Influence in the Linux Kernel Community

Timo Aaltonen and Jyke Jokinen
Institute of Software Systems, Tampere University of Technology
first.last@tut.fi

Abstract. Several success stories of open source (OS) products have been seen during last decade. Due to the economical importance of the products, it is important to know who are the ones who have the largest influence to the products. Is there a dominant player in developing communities? In this paper[1] the aspect is studied with respect to the Linux Kernel community. We show that the influence is centered to a small number of core people, and corporates have a large impact to the development. Moreover, we enumerate the most influential companies.

Key words: data mining, Linux kernel

1 Introduction

Open source (OS) software development has gained much attention lately. During last decade several success stories, like Apache, Mozilla and Linux, has been seen. Apache is the market leader of the world's web servers [2] having over three times the market share of its next-ranked (proprietary) competitor. Internet Explorer has been losing market share to OS web browser, especially to Mozilla [3]. Linux [4] is a free UNIX-type operating system originally created by Linus Torvalds.

Due to the economical importance of open source, it is important to know, who influences the development. Is it carried out by altruistic individuals and what is the impact of large organizations? By knowing these facts one is able to predict the directions how the products evolve in future. This is essential when choosing between different open source and proprietary alternatives.

This paper studies the influence of the developers and leaders of the Linux Kernel. The Kernel was chosen because it is the only operating system challenging Microsoft Windows, the available amount of data is large, and the number of people working for the project is numerous.

The study of influence is based on counting the signers of *Developer's Certificate* o*f Origin* [5] (DCO) for patches. In short, signing a DCO has two main meanings

[1] This paper is a revised version of [1]

1. the original author of a patch certifies that she has the right to submit it under the open source license indicated in the file; and
2. later code maintainers and Linux lieutenants indicate that they accept the patch by adding their own signature.

It is obvious that the signers are influential persons in the Linux Kernel community. The more person signs patches the more influence she has.

We applied a set of measures to the mined data. The most important findings are

1. A large portion of the influence is contributed by a relatively small amount of people.
2. Based on studies on e-mail addresses, corporations seem to have much influence in the Linux Kernel community.
3. The most influential companies can be studied by relating the most influential persons to their employees, and summing up the number of signed DCOs for each company.

We have mined our data from publicly available sources. The DCO signatures have been mined from GIT [6] revision control system used by Linux Kernel developers. Technical details of the GIT data mining are presented in [1]. Personal data of the signers have been searched with Google and from certain public data sources. Section 1 introduces the measures, which are applied to individuals in Section 2 and to companies in Section 3. Section 4 discusses the paper.

2 Measures

We have developed a set of measures to be applied to our data. The measures are divided into two categories: *personal*, *company-related*. The personal measures attempt to highlight various aspects of people in the Linux Kernel community:

- **Influence distribution.** Number of signed patches are counted for each person. Then these (person, amount) pairs are sorted in descending order. The measure illustrates how the control and development work is distributed in the community.
- **E-Mail domain distribution.** The Linux Kernel development is highly geographically distributed. This measure shows where and by which kind of organizations does the decision-making takes place.
- **E-Mail taxonomical distribution.** Measure attaches a category to e-mails from taxonomy: corporate, open source project, ISP, e-mail provider, university, personal domain, and other.

The company-related measures attempt to reflect the role of companies in the development:

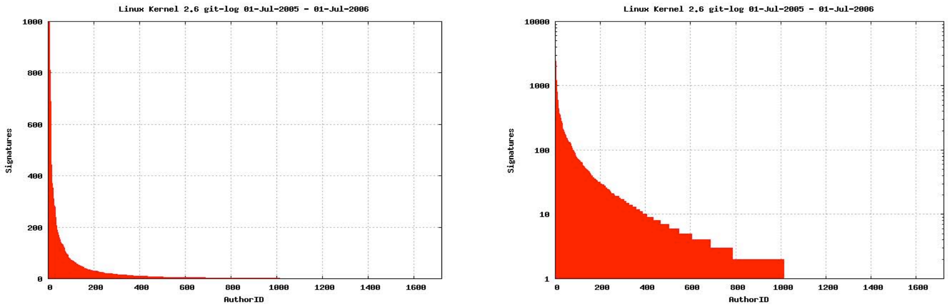
- **Impact of Companies.** Leaders and developers of the Linux Kernel community signing the patches are related to companies they work for. Then the influence of employers of each company are summed together. This sum is the influence of the company.

