Extended Abstract:

Economic Returns to Open Source Participation: A Panel Data Analysis*

Il-Horn Hann
Marshall School of Business, University of Southern California

Jeff Roberts, Sandra Slaughter
Tepper School of Business, Carnegie Mellon University

Roy Fielding
Co-Founder, Apache Software Foundation (ASF)

Relying on volunteer labor, open source projects like the Apache web server create commercial quality software. Why developers contribute freely without direct remuneration has been widely debated. We offer empirical evidence that such participation can be explained by existing theories in labor economics. Analyzing panel data covering a four-year period, we find that increases in human capital, measured as project contribution, do not lead to increased wages. In contrast, credentials earned through a merit-based ranking system are associated with significantly increased wages. Our results suggest that status within an open source meritocracy operates as a credible signal of productive capacity.

* We thank the open source programmers who have contributed to this study. We also thank the participants of the session on "Economics of Open Source Software" at the 2004 Annual Meeting of the American Economic Association in San Diego, the participants of the conference "Open Source Software: Economics, Law and Policy" in Toulouse, France organized by Institut d'Economie Industrielle (IDEI) and the Center for Economic Policy Research (CEPR), and Rebecca Hann for their valuable comments. This research has been generously supported by a Faculty Development Grant from Carnegie Mellon University and a Doctoral Student Research Grant from the Carnegie Bosch Institute at Carnegie Mellon University.

I. INTRODUCTION AND OBJECTIVES

One widely debated question within the open source community is why open source programmers contribute voluntarily, thereby foregoing any direct remuneration that they could accrue while working on a commercial system. Often quoted individual level motivations for participating in open source development projects cover a broad spectrum including scratching a "personal itch" with respect to software functionality, enjoyment, and desire to be "part of a team" [Ghosh 1998; Raymond 1999a; O'Reilly 2000]. Others liken the open source community to a gift culture where the status of a participant depends on "what he gives away" [Raymond 1999b].

More recently, Lerner and Tirole [2002] opined that open source participation can, in part, be explained using economic theories. They argue that open source participation yields two types of rewards: immediate rewards that ensue from the increase in productivity (less the opportunity cost of time), and delayed rewards relating to various career concerns such as one's future marketability. For the latter, a participant motivated by career concerns has incentive to *signal* his or her abilities to the labor market. This signaling incentive is likely to be stronger when performance is more visible to the relevant audience, and performance is informative about talent [Lerner and Tirole 2002]. Such an incentive is particularly relevant in the information technology industry for two reasons. First, programming is often viewed as more of an art than a skill [Weinberg 1998]. Good programming is not confined to learning the syntax and specific features of a programming language and the practice of good documentation. Productive programmers are believed to have a certain aptitude that allows them to proceed logically from problem to solution, and in the process derive the most efficient and general software design possible. Subsequently, he or she has to take the lead in propagating the software design to co-workers and in sharing sufficient insights such that the co-workers in turn can be productive. Hence, it has been documented in the software engineering literature that the productivity of a "star" programmer is an order of magnitude greater than that of an average programmer. Second, the inability to formalize characteristics of highly productive programmers makes the programming process very difficult to evaluate. Not surprisingly, it often proves to be a challenge for employers to evaluate the majority of programmers [Kirsch 1996].

In addition to signaling imperfectly observable abilities, participating in open source projects has the potential to increase a contributor's human capital. In open source projects, contributors can select both the problem they want to attack and the implementation approach or solution. Once implementation is complete, other contributors provide timely feedback on the solution, ranging from identification of software defects (bugs) to suggestions on how to improve the submitted software code [Raymond 1999a]. Hence, contributing to open source projects can be seen as learning experiences that increase the programmer's knowledge. Inasmuch as this knowledge is transferable, open source participation increases the contributor's human capital.

In this paper we empirically investigate whether open source participation is consistent with theories in labor economics. As we have noted, contributing to open source projects can potentially be beneficial to contributors in two ways: (i) participation enhances existing skills or provides opportunities to gain new experience and hence makes contributors more valuable to employers and (ii) participation signals contributors' imperfectly observable productive characteristics to their employers. In order to measure the first benefit, we proxy the knowledge gained by contributors directly from the volume of their software source code submissions. To measure the second benefit, we exploit a unique setting of a specific open source project (the Apache Software Foundation) that ranks its members based on merit. From a signaling perspective, we maintain that certain abilities such as software design understanding and project leadership skills are endowments that are often difficult to evaluate. However, an open source community affords a venue in which such abilities can be discerned by highly skilled peers and rewarded with a higher rank. Using panel data collected on open source contributors and a fixed-effect specification of the standard wage equation to isolate contributors' time invariant qualities, we distinguish

the learning effect and the signaling effect from time invariant characteristics such as intelligence in explaining participation.

