Firms' involvement in the projects of the OS community
Some preliminary empirical evidence and a research agenda

Andrea Bonaccorsi
University of Pisa
a.bonaccorsi@gmail.com

Dario Lorenzi
University of Pisa
darram81@hotmail.com

Monica Merito
Sant'Anna School
merito@sssup.it

Cristina Rossi
Politecnico di Milano
cristina1.rossi@polimi.it

Abstract

At present, more and more commercial firms are getting involved, to various extents, in the Open Source (OS) movement. While several studies have examined incentives and business models of these OS-based software companies, very few works have investigated whether and how firms actively participate to open projects. This paper contributes to the literature by providing empirical evidence on the role and the activities of software houses in community developed projects. The research proposes an original methodology of large-scale primary data collection from OS project repositories and linked Web sites. The findings show how different today's OS movement is from its origins and how important firm involvement has become.

1. Introduction and review of literature

It has been widely acknowledged that the projects of the OS community are impressive successful collective action processes [1]. Indeed, thousand of developers work in a decentralised manner without receiving any direct monetary compensation and have succeed in providing high quality and complex programs [2]. Even more surprisingly, the movement has evolved considerably in recent years. Open Source has gained an increasing economic importance and new agents (public bodies, Universities and research centres, and, even for profit firms) are taking part in the collective action by adopting open standards or using them in their productive processes. All this witnesses how the idea, proposed by the Open Source Initiative (http://www.opensource.org) in 1998, of getting OS world closer to the commercial one, has been extremely farsighted.

At present, the phenomenon of the engagement of software companies' in OS activities is becoming fairly widespread: more and more firms (including also several large market incumbents) are entering the market by using open code downloaded from the Internet as an input for providing to their customers Open Source-based products and services. Several empirical analyses have investigated the topic. Bonaccorsi et al. [3] have extensively described the phenomenon, calling these agents *Open Source firms* (On the contrary, proprietary firms are the ones that entirely base their activity on proprietary programs). Using data from a large scale survey on 146 Italian Open Source firms, the authors have found a wide diffusion of *hybrid business models* that mix the offering of open solutions with the provision of proprietary software.

Nevertheless, up to now, firms' participation in collective action has been poorly investigated by economic scholars: several works have explored as firms lobby for gaining trade protection which benefits, not only those firms that lobby for protection and bore the costs, but also the free riders [4]. Firms' engagement in OS projects represents a valuable chance of addressing the problem: Berdou [5], for instance, has investigated the dynamics of cooperation between paid developers and volunteers. To the best of our knowledge, the studies that have explored how firms contribute the OS community have focused on single projects [6, 7], or on a limited number of companies, often using case studies or other qualitative methodologies [8, 9]. Also the survey data collected by Bonaccorsi and Rossi [10] present several shortcomings. First, they do not provide any information on the projects to which respondents take part. Second, although the authors have distinguished between coordination and simple participation, it would be interesting to know more about firms' activity within the projects and about their evolution over time.

In this framework, this paper contributes to the literature by investigating whether and how firms contribute to the projects of the OS community. An innovative methodology, based on the analysis of the projects hosted on the largest OS repository, SourceForge (http://www.sourceforge.net) is proposed, aiming at providing original empirical evidence about three main research questions (i) Do for profit firms act not only as *takers* but also as *givers*, by directly contributing to OS projects hosted on SourceForge? (ii) If yes, what do firms do within the projects? Do they provide code and undertake coordination activities? Moreover, (iii) Does the presence of firms shape the evolution of the projects? Are there significant differences between projects participated by firms and the others? Finally, basing on the empirical findings, a research agenda for future developments is provided.

