ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

von Engelhardt, Sebastian; Freytag, Andreas; Schulz, Christoph

Working Paper

On the geographic allocation of open source software activities

Jena Economic Research Papers, No. 2010,009

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: von Engelhardt, Sebastian; Freytag, Andreas; Schulz, Christoph (2010) : On the geographic allocation of open source software activities, Jena Economic Research Papers, No. 2010,009, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/32590

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

JENA ECONOMIC RESEARCH PAPERS

2010 – 009

On the Geographic Allocation of Open Source Software Activities

by

Sebastian v. Engelhardt
Andreas Freytag
Christoph Schulz

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

© by the author.

Jena Economic Research Papers 2010 - 009

On the Geographic Allocation of Open Source Software Activities

Sebastian v. Engelhardt[*] Andreas Freytag [*†]
Christoph Schulz [*]

February 2010

Abstract

Open source software (OSS) is marked by free access to the software and its source code. OSS is developed by a 'community' consisting of thousands of contributors from all over the world. Some research was undertaken in order to analyze how global the OSS community actually is, i.e. analyze the geographic origin of OSS developers. But as members of the OSS community differ in their activity levels, information about the allocation of *activities* are of importance. Our paper contributes to this as we analyze not only the geographic origin of (active) developers but also the geographic allocation of OSS activities.

The paper is based on data from the SourceForge Research Data Archive, referring to 2006. We exploit information about the developers' IP address, email address and indicated time-zone. This enables us to properly assign 1.3 million OSS developers from SourceForge to their countries, that are 94% of all registered ones in 2006. In addition we have information about the number of posted messages which is a good proxy for activity of each developer. Thus we can provide a detailed picture of the *world-wide allocation* of open source *activities*. Such country data about the supply-side of OSS is a valuable stock for both, cross-country studies on OSS, as well as country-specific research and policy advice.

JEL Code: L17, C81, L86
Keywords: Open Source Software, Geographical Location, Open Source Activities

---

Financial support from the Klaus Tschira Foundation is gratefully acknowledged.
[*]Friedrich-Schiller-University Jena. Email addresses: Sebastian.von.Engelhardt@uni-jena.de, Andreas.Freytag@uni-jena.de, and Christoph.Schulz@uni-jena.de
[†]European Centre for International Political Economy (ECIPE), Brussels.

1 Introduction

Open source software (OSS) is developed by a community that include hobbyists as well as companies, and the source code—the human-readable recipe—is 'open'. This means that everybody has access, and the right to read, modify, improve, redistribute and use the source code. Thus, OSS appears to be a case of a "private provision of a public good" (Johnson 2002). As the community is often described as being global, OSS seems to be a digital public good with a truly globalized private provision.

However, beside anecdotal evidence for the internationality of certain teams of OSS projects, the question remains how global the OSS community actually is and how the supply-side of OSS differs among countries. Thus, the issue of the geographical allocation of OSS developers comes into focus of research on OSS. It turns out that the most OSS developers come from North America and Europe. This result is quite consistent independently from the method used. Such methods to gather information about the geographic origin of OSS developers can be broadly distinguished into two approaches. Some studies are based on survey-data, while other work is based on specific data drawn from code of certain OSS projects (the credit files respectively), mailing lists or informations from platforms like SourceForge.

Robles et al. (2001) provide a combination of both types of data collection. In Ghosh (2006), David et al. (2003) and Ghosh et al. (2002) one can find survey-based information about the origin of OSS developers. Lancashire (2001) provides information about the world-wide allocation of Linux and Gnome developers, based on data collected form the Linux Credit file and in case of Gnome developer-contact information from the project's web-site. The most recent research dealing with the geographic origin of OSS developers is Gonzalez-Barahona et al. (2008). The article provides a worldwide picture of OSS developers, weighted by population, internet users and GDP. With respect to the data mining regarding developers at SourceForge, Gonzalez-Barahona et al. (2008) build on Robles & Gonzalez-Barahona (2006). Robles & Gonzalez-Barahona (2006) use information about the email addresses of registered users and the indicated time-zone to assign developers at SourceForge in 2005 to their countries. However, for 25% of all cases direct assignment to countries is not possible, because of the combination of a generic, i.e. not country specific Top Level Domain like .com with the country unspecific timezone GMT. Because of this high level of not directly assignable users Robles & Gonzalez-Barahona (2006) develop methods to estimate the geographic allocation of this 25%.

