2010

Success and Abandonment in Open Source Commons: Selected Findings from an Empirical Study of Sourceforge.net Projects

Charles M Schweik, *University of Massachusetts - Amherst*
Robert English
Qimti Paienjton, *University of Massachusetts - Amherst*
Sandy Haire, *University of Massachusetts - Amherst*

Success and Abandonment in Open Source Commons: Selected Findings from an Empirical Study of Sourceforge.net Projects

Charles Schweik[123], Bob English[34], Qimti Paienjton[2] and Sandy Haire[1]

1 Department of Natural Resources Conservation, University of Massachusetts, Amherst. WWW homepage: http://nrc.umass/edu
2 Center for Public Policy and Administration, University of Massachusetts, Amherst.WWW home page: http://masspolicy.org
3 National Center for Digital Government, University of Massachusetts, Amherst. WWW home page: http://ncdg.org
4 Daystar Consulting. WWW home page: http://edaystar.com

Paper submitted for presentation at the 2nd workshop on Building Sustainable Open Source Communities (OSCOMM 2010),
University of Notre Dame, June 2, 2010

Abstract. Some open source software collaborations are sustained over long periods of time and across several versions of a software product, while others become abandoned even before the first version of the product has been developed. In this study, we identify factors that might be responsible for one or the other of these collaborative trajectories. We examine 107,747 open source software projects hosted on Sourceforge.net in August 2006 using data available through the FLOSSmole Project. We employ Classification and Regression Tree modeling and Random Forests statistical approaches to begin to establish an understanding of how various project attributes, especially physical and community ones, contribute to project success or abandonment. We find that factors associated with success and abandonment differ for projects in the early stage of development (pre-first release) compared to projects that have had a first release, and that product utility, project vision, leadership, and group-size are associated with success in open source collaborations. We also find that successful open source projects exist across all types of software and not simply in areas associated with the open source "movement." Other evidence suggests that Sourceforge.net may play an important role in "intellectual match-making."

1. Introduction

This paper presents selected results from a 5-year study funded by the U.S. National Science Foundation. The overarching research question of the study is: What factors lead to success or abandonment of open source software (OSS) projects?

2. Theoretical Factors Related to Success and Abandonment

In Schweik and English (2007) we categorize open source software (OSS) projects into two broad, longitudinally-distinct categories: *Initiation* stage (pre-first public release) and *Growth* stage (post-first public release). This distinction is significant because OSS project success and abandonment could occur at either of these stages, with potentially different implications for the project, and possibly driven by different factors. For example, as we measure it, success in the Initiation stage is achieved when the project posts its first public release. Success in the Growth stage, as we define it, is achieved when a project has at least three "meaningful" releases and exhibits some ongoing development and usage activity. English and Schweik (2007) describe the construction of this dependent variable in detail.

Drawing on theoretical and empirical literature in information systems, software engineering, environmental commons management, virtual teams and other areas, we identified factors thought to influence project success in these two stages. To structure and guide our efforts, we utilize an organizing theoretical structure commonly referred to as the *Institutional Analysis and Development* (IAD) *framework* (Ostrom, 2005). At any point in a project's life cycle, its members (programmers, users), make decisions about how they participate based on three sets of project attributes: physical, community, and institutional (See Schweik, 2005 for more details).

Physical attributes are clusters of variables that are related to the software itself, such as the software's properties (e.g., programming language utilized, the operating system(s) it runs on, database used, and other such components). *Community attributes* are variables that are human-related factors found in open source software projects, such as the degree of user involvement, the size of the development team, whether the project is financed, relationships between developers (e.g., whether they meet face to face) and characteristics of team leadership. *Institutional attributes* are variables related to the management and governance of an open source software project, such as the rules and procedures that govern how the team members work together. Figure 1 provides a graphical summary of these types of variables and those with asterisks are variables we could operationalize using SF project metadata. We established over 20 testable hypotheses (not shown) on how these relate to project success or abandonment.[1] What follows are descriptions of methods and (for brevity) selected results.