The measures *E-Mail taxonomical distribution* and *Impact of Companie*s are the most interesting (and the most controversial) measures. The interesting piece of information they reveal is the role of companies in the Linux community. The former indicates what is the impact of companies together, and the latter shows which individual companies are the most influential. Both measures are controversial, since their evaluation is based on opinion and skills of the researcher(s) evaluating them. Another evaluators might get slightly different values. However, we believe, that the measures describe interesting aspects of the Kernel community, even if their accuracy is not ideal.

33 Measures for Individuals

The influence distribution of the Linux Kernel developers is depicted in Figure 1. The number of signed patches is on the y-axis and individual signers are on the x-axis sorted with respect to the number of sign-offs.

A notable shape of the curve slanting to the left is quite common in open source projects. Actually, the y-axis has been truncated to make the shape of the curve more visible. The curve takes this shape because a small number of core people lead the whole community. In our previous studies [7] we have noticed that a small group of developers contribute more than the rest of the group. We call this phenomenon the *flagpole effect.*



(a) linear

(b) logarithmic

Fig. 1. Influence distributions in linear and logarithmic scales.

To make clearer the strength of the flagpole effect, the influence distribution is redrawn on a logarithmic scale in Figure 1b. It is somewhat surprising, that even now, the curve tends to slant to the left so heavily.

E-Mail Domain Distribution shows that the Linux Kernel development is highly distributed. The measure is based on studying the e-mail addresses of the persons who sign the patches. Figure 2 illustrates the distribution with respect to highest level domains. Not surprisingly, *com* domain is the number one in this measure. The second place is taken by *org*, and the third one is occupied by*de* domain, implying that many of the Kernel developers are from Germany.

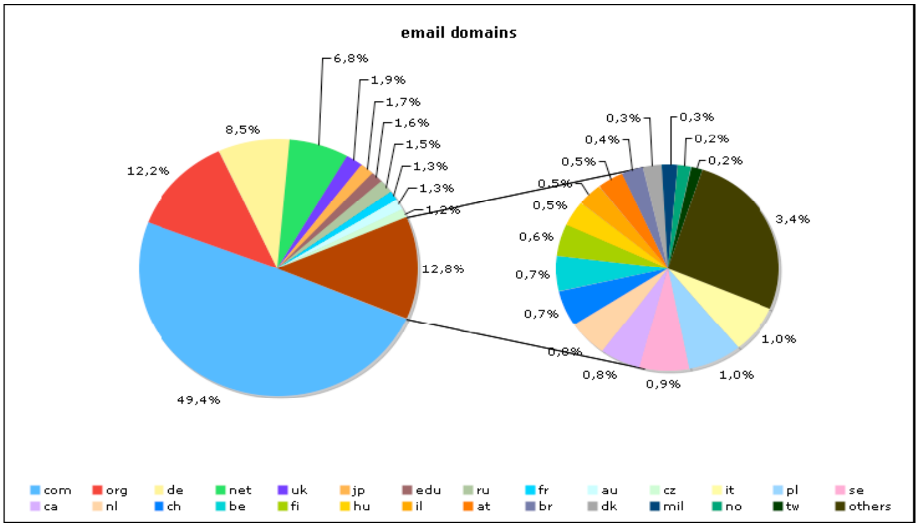


Fig. 2. The e-mail domains of patch

E-Mail Taxonomical Distribution was made by attaching each domain a category from taxonomy: *corporate*, *open source* project, *ISP*, *e-mail provider*, *university*, *personal domain*, and *other*. Google was used manually to attach a category to the domain. The results are illustrated in Figure 3. The distribution has one unexpected result: category *personal domai*n taking the second place is somewhat surprising.

Category	Number	Category	Number
corporate	342	ISP	110
personal domain	207	open source project	78
other	200	e-mail provider	21
university	114		

Fig. 3. The taxonomy of e-mail addresses

4 Measures for Companies

We took a closer at the top 100 signers according to criterium of most signed patches, and used Google search engine to study whether the top 100 leaders were employed by some organization. Then we were able to calculate the size of the impact of the organizations to the Linux Kernel development.