Our work is inspired by Gonzalez-Barahona et al. (2008) and Robles & Gonzalez-Barahona (2006), but proceeds along two lines: First, we do not

have to estimate any geographic origins, as we can directly assign 94% of all developers registered at SourceForge in 2006. We make use of relevant informations delivered by email, time zone and the Internet Protocol address. Combining these, we are able to assign 1.3 million developers to their countries without the need to estimate allocations. We also present indicators for the reliability of the localization methods we use. We cross-checked the results which delivers a good indicator for the validity of each of our methods. Second, we provide information about how active each developer is. With individual data about the number of posted messages we have a good proxy for activity. We can thus distinguish active from non-active (but nevertheless registered) developers, and we are able to show the worldwide allocation of OSS *activities*. Information about activities are of importance, as members of the OSS community differ in their effort levels, numbers of contributions etc. (see e.g. David & Rullani 2008).[1] With the active developers and activity, our study can show a more accurate geography of the supply-side of OSS-development.

2 The Need for Accurate Data about OSS Activities

Software-development is an important part of each country's ICT-Sector. But without having information about the OSS activities the picture of the software industry's supply-side remains incomplete. For example regarding workforce and human capital, data typically count only the paid labor force and thus ignores the—for the most part non-paid—OSS developers. But even adding the number of OSS-developers to the number of paid jobs is not correct, because of two reasons. First, not all registered developers are active: at SourgeForge, only every fifth was active in 2006. Second, some OSS-developers have jobs in the software-sector. This would yield an incorrect double-counting. Taking into account the OSS-activity-level can, together with the numbers of the paid software development, help to get a more accurate picture.

Data including the number of active OSS-developers and their activity level of a country are of importance for both, policy markers and businesses. Government decisions in competition policy as well as its decision to support (or not support) OSS-development, OSS-based business models and start-ups respectively, should be based on knowledge about the national human capital and capacity regarding OSS. The same is true if e.g. an entrepreneur has to decide to do an OSS-based start-up, or if a firm plans to implement an OSS-based business model etc.

In addition to this, a more complete and correct picture regarding the supply-

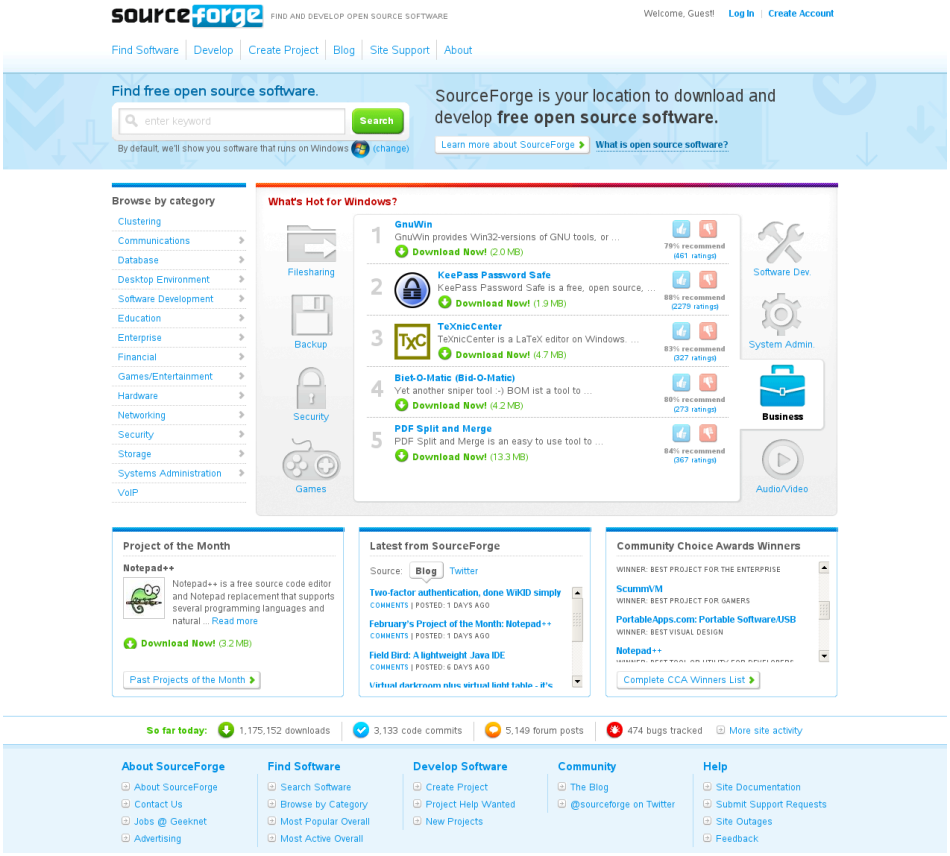
[1] For further literature on the division of labor within open source projects etc. see among others den Besten et al. (2008), Giuri et al. (2008), von Krogh et al. (2003).