[1] This paper is based on sections of a book manuscript we are completing that is tentatively entitled "Success and Abandonment of Open Source Commons." Space limitations keep us from stating specific hypotheses but will be described fully in this forthcoming book.

3. Data

As most readers are aware, SF is a web-based open source project hosting site that provides collaborative tools to support software development. SF project metadata include: date of initial registration; number of developers; intended audience; programming language used; forum post archives; bug reporting information; information on changes to code repository; release dates for software; number of software downloads; and other variables. To investigate the research question posed in the Introduction, we utilize data collected from August through October 2006 on 107,747 OSS projects hosted on the open source hosting site Sourceforge.net (SF henceforth). We combined SF project data gathered by the FLOSSmole project (Howison et al., 2006) with other SF data we "crawled" ourselves.

Dependent Variable. Prior research captures multiple dimensions of OSS "success." For example, Crowston et al., (2003) reviewed traditional information system success concepts, such as: *system or code quality*, *user satisfaction*, *use*, and *individual and organizational impacts*. Other measures of FOSS success and abandonment have been used as well, including: (1) *Project life or death* (e.g., Robles et al., 2003) and (2) *Project popularity*, using web search engine results (Weiss, 2005).

We conceptualized and operationalized measures of success and abandonment for each of our two longitudinal stages, Initiation and Growth. We combined the *use* and *popularity* aspects proposed by Crowston et al. (2003) and Weiss (2005) with *project life and death* metrics. Ours are conservative measures, because they define both projects that are developed and used by very small, specialized groups of people (for example, projects in Bioinformatics) and projects with a large number of potential users and developers as success situations. The operationalization and validation of this success and abandonment measure took us more than a year to develop and is fully described in English and Schweik (2007). It utilizes SF project metadata such as lifespan (calculated as the difference between project start date and data collection date), number of releases, first release date, last release date, and number of downloads.

Independent Variables. As mentioned above, theoretical independent variables that we could operationalize using SF metadata are designated with an asterisk in Figure 1, and consist of both numerical and categorical data types. Numerical data include: the number of developers on the project – associated with our theoretical interest in the effect of 'group-size' on success, the number of "Bug Tracker" requests and forum posts – capturing both the influence of the "collaborative infrastructure used" and "project utility," and lastly, a count of Page Visits to any page on the project's SF website--also capturing a measure of 'product utility.'

In addition to these numerical independent variables, we also operationalized seven categorical independent variable groups using FLOSSmole data. At the time of our data collection, project administrators had the option of selecting over five hundred categorical metadata options. For the sake of parsimony, we consolidated the five hundred options down to fifty-four aggregated subcategories. Each of these new

subcategories is a separate independent variable in our analysis existing within one or another of the seven broader "groups." These independent variable groups include Intended Audience, Operating System, Programming Language, User Interface, Database Environment, Project Topic, and Project License. Many are associated with the 'product utility' hypotheses, and some address the 'developer attributes and motivations' hypotheses. Project license (GPL or non-GPL) is the one variable capturing one Institutional measure. We also constructed an independent variable called the *Project Information Index*, which is simply the total number of categorical variables that a project's administrator had selected to describe the project. Many SF projects have few or no categories selected (suggesting the project might be something more trivial like an assignment for a college course) while others have many categories selected (possibly signaling a project leader who is serious about the project). In sum, the final dataset for analysis included information on fifty-nine independent variables: five numerical independent variables and seven categorical independent variable "groups" containing fifty-four categorical variables.

4. Analysis

In order to determine which factors were associated with success and abandonment in open source commons, we needed a statistical technique that would efficiently divide our data into two groups (successful projects and abandoned projects) based on one or more of our fifty-nine independent variables. We chose a non-parametric approach called "Classification and Regression Trees" (CART) which has among its advantages the potential to work with both categorical and numerical variables, and the ability to model complex interactions between variables. Figure 2 shows a tree generated using our Growth Stage data and provides an example of the output produced using this statistical method.

