical perspective on successful online collaboration. *Organization Studies*, 30(9):987–1008, September 2009.

[12] Eric von Hippel and Georg von Krogh. Open source software and the Private-Collective innovation model: Issues for organization science. *Organization Science*, 14(2):209–223, April 2003.

[13] Shih-Kun Huang and Kang-min Liu. Mining version histories to verify the learning process of legitimate peripheral participants. In *Proceedings of the 2005 international workshop on Mining software repositories*, MSR '05, page 15, New York, NY, USA, 2005. ACM.

[14] E. Kouters, B. Vasilescu, A. Serebrenik, and M.G.J. van den Brand. Who's who in gnome: Using LSA to merge software repository identities. In *2012 28th IEEE International Conference on Software Maintenance (ICSM)*, pages 592–595, September 2012.

[15] Sandeep Krishnamurthy. Cave or community? an empirical examination of 100 mature open source projects. *First Monday*, 7(6), June 2002.

[16] Karim Lakhani and Robert G. Wolf. Why hackers do what they do: Understanding motivation and effort in Free/Open source software projects. SSRN Scholarly Paper ID 443040, Social Science Research Network, Rochester, NY, September 2003.

[17] Jean Lave and Etienne Wenger. *Situated learning: Legitimate peripheral participation*. Learning in doing: Social, cognitive, and computational perspectives. Cambridge University Press, New York, NY, US, 1991.

[18] Nora McDonald and Sean Goggins. Performance and participation in open source software on GitHub. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 139144, New York, NY, USA, 2013. ACM.

[19] Siobhn OMahony. Guarding the commons: how community managed software projects protect their work. *Research Policy*, 32(7):1179–1198, July 2003.

[20] Siobhn OMahony. The governance of open source initiatives: what does it mean to be community managed? *Journal of Management & Governance*, 11(2):139–150, May 2007.

[21] Walt Scacchi. Free/Open source software development: Recent research results and methods. In Marvin V. Zelkowitz, editor, *Advances in Computers*, volume Volume 69 of *Architectural Issues*, pages 243–295. Elsevier, 2007.

[22] Sonali K. Shah. Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science*, 52(7):1000–1014, July 2006.

[23] Sulayman K. Sowe, Ioannis Stamelos, and Lefteris Angelis. Understanding knowledge sharing activities in free/open source software projects: An empirical study. *Journal of Systems and Software*, 81(3):431–446, March 2008.

[24] Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. How social Q&A sites are changing knowledge sharing in open source software communities. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 342354, New York, NY, USA, 2014. ACM.

[25] Georg von Krogh, Sebastian Spaeth, and Karim R Lakhani. Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, 32(7):1217–1241, July 2003.

[26] Yunwen Ye and K. Kishida. Toward an understanding of the motivation of open source software developers. In *25th International Conference on Software Engineering, 2003. Proceedings*, pages 419–429, May 2003.