side of software-development of the different countries can improve cross-country studies, like those dealing with aspects of the digital divide. Finally, our dataset can be used for analyzing the impact of country-specific factors on OSS-activities. The fact that its magnitude differs among countries points to the embeddedness of OSS. Therefore, in von Engelhardt & Freytag (2010) our dataset is used to analyze the role of country-specific culture and institutions.

3 Data Source and Methodology

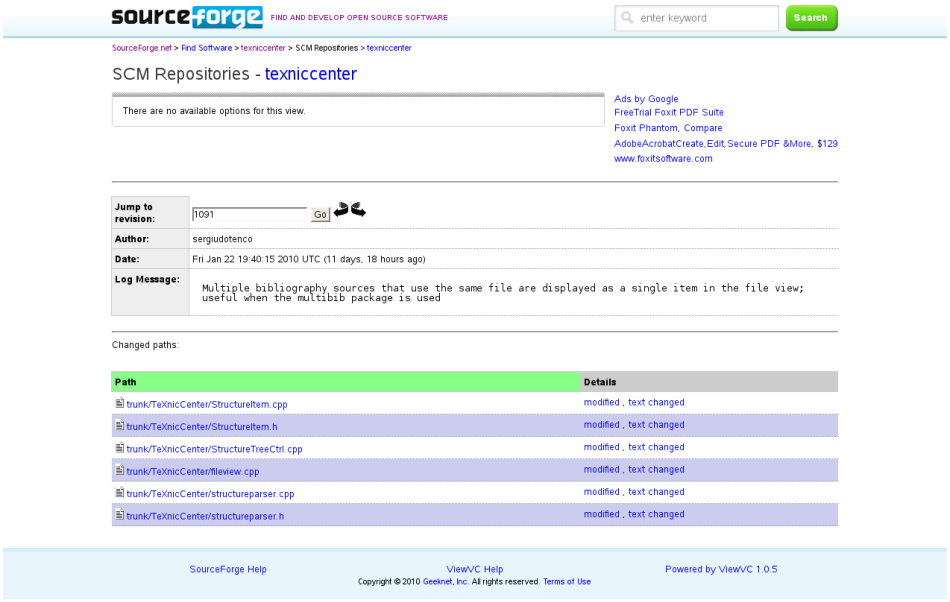
As already mentioned above, we analyze the geographic allocation of activities of OSS developers registered at SourceForge in 2006. SourceForge is an internet platform for developers to control and manage OSS projects—to get an impression see the screenshot of SourceForge's Start Page (figure 1). In

Figure 1: Start Page of SourceForge.Net



a sense it is a virtual center where the developers of a certain OSS project can meet, discuss, coordinate their tasks, upload new developed codes etc. Such activities are documented. For an example see Figure 2 which shows the online documentation of committed code to the software-project 'TeXniCenter'

Figure 2: Information about Comitted Code (TeXniCenter) on SourceForge



at SourceForge. SourceForge is the largest repository of OSS projects. And, while finished version of software can be downloaded by anybody, access to the developer-areas needs registration. When registering users have to provide some personal data, like a valid email address.

SourceForge related data are an often used source for research on OSS.[2] Nevertheless, one has to mention that not all OSS projects are hosted at SourceForge and that there might be a bias to under represent regions from Asia as they have more local communities (see Gonzalez-Barahona et al. 2008, p 358). However, we follow Gonzalez-Barahona et al. (2008) who state that "from an economic perspective it is useful to examine the distribution of participation in global projects". In this respect, data derived from SourceForge make a good indicator.