CART models are built in three steps. First, the data is separated into binary subsets that maximize correct classification of the dependent variable based on values of a selected independent variable. Second, a large, highly accurate tree is built through a process of "recursive partitioning" that is based on any of the independent variables. Third, the tree is then "pruned" back (leaves and branches are reduced) to a size that maximizes classification accuracy while at the same time producing a more parsimonious result. Partitioning is done through the evaluation of a statistical measure called the "Gini index" that maximizes correct classification in the subsets. Pruning is based on a cost-complexity statistic that protects against "over-fitting" the data and producing an overly complicated tree that might not characterize open source project success and abandonment in a useful, interpretable way. For more information on how CART models work, see De'ath and Fabricius (2000).

Initial attempts to use Classification Tree to analyze the complete 107,747 project dataset (separated into two subsets for Initiation and Growth Stage projects) produced an unusual problem: the computational requirements to partition such a large dataset were too high for our relatively high-end computer. To circumvent this issue, we used multiple random samples drawn from the data in order to develop trees for each Stage. Through a systematic investigation, we determined that a sample size of 1000 or greater tended to include enough variability and enough

replicates to produce interpretable and fairly accurate results in most cases. However, the use of these highly variable random data subsets left open the possibility that important variables would be inconsistent across trees – a problem often caused when "surrogate" variables exist (see D'eath and Fabricus, 2000). In order to develop a robust representation of the relative importance of the independent variables, we employed the "Random Forests" classification method which fits many Classification Trees to a data set and then combines the predictions from all these trees (see Cutler et al. (2007) for more information). The Random Forests approach produces as part of its output a "Variable Importance Plot" (VIP) to graphically represent the ranking of the importance of our independent variables. Such plots were produced for both the Initiation and the Growth stage data (Figures 3 and 4).

5. Results

The Project Information Index (or PII), a measure we created by totaling up the number of subcategories selected for all SF categorical variables, and one that captures the concepts of vision and leadership, is the most important variable for discriminating between success and abandonment in the Initiation Stage (Figure 3). Alternatively, for the Growth Stage, the most important variables distinguishing between success and abandonment were Page Visits and Downloads (Figure 4). This contrast between the two stages is illustrative of our overall results. This leads to our first finding:

Finding 1. *Factors associated with success and abandonment differ between pre-first release and post-first release projects.*

While we have a number of insights about success and abandonment in both Initiation and Growth stages, due to space considerations, we focus the remainder of our discussion on key results from the Growth stage component of our study.

As mentioned earlier, Page Visits and Downloads are the most important discriminators of success and abandonment in the Growth stage (Figure 4). As an example of the amount of distinguishing power displayed by these variables, consider a classification tree that we developed using only *complete observations* (i.e. a data subset comprised of only those projects that had at least one value or subcategory selected for each of the seven variable groups). With n=2052, this 2-leaf tree (not shown due to space limitations), split on the basis of 3026 Page Visits, and classifies projects as AG or SG with a 77% accuracy rate. Since the Page Visits variable captures the concepts of product utility and user base, these results lead us to these next two findings:

Finding 2. *Clear utility of the project software for a fairly large number of end users, improves the likelihood of success in the Growth Stage*.

Finding 3. *Larger user communities improve the likelihood of success in the Growth Stage.*

These conclusions are further supported by the fact that Tracker Reports and Forum Posts are also among the most important discriminator variables in the Growth Stage. Not only do these two variables reflect product utility and size/effort of the end user community, but also signal leadership and the use of collaborative infrastructure on the part of developers.

Further analysis revealed two other major findings for Growth Stage projects:

Finding 4. *Slightly larger teams are a causal factor for successful Growth Stage projects, even though, on average, they are still very small (2-3 people) groups.*

In the VIP shown in Figure 4, the "Developers" variable is the next most important variable after Page Visits, Downloads and Tracker Reports. Based on this large body of SF project data, we have strong statistical evidence that shows that successful Growth Stage projects gain slightly larger developer teams (SF project "members") on average *before* they become successful. Therefore, we can make a reasonable case that larger development teams are a causal factor in Growth Stage success.

