OPEN SOURCE ADOPTION IN LARGE US COMPANIES

Spinellis, Diomidis, Athens University of Economics and Business, Patision 76, 10434 Athens, Greece, dds@aueb.gr
Giannikas, Vaggelis, Athens University of Economics and Business, Patision 76, 10434 Athens, Greece, egian@dmst.aueb.gr

Abstract

Various organizations increasingly adopt open source software, both on desktop PCs and servers. Since the first movements in open source in the 1960's its growth has lead to new approaches in software development, licensing, and distribution, as well as in software vendors' business models. The literature includes very interesting studies regarding prospective benefits, business models and case studies. However, the adoption of open source in large, global companies and its relationship with factors such as profitability, revenues and industry sector has not yet been researched. This study aims to answer these questions based on data we collected from Fortune 1000 companies and provides a method that can be applied in similar contexts.

Keywords: *Open source software, adoption in business, server and client applications*

1 INTRODUCTION

Even though the open source software (OSS) has been object of intense research by the academics and other research centers, little attention in scientific research regarding open source software adoption related to business tactics and strategies has been given. This study aims to research this issue, give some first results and provide a methodological framework, which will lead to its expansion. Here we study open source software adoption in large (Fortune 1000) US companies in four areas: the web server and client and the corresponding operating systems hosting these applications.

This study is organized as follows. Section 2 details our research questions while Section 3 discusses methodology. Section 4 contains analysis and hypothesis testing. Section 5 contains a short presentation of related work, based on technology and open source adoption and specific cases of open source usage. We close this study with some general conclusions and issues for further research.

2 RESEARCH QUESTIONS

We formulated our study around the following research questions.

Hypothesis H1: The adoption of OSS is associated with economies of scale.

Larger companies are more likely to adopt OSS than smaller ones due to economies of scale associated with a larger number of IT assets. As a proxy for the size of a company's IT assets (consistent information for which is not directly available) we use a company's annual revenues, arguing that in today's knowledge-intensive economy a fairly fixed size of IT infrastructure is needed to generate a given revenue. We rejected the number of employers and assets as possible indicators, because a) there may be many employees who do not use IT and b) IT equipment as part of a company's assets varies very much, and therefore both cannot be used as a proxy for the company's IT infrastructure size.

Hypothesis H2: The adoption of OSS is positively associated with a company's profitability.

This can be either a direct casual relationship, where the lower cost of OSS software acquisition is directly reflected into the company's profits, or an indirect result of profitable well-run companies adopting OSS as an appropriate IT practice.

Hypothesis H3: The adoption of OSS in one area is positively related with the adoption in another either on the client or the server side.

This may show that there are many advantages in the adoption of OSS that would lead someone, who have already used it in one area, to choose it for another application.

3 METHODOLOGY

3.1 Justification

Focusing on the Fortune 1000 companies benefitted our study in a number of ways. First, the companies form a wide sectoral cover of the economy. Although the large size of the companies can limit the applicability of our study's results to small and medium enterprises, this is offset by the fact that their activity forms a large part of the US economy. Furthermore, their large size increases the visibility of their operations, and makes them more likely to appear in our study's client software radar. Finally, for all the companies we could readily obtain relatively reliable financial data, a sectoral categorization, and an address of an operating web site, and thereby also a probable domain-name address their employees use accessing the web.

One other problem in our sample is the US focus. This does indeed limit somewhat the wider applicability of our study, but the limitation was offset by the data's reliability and the sample's homogeneity. Nevertheless, our study is applicable to companies operating in the global economy: the 2008 revenues of the companies we studied represent a big percentage of the corresponding world GDP, and many of the companies are export oriented. Having mastered the methodological and research issues through the Fortune 1000 sample, a global study could certainly follow.

In comparison to issuing questionnaires our method considerably lessens self-selection bias. With a questionnaire-based study it would be likely that companies with antiquitated IT strategies and systems would be less likely to respond; the same could also be true for companies whose IT management formed a tactical or strategic advantage.

3.2 Data Collection and Processing

We used a variety of techniques to obtain data about software used on the companies' desktops and by their back-office operations. Due to the methods we used, we focused in three types of software: the web client (on the desktop), the web server (on the back office), and the operating system on which the two are running (in both instances).