[2] See for example Au et al. (2009), Giuri et al. (2010), David & Rullani (2008), Eilhard (2008), Gonzalez-Barahona et al. (2008), Fershtman & Gandal (2008), Comino et al. (2007), Robles & Gonzalez-Barahona (2006), Lerner et al. (2006), Xu et al. (2006), and Lerner & Tirole (2005)

We draw our data about OSS developers registered at sourceforge.net from the SourceForge Research Data Archive (SRDA). SRDA is offered by the University of Notre Dame under a special agreement for scientific research (Madey, see also http://www.nd.edu/~oss/Data/data.html). The database consists of monthly dumps containing some of the information stored at the SourceForge web-page. The latest dumps with all information necessary for our analysis are those of the year 2006. As we are able to identify each user by the user-ID we can connect information about the indicated email address and time zone, the saved Internet Protocol address and the number of posted messages of each registered developer.

When OSS developers register at SourceForge they have to indicate a valid email address. Additionally, when registering developers can change the time zone from the default-value to their specific time zone (e.g. "Europe/Berlin"). Email address and timezone are saved in the 'users'-table of SourceForge. However, not all dumps offered by SRDA contain email addresses, as 'email' has been removed by the SRDA-team from the users-table as of the October 2006 dump for privacy purposes.

Furthermore, the SRDA contains tables with the Internet Protocol address of the users logged in. Internet Protocol addresses of registered users can be found in the tables 'user_ip_dl_auth' and 'audit_trail_users'. The first one consists of information about users who have registered in the respective month. The second table consists of data generated by SourceForge in order to be able to restore the data, i.e. are data used for backups. Here data are saved only when something was changed (data changed/uploaded by a user etc.). Nevertheless, in the SRDA only the dumps of July, August, September and October 2006 contain the tables 'user_ip_dl_auth' and 'audit_trail_users'.

The original data of the 2006 dumps delivers approximately 1.4 million datasets which have to be cleaned of all duplicates, fake accounts and non reliable data. Then we assign to each user his or her geographical origin by making use of information we derive from the email address, the time-zone and the IP address:

country coded Top Level Domain (ccTLD) If the Top Level Domain of the respective email address is a country coded Top Level Domain (ccTLD), it can be used to assign the user to a country. For example, emails ending with ".us" will be assigned to the USA, with ".nl" to the Netherlands, or with ".de" to Germany. Thus, the assumption standing behind this is that each user's ccTLD correctly indicates his or her native country or the country of (long-term) residence respectively.

A problem are so-called open ccTLDs. While registration for ccTLDs is

limited to citizens or firms of their respective countries, in the case of *open* ccTLDs registration is possible to any interested registrant subject to a charge (Edelman 2002). The reason is that such Top Level Domains are an attractive domain name for a global website as e.g. ".tv" (for Tuvalu) looks like "television", or ".ws" (Western Samoa) looks like "website". This enables to generate some revenue by selling domains containing to such name spaces to firms etc. (see e.g. `http://worldsite.ws`). But this implies that such open ccTLDs can *not* be used for geographical identification. In fact, these are de facto generic TLDs such as ".org" or ".com". Therefore we exclude all open ccTLDs from the dataset when identifying via country coded Top-Level Domain of the email addresses.

Second Level Domain (SLD) For all email accounts with generic TLDs it is possible to use information from the so-called second level domain (SLD). For example in case of "xyz@yahoo.com" is "yahoo" the SLD. It is possible to identify the location of the domain server of a SLD. Therefore we manually assign to each of the top 1000 SLDs their domain server, and therefore the country of the server. If one assumes that the location of the domain server of the SLD of a user's email address also indicates the country the user lives in, it is possible to assign users with generic TLDs to countries. Clearly this method can be criticized as the probability of mistakes might be high. For example a Spanish developer using an yahoo.com email account would be counted as a citizen of the USA. We will come back to this later.

Time Zone (TZ) Another indicator is the time zone (TZ) indicated. A TZ like "EST" sums up several countries and can therefore not be used for the analysis. The same is true if 'time zone' has its default value, as it is not known whether the option TZ was just ignored, or not. Thus, members with the default or a summarizing TZ can not be geographically identified via this method. But nevertheless, well-defined and unique TZs can be used to assign a country to a user. For example, if one has chosen the TZ "Europe/Berlin", then this can be assigned to Germany. Clearly the assumption standing behind this is, that users report their TZ correctly (when changing it from the default value "CET") and that this indicates their usual place of residence.