Finding 5. *Success in the Growth Stage is not restricted to particular software categories.*

Our analysis shows convincingly that none of the SF categorical variables featured prominently in our classification tree analysis, which implies that factors such as Programming Language, Operating System, and Project License (GPL versus non-GPL) are not useful in discriminating between successful and abandoned Growth Stage projects.

So one might ask, what do these findings suggest for future open source projects?

Finding 2 (utility) suggests that choosing to develop software that has a large body of *potential* users is probably more likely to result in success than developing software for an inherently small audience. Finding 3 (larger communities) suggests that having leaders on the project who are capable in building community (e.g., possesses good communication skills, for example) and making good use of collaborative infrastructure to keep the user community engaged, will be more likely to succeed in the Growth Stage.

Although Findings 2 and 3 may seem obvious, recall that our dependent variable defines success in the Growth Stage in terms of producing multiple releases of software that is useful to *some* number of users. While it is obvious that many successful open source projects have a large developer and/or user community, it could well have been that a *majority* of successful open source projects consisted of small groups of academics or others without much evidence of a significant community or user base. In addition, these two findings provide strong statistical support to assumptions made about open source and have, to our knowledge, not been shown through empirical study. This may be some of the first empirical and strong statistical evidence that reveals these relationships in OSS projects.

Finding 4 (larger development teams) suggests that projects will have a higher likelihood of success, in terms of continued development, if they are able to find an additional developer to join the team. This too might seem trivial at the outset; however, this finding underscores the potential importance of hosting sites like SF that allow people in potentially distant parts of the world to find projects of interest and to connect to others with a similar passions, interests, in a particular project and the skills to collaborate. In further work related to our book project (Schweik and English, in preparation) we have more evidence to support this statement.

Finally, Finding 5 (success found in all categories) provides strong evidence suggesting that open source is broadening in its scope and reach. This provides statistical evidence that open source as a "movement" or "cause" may be diminishing in its importance as a motivator for open source software collaboration, and that open source is now being driven by the broader "ecosystem" of not just volunteers, but also interests by firms, nonprofits and government agencies.

6. Limitations and Conclusions

A limitation of the work is that the SF project metadata only correspond to some theoretical variables thought to drive projects toward success or abandonment. Most variables are *physical attributes* of OSS projects. A few map to *community attributes* (such as the Project Information Index capturing a measure of leadership). The only *institutional attribute* captured in the SF metadata is the GPL/non-GPL licensing variable. The primary reason we explored SF data alone was because we think of historical SF data repositories like FLOSSMole (Howison, et al. 2006) and the SF Research Data Archive (Antwerp and Madey, 2008) as "remote sensors" of OSS. That is, like the satellites that take images of the Earth, these repositories take longitudinal snapshots of OSS projects. From that standpoint, it is important to investigate what can be learned from monitoring them alone. However, knowing that this is an incomplete analysis, in the Fall of 2009 we conducted a survey of nearly 1500 SF developers to gather information on additional community and institutional factors. The findings will be presented a book-length manuscript that we hope will appear in 2011 (Schweik and English, in preparation).

Acknowledgements

Support for this work was provided by a grant from the U.S. National Science Foundation (NSFIIS 0447623). The findings, recommendations and opinions expressed are those of the authors and do not necessarily reflect the views of the funding agency. Special thanks go to Megan Conklin, Kevin Crowston and the FLOSSmole project (http://ossmole .sourceforge.net/) for making their Sourceforge data available, and to Megan for early assistance with their data. Of course, any mistakes are our responsibility alone.

References

Crowston, K., H. Annabi, & J. Howison. 2003. "Defining Open Source Project Success," In Proceedings of the 24th Int.l Conf. on Info. Systems, ICIS, Seattle.

Cutler, D.R., T. Edwards, K. Beard, A. Cutler, K. Hess, J. Gibson, & J. Lawler. 2007. “Random Forests for Classification in Ecology.” *Ecology* 88(11):2783-2792.