2. Data and methodology

The methodology proposed in this paper is based on data collection from SourceForge, currently the largest Open Source repository (as of January 17th 2007, the repository hosted 139,286 projects and 1,485,883 registered users) available. We sampled 300 projects currently hosted on the repository that were selected on the basis of their level of activity: SourceForge itself provides detailed criteria for assessing this dimension. Such criteria are based on several metrics, as the intensity of use of the instruments offered by the repository; the number of downloads and Web hits per day; the age of the most recent file uploaded, and so on[1]

On these bases, the repository provides a classification of the hosted projects: we selected the 300 projects ranking at the top positions, as done in other studies [11][2].

The repository itself was an important source of information about the projects. Indeed, for each one, it provides detailed information on: number of developers and administrators; date of registration on the repository; type of licence under which the code is released; intended audience (e.g. advanced users vs. end users); typology of products (e.g. Internet software vs. Management software); compatibility with different operating systems; use of mailing lists and forums; level of activity of different instruments (bug reporting, patches, feature and support requests, elements in Subversion, SVN); programming languages and databases used; availability of translations into foreign languages; awards won; number of donators; development phase and other statistics (e.g. downloads and visited web pages daily). All these dimensions were included in our database.

Nevertheless, it was not possible to detect firms involved by using only the repository, data on companies' participation has been collected mainly through projects' Web sites and other instruments outside SourceForge, in particular mailing lists and forums. Further details are provided in the Methodological Appendix.

3. Main results

The most important characteristics of the 300 sampled projects are as follows.

Project dimensions. In line with the literature [15], on average, the core development group of the project has demonstrated to be fairly narrow: the median number of programmers is 7, while 15% of the projects count only one participant.

Licenses. As expected [12], the most widespread licence is GNU General Public Licence (57.91%), followed by its derivation (LGPL, 12.84%) and BSD licence (7.76%).

Technical aspects. Almost every project has a forum; 66% of them have, at least, one mailing list and over 50% have a Web-site hosted on SourceForge. The most widespread programming language is Java (30.00%), even if the entire C family (For instance C, C++, visual C++, and so on) is still predominant (57.00%). It is worth noting that 55.67 % of the programs are compatible with Windows operating systems and 16.67 % are developed exclusive for these ones.

Intended audience and products. Projects target mainly developers (26.90 %) and end users (29.16 %). In general, it seems that the average user has high computer science skills: 30% of projects are clearly directed to firms and 10% target system administrators. Solutions provided are fairly heterogeneous (177 different cases of products identified), the most frequent ones are development software (26 cases, 8.67%) and Internet related applications (20 cases, 6.67%).

About firms' participation, 97 projects out of the 300 (32.33%) count the involvement of one or more firms: this result is intriguing and in line with the recent economic literature. It is possible to single out three main kinds of involvement can be: (i) *project coordination*, this is the most frequent way companies' participation, with 60 cases; (ii) *collaboration* to code development, in different phases and at different extents (bug fixing, testing or offering services, 37 cases); (iii) *provision of code* or protocols (For example, communication protocols used to share information among different devices) in 7 cases. It is worth noting that the sum is 104 instead of 97, as in 7 cases there is more than one firm involved in different ways in the same project (in 6 of these cases, there is a firm coordinating the project and one or more firms collaborating to it).

The numerous cases of project coordination witness the good relationships between firms and the Open source community. The ways in which firms succeeded in achieving the leadership have been investigated. In most cases (36), the firm itself founded the project, but there is evidence of companies that entered an existing project and replaced the coordinator (17 cases). Seven coordinating firms were settled up by the members of the initial project coordinating group.

Several disparities have emerged between projects participated by firms (group A) and the others (group B). Table 1 summarises the results of the inferential procedures carried on to detect statistically significant differences.

[1] See:http://sourceforge.net/docman/display_doc.php?docid=14040&group_id=1#rankings
for further details

[2] A complete list of the projects, you can be asked to Dario Lorenzi, who has prepared the database.

Table 1: Comparison between projects participated by firms and the others.