Internet Protocol address (IP) If the Internet Protocol address (IP) of a user is available then this information can be used for geographic location using GeoIP. GeoIP is a technology for IP geolocation. It allows to identify geographic location of internet-connected devices via their IP-range.

Namely the location of servers of internet service providers, universities etc., can easily be identified. Via these providers, the geographic location of internet users can be identified quite correctly, to get an impression, the reader can visit www.maxmind.com/app/locate_my_ip. This technology allows to make use of the information offered by the saved IP, given that such information can be found in the SRDA. We thus use the partially available IPs of users to identify their actual habitation by GeoIP. (Some of the IPs in the data are not useable for our purpose, as they belong to a range that is assigned to regions but not to certain countries.)

Identifying the geographical origin of OSS developers via ccTLD, IP and indicated TZ seems to be quite reliable, with IP to be the most correct one. In order to get an impression of the actual reliability, we cross-check by pair-wise comparison the results that ccTLD, IP and indicated TZ deliver. For example to check ccTLD versus IP, we take the subset of users that have an email address with a ccTLD *and* also a saved IP in the data. We firstly identify the users via ccTLD and then again identify them via IP. The resulting two lists of users with their assigned countries are now cross-checked per each user. This procedure of pair-wise comparison is executed for all methods which delivers the corresponding matching rates. The results are presented in table 1. As

Table 1: Matching rates of the different identification methods

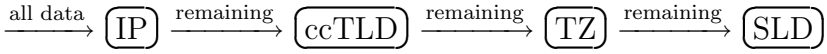
	IP	ccTLD	TZ	SLD
IP	100%	89.16%	87.29%	51.83%
ccTLD	89.16%	100%	80.45%	–
TZ	87.29%	80.45%	100%	56.45%
SLD	51.83%	–	56.45%	100%

the reader can see, ccTLD and IP have a matching of 89.16%, while IP and TZ deliver the same results in 87.29% of all cases, and TZ and ccTLD have 80.45% of matchings. As already mentioned above, identifying the location via SLD is from a theoretical point of view the weakest method. Thus, not surprisingly, checking IP with SLD, and TZ with SLD delivers matching rates of only 51.83%, and 56.45% respectively.

Because of this we combine all four methods in the following way (see also figure 3): First, when possible, we identify users' geographical location via GeoIP. The remaining users are then identified via their ccTLD, if possible. The rest is then assigned to their country using the information about the TZ.

The remaining 283,028 users not located are then assigned to a country using the information about the SLD. Doing so we end up with 1,315,263 users who are assigned to their countries. Thus, we are able to identify 94% of all users, only 83,217 could not be identified. Now, as one might doubt the results using the SLD, we compare the results with and without the identification via SLD. The resulting allocations do not differ much.

Figure 3: Process of geographical identification



As already mentioned, we are also interested in the activity levels of the developers. Therefore we extract information about whether and if, how often, a user posted a forum message in 2006, and use this as an indicator of activity. The SRDA contains information about the number of posted messages, stored in the table 'forum'. This table is delivered by all the dumps from January 2006 until December 2006. The information of the columns 'msg_id' and 'posted_by' of the table 'forum' enables to link each user to his or her posted forum messages. With this information we are able to distinguish active developers (developers who had posted in 2006) from non-active ones. Furthermore, counting the number of messages posted by users from a country deliver us data about the OSS activity that comes from a specific country. The tables 2 and 3 show the top 30 countries with respect to number of active developers and activities. In both cases we present the results with and without the identification via SLD. Both, the number of active developers and number of messages as well as the results with and without the SLD-identification are highly correlated (about 0.99).

Weighting all these information by the number of inhabitants in 2006 (source: World Bank 2007), we finally end up with country-specific informations about the number of OSS developers per 1,000 inhabitants, the number of *active* OSS developers per 1,000 inhabitants, and the *level of OSS activity* (Number of posted messages per 1,000 inhabitants). As we have the information about the activity of each developer, our data offers more information about global OSS activities than any other non-survey data we are aware of. The next section gives details on these results.