To determine the desktop operating system and web client software used by each company we examined web server logs. We collected about 55GB of log files from three sources: our own servers, servers of our contacts, and located in the wild. We started by collecting and processing log files of servers we administer. This allowed us to get a feeling of the type and quantity of data we were looking for. Having processed about 4,7 GB of our own data allowed us to present convincingly what our analysis tools and argue that the aggregation we performed would not raise any issues regarding privacy and confidentiality. This allowed us to branch out to the second source of data, namely log files given to us for our research by administrators we contacted for this purpose. Through this road we collected another 11,6 GB of log data. As a final step, having seen many types of logs, we used this knowledge and tried our luck by issuing various Google queries that should, in principle, locate publically available log files. To our surprise, we located many such files totalling 33.8 GB of log file data; this 67% of our data came from publically available web logs.

In total the log files contained 278 million entries. Most web servers record a log entry for every file a web client retrieves from it in a standardized format. The log entry contains the IP address of the client, the date and time of the access, the file retrieved, the operation's result, page's referrer, and details about the client's software. For the purposes of our study the important fields were the IP address, the date, and the client's software. As a first step we processed each entry to convert the (typically) numerical IP address, like 195.212.29.137 into a host name like blueice18n5.uk.ibm.com. (Few web servers are configured to make this conversion when they record the entries, in order to save the corresponding lookup time). We then went through all log entries looking for those where the last two parts of a client's hostname matched those of a Fortune 1000 company's web site address. For instance, the above host name would match IBM's web site address www.ibm.com. Finally, for each matching entry we examined the client software details to determine whether the web client and the underlying operating system were proprietary or open source. As an example, the following client identification string

```
Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9)

Gecko/2008052906 Firefox/3.0
```

corresponds to an open source browser (Firefox) running on a proprietary operating system (Microsoft Windows XP). We tabulated the results by company and year in a list specifying whether a company was found to use a proprietary or open source (or both) operating system or browser.

To determine the web server used by each company we retrieved the company's top web page using the *wget* tool, and logged the HTTP protocol headers. One of those headers contains an identification string of the web server, which we used for establishing whether the company used a proprietary or an open source product.

Finding out the operating system that hosts a company's web server proved to be trickier. At the time of writing, the only standard to describe to the outside world a machine's operating system is the domain name system (DNS) HINFO record. Unfortunately, this facility has fallen into disuse: system administrators seldom fill the corresponding DNS field and even the entity that oversees the official names (the Internet Assigned Numbers Authority — IANA) last updated the database of operating system names in 2002. Therefore, to determine the operating system type, we employed *nmap*, a network exploration and port scanning tool, *Nmap* works by sending specific network packets to the host, and analyzing minute accidental differences in the responses that can be traced back to the responding computer's operating system. It then matches those results against a database of 1503 (for the version 4.76 we used) so-called operating system fingerprints. The match is probabilistic in nature and can often fail.

3.3 Threats to Validity

There are several threats to the validity of this study; many are associated with the data we used for identifying companies using open source operating systems and browsers.

The first problem concerns the small number of software systems we examine. A company may use hundreds of software systems for a variety of purposes, but we examine just four: the web browser, the web server, and their corresponding operating system hosts. Here we can argue that these are ubiquitous and highly-visible systems, from which we can derive generalizable lessons for desktop applications and system software. Nevertheless, lessons from these systems cannot apply to specialized vertical applications, and this remains a limitation of our study.

Another serious problem concerns the resolution of numerical IP addresses into host names. For the sake of performance, web servers typically only record the numerical IP address when a host records a page, thus saving the overhead of a DNS lookup operation. To match web server requests with specific companies we performed this resolution before processing the logs. However, our name lookup lagged the actual client's fetch operation by as much as four years (some of the requests date from 2004). During that period a company could have changed its internet provider, and therefore its assigned IP address. Due to the disruption associated with an IP renumbering, we believe that such changes would not be frequent, particularly for large companies with a significant internet presence. Due to the small fraction of the available IP address space occupied by the companies in our study, address changes are more likely to yield false negatives (our study missing data from companies that changed their IP address) than false positives (mistakenly associating an IP address with a specific company).