De'ath, G. & K.E. Fabricius. 2000. “Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis.” *Ecology*. 81(11): 3178-3192.

English, R. & C.M. Schweik. 2007. "Identifying Success and Abandonment of FLOSS Commons: A Classification of Sourceforge.net Projects" *Upgrade: The European Journal for the Informatics Professional VII.6*. http://www.upgrade-cepis.org/issues/2007/6/upg8-6English_Schweik_v2.pdf

Howison, J., Conklin, M., & Crowston, K. (2006). “FLOSSmole: A Collaborative Repository for FLOSS Research Data and Analyses.” *International Journal of Information Technology and Web Engineering*, 1(3), 17–26.

Ostrom, Elinor. 2005. *Understanding Institutional Diversity*. Princeton, NJ: Princeton University Press.

Robles, R., G. Gonzalez,- J.M. Barahona, J. Centeno-Gonzalez, V. Matellan-Olivera, & L. Rodero-Merino. 2003. “Studying the Evolution of Libre Software Projects Using Publically Available Data,” In J. Feller, B. Fitzgerald, S.Hissam, and K. Lakhani (eds.) Taking Stock of the Bazaar:Proceedings of the 3rd Workshop on Open Source Software Engineering. http://opensource.ucc.ie/icse2003.

Schweik, C.M. 2005. "An Institutional Analysis Approach to Studying Libre Software “Commons”." *Upgrade: The European Journal for the Informatics Professional* VI.3 (2005): 17-27.

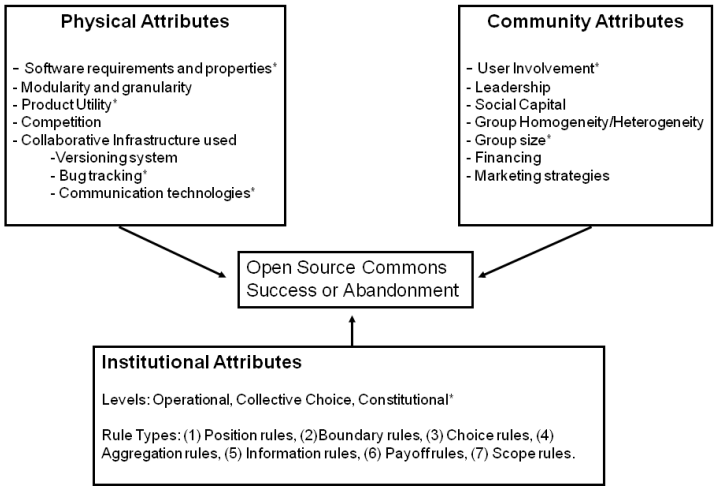
Schweik, C.M. & R. English. 2007. "Tragedy of the FOSS Commons? Investigating the Institutional Designs of Free/Libre and Open Source Software Projects" *First Monday* 12.2. http://firstmonday.org/htbin/cgiwrap/bin/ojs/ index.php/fm/article/ view/1619/1534.

Schweik, C.M., and English, R. In preparation. *Success and Abandonment of Open Source Commons*. Book manuscript expected to be published in 2011.

Van Antwerp, M. and G. Madey, "Advances in the SourceForge Research Data Archive (SRDA)", *The 4th International Conference on Open Source Systems - (WoPDaSD 2008)*, Milan, Italy, September 2008. http://www.nd.edu/~oss/Papers/srda_final.pdf

Weiss, D. 2005. “Measuring Success of Open Source Projects Using Web Search Engines,” Proceedings of the First International Conference on Open Source Systems, Genova, 11th-15th July 2005. Marco Scotto and Giancarlo Succi (Eds.), Genoa, 2005, pp. 93-99.

