The Effect of “Following” on Contributions to Open Source Communities

Research-in-Progress

Mohammadmahdi Moqri
University of Florida
Gainesville, FL 32611
mahdi.moqri@warrington.ufl.edu

Liangfei Qiu
University of Florida
Gainesville, FL 32611
liangfei.qiu@warrington.ufl.edu

Subhajyoti Bandyopadhyay
University of Florida
Gainesville, FL 32611
shubho.bandyopadhyay@warrington.ufl.edu

Ira Horowitz
University of Florida
Gainesville, FL 32611
ira.horowitz@warrington.ufl.edu

Keywords: Open Source Software, Open Source Communities, Online Communities, Social Interaction, Online Following,

Abstract

Online communities have been growing rapidly in the last two decades and at the same time becoming more diverse. Millions of internet users actively contribute to various types of online communities ranging from user generated contents and social media to collaborative content communities, on a daily bases. Although numerous studies have examined members’ motivation to contribute to such communities, the positive effect of social factors has not been unanimously confirmed in different settings. In this study, we estimate the effect of “online following,” a basic form of online social interaction, on members’ contributions in open source software (OSS) communities. This estimation is based on analysis of a large-scale dataset of 4 million online members and their interactions over 7 years, collected from the largest OSS community. The results have implications for online community designers and OSS scholars.
Introduction

Over the last decade, numerous studies have tried to answer one of the most fundamental questions in open source software (OSS) development formulated by Lerner and Tirole (2002): “Why would thousands of top-notch software developers contribute for free to the creation of a public good?” The extant research has identified three categories of motivations for developers to contribute to OSS projects: a) intrinsic motivations, such as a sense of enjoyment or accomplishment in the performance of the task, which are linked to the satisfaction of human needs for autonomy and competence; b) extrinsic motivations, such as reputation, status, and monetary rewards as a result of the outcome of the task, which relate to incentives and reward contingencies; and c) internalized extrinsic motivations, such as contributing to solve a problem of personal use benefit, which are somewhere in between two extremes of internal and external motivations. This study focuses on a subset of extrinsic motivations related to developers’ social interactions. (For a comprehensive review of OSS literature, we direct interested readers to Höst et al., 2011, Crowston et al., 2012, and Aksulu and Wade, 2010)

Several theoretical studies, rooted in behavioral science, social psychology, have suggested that social factors such as reputation and identification have substantial effect on developers’ contribution to OSS communities (Osterloh and Rota, 2007; Roberts et al. 2006). However, very few large scale empirical studies have been conducted to verify or quantify the effect of these factors on contribution. Moreover, the results of these studies are mixed. For example, Grewal et al.’s (2006) results suggest a significant relationship between developers’ network centrality and their contribution, in their constructed network. However, the more recent study of Singh (2010) does not find any significant effect confirming the previous results. We believe the lack of an empirical agreement in such studies arises for two reasons. First, the social factor measures in such studies are almost always approximated by other measures. For example, in the two above mentioned studies, two developers are assumed to be socially connected if they both have contributed to the same project at any point in time. Although in the absence of the true social factors, using such proxies are the only viable option, they are not always accurate measures of those constructs. Second, OSS projects are extremely diverse: software projects range from mobile applications to desktop software to online games. Similar to commercial software packages, OSS are designed and developed for different purposes and require different sets of skills and resources. Therefore, a relationship or pattern identified in one study based on one sample of homogenous OSS projects may not hold in other settings.

Both the limitations of approximation and sampling are primarily caused by the lack of data. Until recently, a large scale dataset on OSS environment which a) contains information on developers direct social interactions, b) span over a heterogeneous set of OSS projects and settings, and c) is well maintained and publicly accessible has not been available to OSS scholars.

In this study, using a unique longitudinal data set containing information on over 4 million OSS developers and their social interactions over 7 years, we empirically investigate the effect of online social interaction in OSS developers’ contribution patterns. Specifically, we wish to examine the effect of reputation social ties on developers’ activities over time.

Context

We obtain data from GitHub.com (GitHub). GitHub is currently the largest open source community and code host in the world, with more than 4 million software developers working on more than 9 million OSS projects. Desktop software, web development, mobile applications, games, and operating systems are among the various types of OSS projects hosted on this platform. These projects are developed using various programming languages such as Java, C#, Python, Php, Javascript, Ruby. Many well-known OSS project such as Linux, Bootstrap, JQuery, and Firefox for iOS are mainly or partially hosted on GitHub.