One other validity threat concerns the log data's provenance. A large percentage of the data we used came from publically available log data located through Google queries. Consequently, had no control over how that data was collected and stored. It is even conceivable, though unlikely, that the logs were doctored before being placed online in order to mislead those downloading them. Along the same lines, there is an element of self-selection in the data we collected. Our data doesn't cover all companies we wanted to study, and it is more likely that it covers companies whose employees are allowed or even encouraged to frequently use the internet. Therefore our sample is likely to be skewed toward the inclusion of knowledge and service-oriented companies and the exclusion of companies dealing with, say, commodities, metal bashing, and construction. Comparing our sample against the population of our study confirms this fact. Along these lines one can argue that the intake of open source software in our sample is likely to be higher than that in the population, because more internet-savvy companies are also more likely to appreciate and be able to realize the benefits of OSS adoption.

Finally, there is a chance that pristine logs may contain incorrect data concerning the client's operating system and browser. This can occur when a user or administrator changes the contents of the HTTP User-Agent header that the client presents to the browser, typically for one of two reasons. The change can be performed as a security measure (Kumar, 2008) (directly at the client, or through an application level proxy firewall), in order to minimize the chance of exposure to operating system or client–specific attacks. Also, occasionally, a client browser's identity may be changed as a way to increase a particular browser's compatibility with browser-specific web pages (Dennis and Harrison, 1997). There is no evidence that any of these two practices is particularly prevalent.

Other problems are associated with the operating system fingerprinting technique we used for determining the operating system hosting each company's web server. The tool can be fooled by equipment residing in front of the server, such as firewalls, content delivery networks, and load balancing systems. However, this discrepancy does not significantly affect our arguments, because we consider these elements a part of a company's core IT infrastructure.

A concern voiced by some of this paper's reviewers is whether the use of a particular operating system or browser reflects a company's policy rather than choices of individual employees. This criticism is justified, because we academics and researchers are blessed with virtually unlimited freedom regarding the choice, setup, and configuration of our computing infrastructure. However, the situation in industry is different. There, automated mass installations from a single stable configuration image, a severely constrained user ability to install new software, and rigidly enforced IT policies are the rule. In large listed companies externally imposed legal requirements and standards (Larsen et al, 2006), the provision of a standard operating environment cite (Halprin, 2000), and the imposition of change management procedures (Sellens, 2001) align the software used by a company's employees with its policies.

4 FINDINGS

We analyzed our data with the help of the open source statistical analysis application R (version 2.8.1). The main findings appear below.

4.1 Data Analysis

In order to test our first two hypotheses we started searching the difference between the means of each variable (OSS users vs. non OSS users) using the t-test method. This gave us the opportunity to get a feeling at the results we expect. We next used the logistic regression model (Ross, 2004) to find the specific relation between our variables (Revenues, Profits ~ Open source adoption). We chose this model because of the particular type of the dependent variable (binary). You can find the results in table 1. In the last column there is the t-test based on the independent variable.

Variable		Coefficient	Wald Z	p-value	t - statistic
Dependent	Independent				
Open source software adoption	Revenues	1,09e-05	2,8714	0,004085**	3,0362***
	Profits	4,17e-05	1,5727	0,115773	1,6208
	Profits (>0)	1,38e-04	2,8800	0,003975**	2,9192***
* a=0,05 , ** a=0,01 , *** a=0,005					

Table 12: Statistical results of data analysis (H1-H2)

Trying to search hypothesis 3, we used contingency tables. Having these we performed the appropriate Chi-square test for independence and then the Cramer's V to identify the correlation between applications and operating systems either on the client or on the server side. We also searched the proportion of an OSS user to use more than one OSS using a simple z-test.

Variables	Pearson Chi-square	p-value	Cramer's V
Web browser – Client OS	44,3858	2,70e-11***	0,31088
Web server – Server OS	70,3134	5,06e-17***	0,4458

Table 13: Statistical results of data analysis (H3)

Number of known applications	Observations	OSS adoption percentage	P(OSS_adoption > 50%)
At least 2	446	51,57%	62,69%
At least 3	353	55,24%	98,61%*
All	119	63,87%	99,92%***

Table 14: z-test for proportions

Finally we give some statistics of on the adoption of open source for each application and in total.

Open source software	Observations	Low limit (95%)	Percentage	High limit (95%)
Client operating system	477	17,72%	20,33%	22,94%
Web browser	477	69,64%	72,54%	75,44%
Server operating system	381	25,29%	28,87%	32,45%
Web server	905	31,88%	32,82%	33,76%
Total	964	40,07%	40,66%	41,25%

Table 15: Percentages of open source adoption

As one can see from table 4 our sample is pretty big reaching 96,4% of population for at least one software type mainly because of the contribution of data related to web servers.