We find that, in the context of the Apache open source projects, greater open source experience, as measured in contributions made, does not result in wage increases for contributors. This suggests that employers do not reward the gain in experience through open source participation as an increase in human capital. On the other hand, achieving a higher status in the merit-based ranking within the Apache open source community is associated with a 13-27% increase in wages, depending on the rank attained. Our results are consistent with the notion that a high rank within the Apache Software Foundation is a credible signal of the productive capacity of a programmer.

II. Data Setting and Sources

Our data set combines several primary data sources including archival data from large open source software projects, and two targeted surveys of open source participants. We investigated three major open source projects and its subprojects under the control of the Apache Software Foundation (ASF); the Apache server project, the Jakarta (Java) project, and the XML project.

As a meritocracy, status, responsibility, and benefits are commensurate with contribution within the ASF. There are five observable levels of recognition or rank. In order of increasing status, these are *developer*, *committer*, *project management committee member*, *ASF member*, and *ASF board member*. This hierarchy within the ASF makes the Apache projects particularly appropriate for an evaluation of open source participation. As observed by Tyler, et al. [2000], data for identifying economic returns to a variable serving as a signal in labor markets should contain exogenous variation in the signal status among individuals with similar levels of human capital. Participants in ASF projects possess such a variable or credential – their rank within the ASF.

One of the basic tenets of open source software is that the development process and resulting products are "open" and freely available. Apart from the source and binary codes of the actual programs, Apache products include developer web sites, change logs, documentation, and developer communications in the form of email archives. From these products, we extracted two types of information: information pertaining to each individual's progression along the Apache career path, and information about each individual's source code contributions to the project.

We constructed a longitudinal data set of participant contributions by year. The longitudinal contribution data encompassed contributions made and accepted into any of our three target Apache projects. Data collection was completed in January 2003 and included all contributions from January 1, 1998 through December 31, 2002. To augment the longitudinal contribution data set outlined above, we collected demographic and job history data in two waves. Two secure web-based surveys of Apache contributors were conducted for this purpose. Of primary interest in each survey was the respondent's wages for the current and prior year. Dr. Roy Fielding, the then chairman of the ASF, introduced the first survey to 1,301 uniquely identified contributors via e-mail in November 2001. Two hundred thirty-three e-mail invitations were undeliverable. Of the remaining 1,068 contributors, 325 completed the instrument, yielding a response rate of 30%. The second wave involved the 237 respondents from the first survey who agreed to participate in another round of data collection. The second survey was introduced via e-mail in January 2003. Eleven e-mail invitations were undeliverable. Of the remaining 226 contributors, 122 completed the instrument yielding a response rate of 54%. Finally, we retain only those cross-sections where both the dependent and independent variables result from a common labor market experience, viz. the U.S. Applying the above constraints yields a cross-sectional time series panel of 147 cross-sections (individual respondents) each having at least two years of reported wages for a total of 360 observations for any of the years 1999 through 2002. We performed sample bias tests with respect to distribution of rank and rank patterns over time. In all cases, the results indicate that we cannot reject the hypotheses that the respondents and non-respondents are drawn from the same underlying population.

III. Empirical Methodology and Results

A. Estimation

Our approach is to employ econometric models that take advantage of our repeated measures data to control for time-invariant participant endowments. A common concern in the human capital literature is the potential correlation of some unobserved person-specific variable, say $u_{i,}$ with one or more of the regressors. If we assume that u_i contains some time-invariant heritable characteristic, such as intelligence, and to the extent that u_i is correlated with one of the other regressors, both OLS and GLS will yield biased and inconsistent parameter estimates [Chamberlain 1984]. In the present case, our concerns are focused on accounting for unobserved skill or quality differences across contributors. Potential quality differences include inherent programming and design capabilities, the ability to succinctly explain complex technical issues, or the ability to self-motivate and work in an unstructured, often chaotic, environment. Measures of work or programming experience may not adequately reflect such skills. Ideally, one would like to directly control for such individual effects, and indeed this is a goal of many empirical studies of human capital [Taubman and Wales 1973]. If, however, we assume that such abilities are rooted in the individual, and thus constant over time, then a fixed-effect (FE) model solves the omitted variables problem. By differencing away time-invariant variables, whether observed or unobserved, the FE model produces consistent parameter estimates, purged of heritable individual effects. Hence, we make use of our cross-sectional time-series (panel) data to fit a FE regression model to explore the relationship between open source participation and wages over time.