Figure 1: Independent Variables Thought to Affect Success or Abandonment of Open Source Commons



Note: "*" denotes concepts that we could operationalize using the Sourceforge.net dataset

Figure 2: Representative Growth Stage Classification Tree

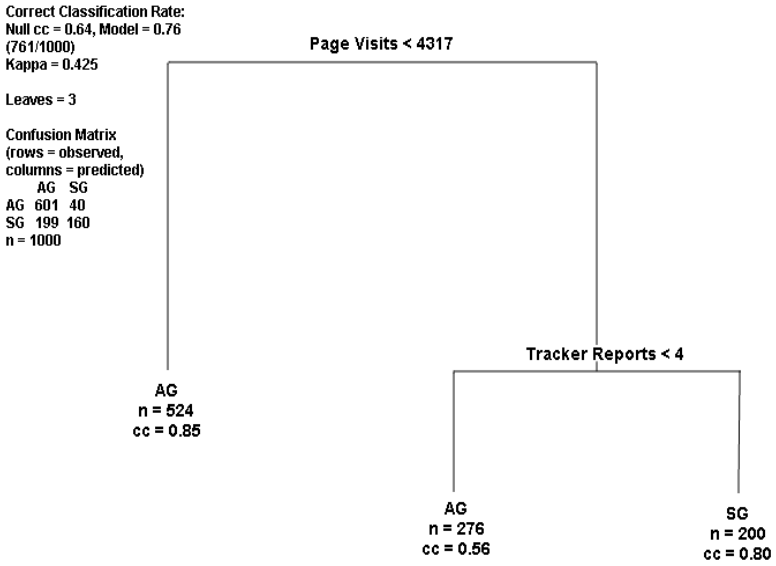


Figure 3: Initiation Stage Variable Importance Plot
(n = 2000; 500 Trees)

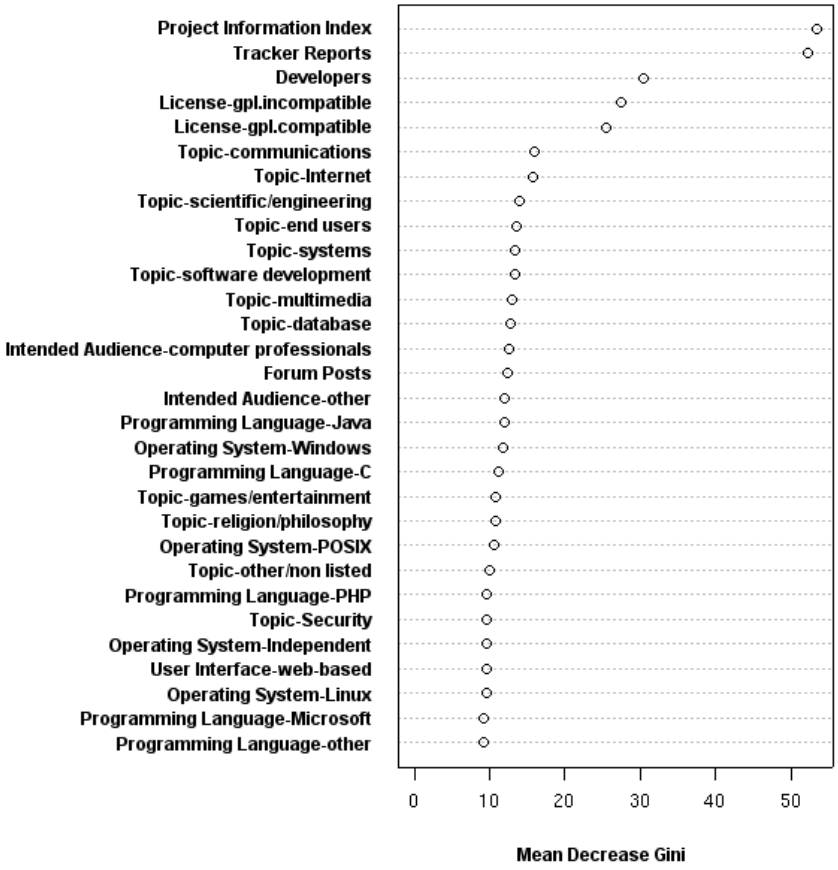


Figure 4: Growth Stage Variable Importance Plot
(n=1000; 500 Trees)