GitHub uses a distributed version control system. Version control, or revision control, is a system that records changes over time so that any specific state (version) can be recalled later. In the simplest case, version control is analogous to Wikipedia’s “page history” that enables everyone to track each change to the page. GitHub uses the “fork & pull” model for collaboration. A developer may fork an existing OSS project to create a personal copy of the project without requiring access to be granted to the project. Any changes or “commits” to the project is first performed on this personal copy. The developer may then send a pull request to the project maintainer to merge the changes into the source project.
For example, as shown in Figure 1, the 4,756 developers who currently (March 15th, 2015) contribute to the Linux Kernel project hosted on GitHub have submitted 506,519 commits to the project (https://github.com/torvalds/linux). These commits include adding, modifying, or deleting pieces of code or explanatory comments. Since version control requires all information about each contribution to each project be recorded, the detail information about each commit is publicly available on the platform.

In addition to developers’ contributions, GitHub collects all information on developers’ interactions such as “following” a developer. As shown in Figure 2, for each developer on the platform, a profile page exists which presents three types of information: 1) general information such as developer’s name, location, company and personal website (left side of the page); 2) information on each public activity that a developer performs, including his contributions to the projects, from the time he registers on the website (middle of the page); and 3) the developer’s social interactions including total number of “followers,” “following,” and “stared” by other developers (lower left of the page). The detail information on each social interaction with timestamps can be retrieved using GitHub’s application program interface (API). This host of information provides a valuable resource for research on OSS.
Social Interaction and Contribution to Electronic Communities

The effect of members’ social interactions on their contribution to electronic communities has been of great interest to scholars and practitioners in different contexts. In this section, we briefly review some pioneering studies on the effect of members’ social interactions on their contributions to organizational electronic communities (Kankanhalli et al. 2005; Wasko and Faraj 2005; Wasko and Faraj 2000), online/virtual communities (Ma and Agarwal 2007; Chiu et al. 2006), user generated content (Goes and Lin 2014, Chen et al. 2010), and collaborative contents communities such as open content (Okoli and Oh 2007) and OSS communities (Osterloh and Rota, 2007; Roberts et al. 2006). Figure 3 summarizes these examples of electronic communities.
Intraorganizational and interorganizational electronic communities have been particularly of interest to organizations as a means of communication, knowledge sharing, and knowledge acquisition for employees. Electronic (online) networks of practice is an example of such interorganizational communities which provides a forum where people with a shared interest voluntarily exchange ideas and solutions to common problems. Similarly, electronic knowledge repositories (EKR) provide internal stores for knowledge sharing, project reviews, case studies, lessons learned, and best practices among the organization’s employees. Several empirical studies have been conducted to examine potential factors, including social factors, in members’ contributions to these platforms. For example, Wasko and Faraj (2000) and Wasko and Faraj (2005) conducted multiple surveys on computer professional and legal professional online networks of practice to answer why individuals help strangers in these electronic networks. The surveys’ results suggest that people contribute to online networks of practice more often and with a higher quality when they perceive that it enhances their professional reputations and standing or status. Similarly, Kankanhalli et al. (2005) survey 150 knowledge management executives, covering 7 industries, to find important factors affecting knowledge contribution to organizational knowledge repositories. Their findings, in contrast to Wasko and Faraj’s (2000 and 2005) results, suggest that reputation is not a significant indicator of contribution to the knowledge repositories.

Aside from their importance to organizations, user-generated contents are fundamental to online and virtual communities in the context of social communities. Several studies have been recently conducted to identify the motivations to share and contribute to online communities. For example, Ma and Agarwal (2007) survey 650 members of two different online communities (an emotional support community and a sport car owners) to find factors influencing and facilitating knowledge sharing. Similarly, Chiu et al. (2006) survey 310 members of a virtual community of computer experts to identify the motivation underlying individuals’ contribution to such communities. Both of these studies support the hypothesis that social factors such as reputation significantly and positively affect quantity of knowledge sharing. Social ties were identified as insignificant factor in the “quality” of contribution by Chiu et al. (2006).

Online communities are not always centered on sharing technical knowledge and expertise or providing emotional support. Increasingly, online product review communities are becoming more popular and more social. Such websites create value by presenting an aggregated opinion of consumers on the

**Figure 3. Different electronic communities and members’ motivations to contribute**
products they review. Gathering a higher number of reviews is of a great interest to such platforms. Many websites encourage social interactions among users in order to solicit more reviews (Goes et al. 2014). Goes et al. (2014) empirically study the contributions and social interactions of 92,094 members of a product review website. They observe that receiving more incoming ties increases the number of product review articles that a user posts but with a decreasing rate. Similarly, Chen et al.’s (2010) natural field experiment involving 398 members on a movie review website shows that revealing social information such as social comparison leads to 530 percent increase in the number of monthly movie ratings in one group but a 62 percent decrease in another group.