Because we are interested in the nature of the relationship between open source participation and the market for information technology labor, the human capital model provides a natural structure for assessing the returns to open source software participation. Accordingly, we formulate essentially Mincerian wage models traditionally used to test the impact of education on log-earnings [Mincer 1974]. We estimate the returns to open source participation using the following equation:

$$(1)\quad \begin{aligned} LWAGE_{i,t} = {} & \alpha_i + \beta_1 CNTRB_{i,t-1} + \beta_2 DEV_{i,t-1} + \beta_3 COM_{i,t-1} + \beta_4 PMC^{+}_{i,t-1} + \beta_5 EXPR_{i,t-1} + \\ & \beta_6 EXSQ_{i,t-1} + \beta_7 LEDU_{i,t-1} + \beta_8 JSWCH_{i,t} + \beta_9 FPUB_{i,t} + \beta_{10} FSWIN_{i,t} + \beta_{11} STDNT_{i,t} + \\ & \beta_{12} PDAPC_{i,t} + \beta_{13\text{-}15}(TIME_n)_{i,t} + \varepsilon_{i,t} \qquad (i = 1, \ldots N;\ t = 1 \ldots T); \end{aligned}$$

where *i* represents cross-section *i* observed at time *t*. The individual effect α_i is assumed to be an estimable cross-section specific constant term.

In our setting, total wage is a function of accumulated Apache contributions, rank within the Apache Software Foundation, accumulated work experience, programming skills, education, firm size, firm type (publicly listed or private), firm industry, and job switch. The dependent variable, LWAGE, is the natural logarithm of the sum of each participant's annual wages and bonuses. To account for inflation, each year's wages are expressed in constant 1998 U.S. dollars. We operationalize open source experience (CNTRB) as a participant's cumulative number of lines of code contributed and accepted by the Apache project. If CNTRB is a good proxy for the learning experience of an open source developer, we expect CNTRB to be positively correlated with LWAGE.

The dichotomous variables NORANK, DEV, COM, and PMC^{+} (collectively referred to as RANK) operationalize the observed levels of contributor rank naturally occurring within the Apache meritocracy, that is, latent contributor, developer, committer, and project management committee member or above, respectively. Promotion within the meritocracy is awarded after a positive peer review of one's tangible *and* intangible contributions to the project. RANK may then, in part, reflect sought after (yet hard to observe) traits valued by information technology labor markets, such as the depth of developers' understanding, their efficient designs, or their ability to persuade, to get people "on board" with their ideas and strategies. If Apache RANK is a signal of productive capacity in the open source environment, we would expect our RANK variables to be positively correlated with LWAGE.

EXPR and LEDU are the traditional human capital variables. EXPR is the total number of years of work experience of a contributor at time *t-1*. Consistent with the human capital literature, we expect wages to increase with work experience, but the percentage increase to decline with higher work experience. Thus we expect EXPR to be positively correlated with wages, and EXSQ to be negatively correlated with wages. LEDU is represents the number of years of schooling for a participant at time *t-1*. Education is typically represented as time invariant in studies of human capital accumulation; however, the presence of students in our sample makes it possible to infer accurate levels of schooling within subject by tracking a respondent's declaration of full-time student status within each observation period. Schooling is often the variable of primary interest in studies of human capital, and returns to schooling are expected to be positive.

The variable PDAPC is a qualitative variable that assumes the value of 1 when a participant's paying job in time period *t* involves contributing to any one of the Apache projects. STDNT, FPUB, and FSWIN are qualitative variables that assume the value of 1 if the participant is, respectively, a full-time student, works for a publicly traded firm, or works for a firm operating in the software or e-commerce industry in period *t* and is 0 otherwise. Lastly, the TIME*n* variables are dichotomous controls representing the observation period in which we observe the dependent variable. These time variables capture any systematic changes in our data attributable to general economic conditions.

B. Results

In evaluating our results, we first examine the impact of those variables in our model that are unique to our open source software setting. The coefficient for CNTRB (β_1=-.0001, p=.07) while significant at the 10% level, is, for all practical purposes, zero. Alternative specifications of CNTRB, such as the percent of overall contributions submitted, the median absolute deviation from the median, the number of standard deviations from the mean, and the amount of time spent working on Apache projects also failed to yield a substantive change in the coefficient. Our interpretation of this result is that open source project experience, expressed as cumulative contributions is not, per se, associated with an increase in wages. Given the size and complexity of ASF projects and the use of relatively unsophisticated software development tools and methods, it is not hard to imagine that employers would find it difficult to judge, first hand, the merit of an open source job candidate's contributions. The relationship between our measures of Apache rank or status within the project, however, tell a different story.