4.2 Hypothesis Testing

Hypothesis 1: As mentioned before, in order to check this hypothesis we used revenues as the company's size figure. Both tests (t-test and logistic regression) give statistical significant results. The t-test shows means difference with 99,5% proportion and the coefficient's value is significant at a 99% level. Moreover coefficient is positive, so our hypothesis is confirmed. Larger companies are more likely to adopt open source software. We want to notice that the very small values regarding coefficients are due to the very big difference between the variables (0/1 for the adoption and thousands for revenues). This also occurs in the other regression tests.

Hypothesis 2: Here, we checked two cases. Firstly all the firms, independently if they had profits or deficits, and secondly only the profitable companies. In the first case our hypothesis is not confirmed. Both tests showed no significant correlation. On the other hand, when searching only companies with profits we find a strong statistical correlation between the variables (99,5% in t-test and 99% in regression). So we can support that our hypothesis is partly confirmed for those companies that are profitable.

Hypothesis 3: The analysis that came before confirms our third hypothesis. Not only the chi-square and Cramer's V figures show a positive and strong (99,5%) correlation between each side's applications (client and server) but the z-test for proportions we performed lead as to urge that users of open source choose it for more than one application. The percentages may vary a lot according to the data we used by they are all above 50%, getting higher as we check more applications and at least in two cases they are giving very significant results. This may show that the benefits coming from OSS use are strong enough for the user to prefer it, as a solution for another need, to a proprietary one.

5 RELATED WORK

Related work in our area can be broadly classified into the study of technology adoption in general, open source adoption in particular, and papers providing concrete evidence of open source software use in specific contexts.

5.1 Technology Adoption

Open source adoption from the aspect of the user has been researched very much in the last decades. This has led to the creation of many theories and models, independent (or not) of technology or application, trying to search this issue. In a recent study, eight such kinds of models are presented and compared (Venkatesh et al., 2003). These models go by the following names: "Technology Acceptance Model-TAM", "Theory of Reasoned Action-TRA", "Motivation Model- MM", "Theory of Planned Behavior-TPB", "Combined TAM and TPB - C- TAM - TPB", "Model of PC Utilization-MPCU", "Innovation Diffusion Theory- IDT" and "Social Cognitive Theory-SCT". Many of them have been revised and improved since their first version.

The research question here is in which level these models can been used, not only for a single user but for a whole business. The final software will be used by end users but very often the choice is been made by managers that may take under consideration factors as cost, support etc.

5.2 Open Source Adoption

Several surveys and researches have been made by advisory companies and other organizations on the adoption of open source software giving us interesting information for our research. Through these we can find useful statistics and trends that can illustrate the current situation. Even though most of them are oriented to analysis per software category or type, there are reliable surveys that search the issue per continent, in the public domain or refer to specific use cases. Suggestively we mention up-to-date reports that show open source adoption in web server applications with the percentage of 50% with an increase of over 800% in the last 10 years (Netcraft, 2009). A few years ago, in a survey in USA, 87% of the companies that took place used or evaluated open source software in their activities (Wall et al., 2005).

In another category of researches, we can find discussions on fundamental issues in open source adoption. For example, many useful studies tried to answer the question why should someone choose open source software (Deek and McHugh, 2008) for their needs. Other issues are which is the right software to be chosen (Wang and Wang, 2001), what should the firm's characteristics be in order to adopt open source successfully (Dedrick and West, 2003), which is the open source growth in the last years and its usage in new software production (Haefliger et al., 2008)

Finally there are also studies with predictions about open source usage in the next years. In one of them, it is mentioned that since 2012, 9 over 10 enterprises will have adopted, in some way, open source which will have lost its competitive advantage as low cost software when the concept of software as a service will have grown up (Gartner, 2008).

5.3 Evidence of Use

In this section we have collected some specific cases of open source adoption in companies and pubic organizations around the world. The result was a list of about 40 cases to specific evidence of use along with the software, the application and the industry it was used in. The sad fact is not only that the amount of cases found is small but also that in most cases one of the above (section, organization, software and reason and used) is undefined in the source. So, unfortunately, it is really hard to use this table for reliable conclusions. On the other hand, here someone can find something more that statistics. Below you can see a small number of these cases.