A different type of UGC, social media platforms, is also built on the basis of creation and exchange of user-generated contents. Such platforms typically revolve around sharing a broad set of contents particularly to an individual's “network” or “followers”. Toubia and Stephen’s (2013) conducted an interesting field experiment on Twitter to examine the effect of social (image-related) factors on members' contribution. They created 100 fake Twitter accounts and gradually added followers to a group of real members over a 50-day period. They observed that although the level of activity was increased in the treated group, the effect was not statistically significant.

The focus of this study is on OSS communities, a type of collaborative content communities where digital contents are created by a group of individuals working together. In our classification of electronic communities, we place OSS communities next to open content communities (see figure 3). Open contents are defined as creative work that others can copy or modify (Wiley, 1998). The contents in both types of communities are created by voluntarily collaboration of knowledgeable/skilled individuals and are managed by version control systems. Currently, Wikipedia is the largest open content community. As empirically shown by Okoli and Oh (2007), social ties have a significant effect on users’ performance on Wikipedia. In this work, we seek to investigate if such effects exist in OSS communities. In other words, we seek to answer whether OSS development has become a social activity and to what extent.

## Data Source

The data was collected from the GitHub’s website (www.github.com). For each OSS project, not only all the source codes are publicly available, but also information on every change to the project, every issue reported and/or solved, and every discussion around the project is recorded with their timestamps. In addition, for each developer, three sets of information are recorded: 1) general information such as developer’s location, company and personal website; 2) information on every public activity that a developer performs, including his contributions to the projects; and 3) the developer’s social interactions including “following” and being “followed” by other developers. The exact time for each user's development activity or social interaction is recorded and stored on the platform.

All these recorded information is publicly available, which provides an invaluable resource for researchers. In particular, what makes GitHub uniquely ideal for the purpose of this study is the information on the exact time each “tie” (i.e. action of following a developer) has been created between developers, which has been recently released by GitHub. Using this piece of information, we were able to replicate the dynamic social network of the developer from the time the first developer joined the website (in October 2007) until the present (March 2015). An archive of all public activities on GitHub can be found at https://www.githubarchive.org/.

The empirical analysis is performed using the data covering all activities on the platform from January 2012 to November 2013. The level of social interactions on the platform sharply declined (by around 75%) during December 2013. This large decrease was likely to be caused by a structural change in the design of the platform. To avoid estimation bias due to this shock, the study period is chosen to include all the activities until December 2013.

## Hypotheses, Model and Results

### Hypothesis

We expect to observe an increase in a developer's activities when he obtains more followers. Each act of “following” could be an indicator of a peer developer’s interest in the work of the developer who is being “followed.” Obtaining more “followers” over time might increase the developer’s reputation or recognition
Effect of Following on Contribution to OS Communities

on the platform. However, as the number of followers increases, the marginal effect of an obtaining a new follower should decrease since this increase is less noticeable (e.g. an increase from 92 to 93 followers is less noticeable compared to an increase from 2 to 3 followers). Hence we formalize our hypotheses as follow.

H1: Receiving more incoming ties should increase the level of a developer’s contribution to open source community.

H2: The marginal effect of more incoming ties should be decreasing

Model

Developers on OSS contribute to projects in different ways. Writing a piece of code, revising a code written by other developers, reporting an issue in a project, and resolving a reported issue are common types of contribution. In GitHub, a contribution to a project is called a “commit”. Consistent with previous literature on OSS development (Grewal et al. 2006; Singh 2010), we consider number of commits in a specific time period as an indicator of a developer’s performance.

Aggregating developers’ activities over month, we construct a panel where each unit of observation is a developer and each time period is one calendar month. We include the user’s number of commits as the dependent variable and the number of new individuals that the developer follows and the number of new individuals that follow the developer during each month as main dependent variables. In addition, for each developer, the total number of his followers, the total number of the people followed by him, the number of months he had been in the system, and a binary variable for affiliation with a company are included in the model as control variables.