The coefficient of the rank variable, DEV, is negative but not significant (β_2=-.055, p=.20). It appears that the minimum threshold for attaining rank DEV, a single contribution, is not significantly different from no contributions at all and hence may not provide any additional insights into the productive capacity of the respondents at that rank. In addition, we find that the coefficients for COM and PMC+ are positive and significant (β_3=.132, p=.04; and β_4=.257, p=.01, respectively). We can calculate the percentage change in LWAGE associated with respondents having rank COM as $100\cdot(e^{(.1320)} - 1) = 14.11\%$. Similarly, the percentage change in LWAGE associated with respondents having rank PMC+ is 29.32%. That is, after controlling for open source experience, education, work experience, job switch, and firm characteristics, and latent individual effects via the FE estimators, respondents having an Apache rank of COM enjoy wages that are, on average, 14.11% higher that those having no rank at all. Likewise, respondents having an Apache rank of PMC+ enjoy wages that are, on average, 29.32% higher than respondents having no rank. Also note that the difference between the COM and PMC+ coefficients is significantly different from zero (β_4-β_3= .125, $p < .001$). A respondent at rank PMC+ enjoys wages that are on average 13.3% greater than that respondents have a rank of COM.

IV. CONCLUSION

The research presented here seeks to explore one of the more puzzling aspects of the open source phenomenon – why do developers participate? Specifically, we explore whether participation is consistent with well-established theories from labor economics. From this literature, we establish two

plausible theoretical bases for the existence of returns to open source participation; viz., human capital theory and signaling theory. A human capital explanation of participation suggests that open source experience serves a "job training" function. In contrast, signaling theory suggests that *successful* open source participation serves a signaling or sorting function for IT labor markets. Our analysis shows that employers do not reward the accumulation of experience in open source projects per se. Rather, successful open source participation, measured as higher open source rank, is associated with higher wages, even after controlling for work and programming experience. This finding is robust: related specifications for CONTRB as well as other model specifications yield similar results.

That wages do not increase with contributions to the Apache Project is consistent with the notion that employers find it difficult to assess the performance of programmers and hence changes in their human capital. It follows that employers would have even greater difficulty evaluating open source contributions in order to assess performance even though the source code is freely available. Even so, open source participation absent accompanying increases in rank may yet hold career advancement potential. Inasmuch as a contributor can apply his or her gained knowledge on the job, the programmer may be rewarded in the long run.

Our findings suggest that in the case of the Apache Project, the open source community effectively screens programmers based on their productive capacity. Employers appear to recognize Apache's merit based ranking as a reliable proxy that is correlated with desirable, but imperfectly observably productive abilities.

REFERENCES

Apache, "The Apache Software Foundation," Accessed September, 2002, (http://www.apache.org/foundation/).

Chamberlain, Gary, "Panel Data," *Handbook of Econometrics*, Z. Griliches and M. D. Intriligator, Eds.(Amsterdam:North Holland, 1984), 1247-1318.

Ghosh, Rishab Aiyer, "Interview with Linus Torvalds: What Motivates Free Software Developers?," *First Monday*, III (1998).

Kirsch, Laurie J., "The Management of Complex Tasks in Organizations: Controlling the Systems Development Process," *Organization Science*, VII (1996), 1-21.

Lerner, Josh and Jean Tirole, "Some Simple Economics of Open Source," *The Journal of Industrial Economics*, L (2002), 197-234.

Mincer, Jacob, *Schooling, Experience, and Earnings.*, (New York:Columbia University Press, 1974).

O'Reilly, Tim, "Open Source: The Model for Collaboration in the Age of the Internet," *Proceedings of the Computers, Freedom and Privacy*, (Toronto, Canada:2000).

OSI, "The Open Source Definition," The Open Source Initiative, Accessed May, 2001, (http://opensource.org/docs/definition_plain.html).

Raymond, Eric, "The Cathedral and the Bazaar," *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, (Cambridge, MA:O'Reilly, 1999a), 19-64.

____, "Homesteading the Noosphere," *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, (Cambridge, MA:O'Reilly, 1999b), 65-112.

Taubman, Paul J. and Terence J. Wales, "Higher Education, Mental Ability, and Screening," *Journal of Political Economy*, LXXXI (1973), 28-55.

Tyler, John H., Richard J. Murnane and John B. Willett, "Estimating the Labor Market Signaling Value of the Ged," *Quarterly Journal of Economics*, CXV (2000), 431-468.

Weinberg, Gerald M., *The Psychology of Computer Programming. Silver Anniversary Edition*, (New York:Van Nostrand Reinhold, 1998).