Characteristics of the projects	GROUP A	GROUP B	Test	P value
PARTICIPATION				
Average number of developers	19	5	a, b, c, e	≥ 0.01
Average number of project coordinators	11	1	a, b, c, e	≥ 0.01
TECHNICAL ASPECTS				
Average number of mailing lists	2.61	1.69	a, b, c, e	≥ 0.01
Bug reporting activity	744	358	a, b, c, e	≥ 0.01
Future requests	222	144	a, b, c, e	≥ 0.01
Elements in SubVersion	605	391	a, b, c, e	≥ 0.01
Patches	189	60	a, b, c, e	≥ 0.01
Programming language: C family	40.21%	65.02%	d, e	≥ 0.05
Programming language: Java	47.42%	21.67%	d, e	≥ 0.05
Number of translations into different languages	9	5	a, b, c, e	≥ 0.01
LICENSE				
Usage of the GNU General Public	45.36%	73.89%	d, e	≥ 0.05
USERS AND PRODUCTS				
Intended audience	(1) Developers (2) End Users	(1) End users (2) Developers	d, e	≥ 0.05
Companies as targeted users	39.92%	12.42%	d, e	≥ 0.05

Note: Hartley's Test (a), t Test (b), Mann-Whitney U Test (c), Chi-Square Test (d) and Pearson Correlation Coefficient (e)

In general, projects participated by firms are larger: they are joined by more developers and have more coordinators. Data seem to highlight that they show a higher level of activity, as it is witnessed, for instance, by the more intense patches activity and by the wider use of mailing lists.

Firms' presence has an impact on the management of Intellectual Property Rights (IPRs). The use of the General Public License is less common in projects joined by firms: the percentage of GPLed solutions decreases from 73.89 % in group B to 45.36% in group A. At the same time, commercial companies seem to shape also the typology of software provided: products targeted to companies are more diffused in group A and there is evidence that the average user of the software produced in group A has higher computer science skills. Other technical differences deal with the use of different programming languages, with a wider presence of the Java language.

Another aspect deals with the differences in the project advancement stage: projects in group A seem to be in more mature phases and more often reach a stable production (p value ≥ 0.1).

4. Conclusions and further developments of the research

Our empirical results acknowledge the role of for profit firms within the OS framework and confirm as the OS movement differs from its origins.

In almost one third of the cases there is some form of firms' participation and this seems to have an impact on projects' evolution. It has emerged that project participated by companies are larger and more active, make less use of the GPL licenses, show several technical peculiarities, and, in general, produce software solutions targeted mainly on companies and high skill-users.

Summing up, the issue is undoubtedly challenging and it calls for the definition of clear research hypotheses that should drive the empirical analyses:

(i) Are projects with firms more successful than the others? Do companies contribute crucially to the achievement of the OS movement? Moreover, is it firms' involvement to determine the success of a project or, on the contrary, do successful projects tend to attract companies? Clearly, inferential procedures that we used up to now, are not suitable to address these issues, more sophisticated techniques are needed

(ii) What are the characteristics of the firms involved in OS projects? Data should be gathered on their structural characteristics (e.g. size, age, competences, product/service portfolio, etc.) in order to inquire whether and how they differ from those following a

traditional software production process and how these differences are related to project participation.

(iii) Moreover, a wide literature [13] is now exploring the so called *open innovation model*, according to which firms can achieve a greater return on their innovative activities by using a broad range of sources [14]. In this framework, it is then of interest to explore how project participation shapes the innovation activity of the involved firms. In short, are these firms more innovative than the others?

(iv) Finally, it has been widely acknowledged that Open Source movement was born in Universities and research centres [9]. Our data have shown that, besides firms and individuals developers, several Universities and research centres are involved in OS projects. It is important to provide evidence on the impact of public research and also to explore whether and how participating in OS projects affects the scientific activities of University researchers: does the access to this community have a positive impact on academic performances (e.g. publications)?