A Fixed Effect regression model (see Equation 1) is employed since some unobserved factors such as developers’ creativity or ability might be directly correlated with the independent variable.

\[
\text{Commit}_{it} = \alpha_0 + \alpha_1 \text{Commit}_{i(t-1)} + \alpha_2 \text{NewFollowers}_i(t) + \alpha_3 \text{NewFollowers}_i(t-2) + \alpha_4 \text{NewFollowing}_i(t) + \alpha_5 \text{NewFollowers}_i(t-1) + \alpha_6 \text{NewFollowers}_i(t-1)^2 + \alpha_7 \text{NewFollowers}_i(t-1) + \alpha_8 \text{TotalFollowers}_i + \alpha_9 \text{TotalFollowing}_i + \alpha_{10} \text{Tenure}_i + \alpha_{11} \text{Company}_i + v_i + \epsilon_{it}
\] (1)

The lagged of the dependent variable is included in the model to correct for possible bias in coefficients’ estimates due to an autocorrelation in observations. In addition, a generalized least-squares regression (robust regression) method is used to account for possible heteroscedasticity. Since in our constructed panel, the unobserved panel-level effects might correlate with the lagged dependent variables” which results in biased estimates, we also use Arellano–Bond (a Generalized Method of Moments) method to estimate the unbiased coefficients.

Results

The results show a statistically significant positive effect of the number of new followers (incoming ties) in the previous month on the level of contribution in the current month. On average, obtaining one new follower corresponds to an increase in contribution by .69 commits (based on Arellano Bond regression results). Note that the mean contribution in the community is 3.4 commits per month. Also a negative coefficient of the squared of lagged new followers’ suggests this effect is marginally decreasing.

In addition, the results suggest a statistically significant positive effect of the number of new following (outgoing ties) in the previous month. This might be due to a peer pressure effect where developers are motivated to increase their contributions when they observe other developers contributions more often (as a consequence of following more developers).
<table>
<thead>
<tr>
<th>INDEPENDENT VARIABLES</th>
<th>Random Effect</th>
<th>Fixed Effect</th>
<th>Arellano Bond</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contribution</td>
<td>Contribution</td>
<td>Contribution</td>
</tr>
<tr>
<td>Lagged contribution</td>
<td>0.035</td>
<td>0.009</td>
<td>-0.085</td>
</tr>
<tr>
<td></td>
<td>(1.19)</td>
<td>(1.14)</td>
<td>(9.08)**</td>
</tr>
<tr>
<td>Lagged new followers</td>
<td>0.958</td>
<td>0.812</td>
<td>0.691</td>
</tr>
<tr>
<td></td>
<td>(7.47)**</td>
<td>(7.39)**</td>
<td>(6.63)**</td>
</tr>
<tr>
<td>2nd lagged new followers</td>
<td>0.423</td>
<td>0.086</td>
<td>-0.048</td>
</tr>
<tr>
<td></td>
<td>(3.74)**</td>
<td>(1.30)</td>
<td>(0.67)</td>
</tr>
<tr>
<td>Lagged new following</td>
<td>0.586</td>
<td>0.468</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>(3.51)**</td>
<td>(5.78)**</td>
<td>(4.70)**</td>
</tr>
<tr>
<td>Lagged new followers squared</td>
<td>-0.001</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(2.91)**</td>
<td>(6.52)**</td>
<td>(7.10)**</td>
</tr>
<tr>
<td>2nd lagged new followers squared</td>
<td>-0.001</td>
<td>-0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(1.65)</td>
<td>(1.94)</td>
<td>(1.19)</td>
</tr>
<tr>
<td>Lagged new following squared</td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(4.05)**</td>
<td>(5.73)**</td>
<td>(4.46)**</td>
</tr>
<tr>
<td>Total followers</td>
<td>0.003</td>
<td>_</td>
<td>7.611</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
<td></td>
<td>(32.88)**</td>
</tr>
<tr>
<td>Total following</td>
<td>0.031</td>
<td>_</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(2.39)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tenure</td>
<td>0.007</td>
<td>-1.445</td>
<td>-5.834</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
<td>(60.15)**</td>
<td>(42.22)**</td>
</tr>
<tr>
<td>Company binary</td>
<td>5.288</td>
<td>_</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(14.63)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>20.096</td>
<td>58.758</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(26.66)**</td>
<td>(78.63)**</td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>1,489,631</td>
<td>1,489,631</td>
<td>998,264</td>
</tr>
</tbody>
</table>

Table 1. Random Effect, Fixed Effect, and Arellano Models

Discussion

Our empirical analysis of contribution in OSS communities suggests that social factors significantly affect developers’ contribution patterns. Therefore, incorporating and highlighting social features, such as the ability to follow and observe other developers’ activities, could be an effective approach to boost individuals’ contributions using non-monetary incentives.

The main limitation of the current study is that the random effect, fixed effect, and Arellano Bond models do not take into account a possibility of a reverse effect—the effect of developers’ contribution on the number of new followers. We are currently extending this study by developing a Panel Vector Autoregression (Panel VAR) model to simultaneously estimate the effect of contribution and new followers on each other. Overall, the Panel VAR model’s results seem to agree with the reported results; however, the estimated effects seem to be slightly smaller.
References