Area	Software	Organization
Back – Office	Linux	Amazon
		Google
	Apache Web Server	IBM
	FreeBSD	Yahoo!
Sales	Linux	Toyota Motor Sales
R&D	Eclipse	Ford Motor Company
		Motorola
Undefined	Linux	McDonalds, Chevron, Amazon, Pixar, Dreamworks, Salomon Smith, Barney, Morgan Stanley and Credit Suisse First Boston
		Disney, Merrill Lynch, Pixar, the US Postal Service, Siemens and Mercedes-Benz

Table 16: Open source evidence of use

Reading the whole list we notice that Linux is the software mentioned in most cases mainly in Back-Office IT section. This doesn't mean, for sure, that Linux is the most popular open source software in organizations. Neither that open source is preferred to be adopted in back-office applications. We believe that this is because of the size and the importance of a change in this type of software and section. We support that a change in a company's operating system is more worth to be published than a small office suite change on an employer's work station. Moreover, employees may often not even know that their employers use open source systems in their everyday tasks. This may explain why some popular open source applications, like Mozilla Firefox and OpenOffice are hardly mentioned in scientific bibliography.

Talking about bibliography we have to speak about what is called grey literature. We have found plenty references on the net, in internet-based journals or nameless articles. For us, these are not trustworthy enough to be included in our research. In some cases where we use internet published references, we have checked that these are already published in scientific journals, conferences etc., which means that they are already checked for their reliability. In fact we quote, only in a couple of cases, the scientific source and not the original. Searching there you can find the internet link..

Certainly we can't maintain that these are the only cases that firms use OSS. As mentioned before there are many applications that either firms may not want to publish the adoption of or don't even know they are used by their staff. The fact that there isn't evidence of use in literature doesn't mean that enterprises don't actually adopt this kind of software. This is why we tried to find data for our research that are independent of a company's will to publish or even talk about them.

6 CONCLUSION

The goal of our study was to provide principal research findings in the area of open source adoption by large companies. To the best of our knowledge this has not yet been studied in the scientific literature. Even if someone can claim that the small number of applications we checked is a threat to the validity of our results, the extend of OSS usage in business remains a significant finding. Just ten years after the first release of a product as open source it seems that open source benefits make it competent and efficient enough for a company's needs.

In the other part of this study, the methodology framework it proposes, it provides a useful tool for obtaining reliable and objective data. New ideas to obtain data, additional companies and new software types can be used for further research. We tried one of the first studies on this issue and the results show that related work should certainly follow.

References

Dedrick, J. and West, J. (2003). Why Firms Adopt Open Source Platforms: A Grounded theory of innovation and Standards Adoption. MIS Quarterly, 236-257.

Deek, F. P. and McHugh, J. A. M. (2008). Open Source Technology and Policy, Cambridge University Press, chapter Why Open Source.

Dennis, B. and Harrison, M. (1997). Grendel: a Web browser with end user extensibility, in 'Compcon '97', pp. 74-79.

Deshpande, A. and Riehle, D. (2008). The Total Growth of Open Source. 4th Conference on Open Source Systems (OSS 2008), 197-209.

Gartner (2008). The State of Open Source, 2008.

Haefliger, S., von Krogh, G. and Spaeth, S. (2008). Code Reuse in Open Source Software, Management Science 54(1), 153-180.

Halprin, G. (2000). A System Administrator's Guide to Auditing. Short Topics in System Administration. USENIX Association, Berkeley, CA.

Larsen, M.H., Pedersen, M.K. and Andersen, K. V. (2006). It governance: Reviewing 17 it governance tools and analysing the case of novozymes a/s. Hawaii International Conference on System Sciences.

Netcraft Ltd (2009). February 2009 Web Server Survey, Available online: http://news.netcraft.com/archives/2009/02/18/february_2009_web_server_survey.html. Current March 2009, Bath, UK.

Kumar S. P., A (2008). Service Cloaking and Authentication at Data Link Layer, arXiv 0804.3796v1.

Ross, S.M. (2004). Introduction to Probability Models, Third Edition. Academic Press.

Sellens, J. (2001). System and Network Administration for Higher Reliability. Short Topics in System Administration. USENIX Association, Berkeley, CA.

Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D. (2003), User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 27(3), 425--478.

Walli, S., Gynn, D. and von Rotz, B. (2005). The Growth of Open Source Software in Organizations, Technical report, Optaros.

Wang, H. and Wang, C. (2001). Open source software adoption: a status report. IEEE Software 18(2), 90-95.