These are only part of the questions raised by the fast evolution of the Open Source software. In order to provide rigorous answers, we are actually working on enlarging our sample from 300 to 1,000 SourceForge projects. The first analyses taken on this new sample seem to confirm our preliminary results. We are also exploring the possibility to collect information on participation directly from the code posted on SourceForge, using software scanning applications (as for instance CODD) as it has been done for individual developers [15].

5. References

[1] Von Hippel E., von Krogh G. (2003), "Open Source software and the 'private-collective' innovation model: Issues for organization science", Organization Science 14: 209-223

[2] L. Benussi "The history of the free/libre open source software: stories from the open source evolution." , *In A. Bonaccorsi and C. Rossi (eds.) Economic Perspective on Open Source software.* Franco Angeli, Milano, Italia, 2006.

[3] A. Bonaccorsi, S. Giannangeli, C. Rossi, "Entry Strategies Under Competitive Standards: Hybrid Business Models in the Open Source Software Industry", *Management Science 52*, 2006, pp. 1085-1098.

[4] K.M. Olson, "Free riders among the rent-seekers: a model of firm participation in antidumping petitions", *SSRN Working Paper*, 2004.

[5] E. Berdou, "Insiders and outsiders: paid contributors and the dynamics of cooperation in community led F/OS projects", *Open Source Systems-IFIP Working Group 2.13, Foundation on Open Source Proceedings*", 2006, pp. 201-208

[6] J. Henkel, "Selective revealing in open innovation processes: The case of embedded Linux", *Research Policy 35 (7)*, 2006, pp. 953-969.

[7] L. Dahlander, M.W. Wallin, "A man on the inside: unlocking communities as complementary assets", *Research Policy 35* , 2006, pp. 1243–1259.

[8] L. Dahlander, M.G. Magnusson, "Relationships between Open Source software companies and communities: Observations from Nordic firms", *Research Policy 34*, 2005, pp. 481-493.

[9] Y. Lin, "Hybrid innovation: how does OSS firms collaborate with the FLOSS community", *Knowledge, Technology and Policy 18(4)*, 2006, pp. 86-100.

[10] A. Bonaccorsi, C. Rossi, "Contributing to the common pool resources in Open Source Software. A comparison between individuals and firms", *SSRN Working Paper*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=430920, 2003, accessed on January 19th 2007.

[11] K. Klincewicz, "Innovativeness of Open Source software projects", *MIT Working Papers*, 2005.

[12] J. Lerner, J. Tirole, "The economics of technology sharing: Open Source and beyond", *Journal of Economic Perspectives 19(2)*, 2005, pp. 99–120.

[13] H. Chesbrough, W. Vanhaverbeke, and J. West, "Open Innovation: researching a new Paradigm", *Oxford University press*, Chambridge, 2006.

[14] H. Chesbrough, "Open innovation: the new imperative for creating and profiting from technology", *Harvard Business School Press*, Boston, MA, 2003.

[15] R. Ghosh, V. Prakash, "The orbiten Free Software survey", *First Monday, Peer Review Journal of Internet 5 (7)*, 2000.

Methodological Appendix

Some explanations are needed about the procedures of data gathering, particularly regarding how information on firms' involvement has been collected.

First of all, it is worth noting that every project has its own specificities: it provides different instruments outside SourceForge (Web sites and, less often, mailing lists and forums), organized in different ways.

Hence, the first step was to explore in deep the Web site of each project, using the site map where offered. In this phase, the idea was to watch at every Web page and to its external links (which have been explored in depth too) to find possible elements related to firms' participation. Climbing the learning curve, it was possible to discover that some sections contain more often relevant information, such as the history of the project, developers' personal Web pages, Frequent Asked Questions, and so on. We have also looked at indications of registered trademarks.

The same approach was followed with mailing lists and forums, in some cases we asked directly if there were firms involved in the project.

In brief, given the wide pool of information that we have taken into account, we are fairly confident that there is no firm's participation, when no element suggesting involvement were found. On the other hand, where these elements were discovered, we have also visited the company's Web site, particularly to understand if there

was an effective participation and its main characteristics.