The search techniques we used were various. We had two obvious starting points: a name and an e-mail address. If a developer had a company-related e-mail then it is quite obvious that she works for the company. Few developers had their CV on www, which was easy to find with a simple search. Book publishers and organizer of open-source-related conferences maintain lists of their contributors with a small description of people's careers on www. Often, these people were among the top 100 leaders to the Linux Kernel. One surprisingly fruitful technique was search with the name part from an e-mail address. People seem to preserve their original e-mail names in their e-mail addresses. This way the employer was joined to a set of contributors. Some people were found from Wikipedia [8]. Moreover, several creative searches were carried out.

The results of **Impact of Companies** measure are shown in Figure 4. The company with the largest impact during our time interval has been SteelEye Technology. Actually, all 928 signatures related to the company have been signed by a single person. Obviously SteelEye Technology has been very active during our time window, and perhaps all patches from the company are signed by the person. After SteelEye Technology, the next companies should not be a surprise. Google's rank has been improved by Andrew Morton's migration to the company [9].

5 Discussion

We studied the Linux Kernel development by mining data from GIT repository, and applying four measures to the mined data. The measurements show that relatively small amount of people control the development. Similar results have been reported earlier in [7]. E-Mail domain distribution and taxonomical distribution show that the Linux Kernel is mostly developed in western countries and corporations have half of the influence. Most of the most influential companies are quite expected. However, the most influential company being SteelEye Technology was not expected result.

Similar research to ours has been published in [10]. In this study patches between version 2.6.19 and 2.6.20 of Linux Kernel has been analyzed (our study used a one year time frame between July 1st 2005 and July 1st 2006). The study contains more measures than our study, and they are partly the same (Influence distribution). Similarly to us, the people have been joined to the companies they work for, and this information has been used to compute measures for companies. However, the joining method is simpler than ours, since it is based only to the email domains, whereas we have used more sophisticated searches.

Company	impact	Company	impact
SteelEye Technology	928	Oracle	136
IBM	924	Symantec	135
Google	759	Academic (all universities)	135
Intel	742	MISC	133
Novell	665	Broadcom Deep Blue	131
		Solutions Limited	121
OSDL	588	Qlogic	114
UNKNOWN	453	CoopTel	107
Cicso (Topspin)	421	MontaVista Software	105
Debian	376	Freescale	98
Alcatel	322	Hewlett-Packard	94
Red Hat	302	Network Appliance	92
Netfilter (a project)	293	Circle Computer Resources	86
Linutronix	283	Mellanox Technologies	85
Conectiva	280	Ultra	79
Ameritech	260	Toshiba	77
Dunvegan Media	184	Motorola	74
Simtec Electronics	165		
Wise Riddles Software	164		
SGI	155		
Levanta (previously Linuxcare)	138		

Fig. 4. Companies and the number of patches signed by the personnel.

Acknowledgements

This work has been partly funded by Academy of Finland, project Managing Open Source Software as an Integrated Part of Business (OSSI).

References

1. T. Aaltonen and J. Jokinen, "Demography of linux kernel developers," Tech. Rep. 41, Tampere University of Technology, Institute of Software Systems, 2006.
2. Netcraft, "Web server survey." http://news.netcraft.com/archives/web_server_survey.html/, 2006.
3. R. McMillan, "Mozilla gains on IE," *P*C *World*, 2004.
4. "Linux online." http://linux.org, 2006.
5. L. Torvalds, "Developer's certificate of origin 1.1." http://www.osdl.org/newsroom/press_releases/2004/2004_05_24_dco.html, 2006.
6. L. Torvalds and J. C. Hamano, "GIT -fast version control system." http://git. or.cz, 2006.
7. T. Aaltonen, J. J¬arvenp¬a¬a, and T. Mikkonen, "Oss architecture and implications," tech. rep., eBRC, 2006.
8. "Wikipedia, the free encyclopedia." http://en.wikipedia.org/wiki/Main_ Page/, 2006.
9. B. Pro_tt, "Morton gets googled," *Linu*x *Today*, 2006.
10. Corbet, "Who wrote 2.6.20?." http://lwn.net/Articles/222773/, February 2007.