Table 2: Active Developers, Top 30 Countries

without SLD			with SLD		
Rank	Country	Active	Rank	Country	Active
1	United States	85,485	1	United States	112,981
2	Germany	23,267	2	Germany	24,197
3	United Kingdom	13,031	3	United Kingdom	14,051
4	Canada	11,238	4	Canada	11,524
5	France	10,525	5	France	10,987
6	Australia	7,897	6	Australia	7,945
7	Netherlands	6,666	7	Netherlands	6,687
8	Italy	6,185	8	Italy	6,200
9	Spain	4,563	9	Spain	4,760
10	Sweden	4,546	10	Sweden	4,642
11	Brazil	4,028	11	India	4,163
12	India	3,824	12	Brazil	4,038
13	Russia	3,184	13	China	3,793
14	China	3,149	14	Russia	3,217
15	Belgium	3,026	15	Belgium	3,034
16	Switzerland	3,007	16	Switzerland	3,033
17	Austria	2,537	17	Austria	2,549
18	Poland	2,514	18	Poland	2,520
19	Denmark	2,314	19	Denmark	2,314
20	Hong Kong	1,861	20	Hong Kong	1,894
21	Norway	1,814	21	Norway	1,883
22	Finland	1,805	22	Finland	1,842
23	Singapore	1,685	23	Singapore	1,685
24	New Zealand	1,635	24	New Zealand	1,635
25	Israel	1,458	25	Israel	1,467
26	Argentina	1,456	26	Argentina	1,466
27	Czech Republic	1,443	27	Czech Republic	1,443
28	Mexico	1,401	28	Mexico	1,401
29	Japan	1,331	29	Japan	1,357
30	South Africa	1,211	30	South Africa	1,216

Table 3: Activity (Messages), Top 30 Countries

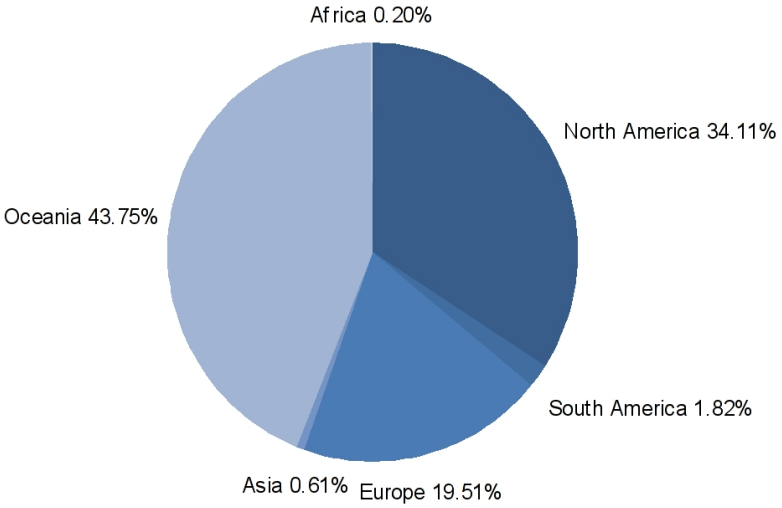
without SLD			with SLD		
Rank	Country	SumMsg	Rank	Country	SumMsg
1	United States	7,734,231	1	United States	8,842,906
2	Germany	1,807,233	2	Germany	1,839,842
3	United Kingdom	1,238,922	3	United Kingdom	1,282,156
4	Canada	1,064,001	4	Canada	1,078,637
5	France	826,659	5	France	845,302
6	Australia	720,326	6	Australia	721,693
7	Netherlands	570,732	7	Netherlands	571,198
8	Italy	433,354	8	Italy	433,793
9	Sweden	361,550	9	Sweden	367,379
10	Spain	329,655	10	Spain	340,567
11	Belgium	260,754	11	Belgium	261,111
12	Switzerland	244,784	12	Switzerland	245,485
13	Russia	238,709	13	Russia	239,824
14	Austria	220,552	14	Austria	220,814
15	Brazil	205,466	15	India	211,735
16	Denmark	198,132	16	Brazil	206,042
17	India	197,009	17	Denmark	198,132
18	China	160,279	18	China	184,113
19	Japan	148,305	19	Japan	149,500
20	Norway	136,902	20	Norway	140,096
21	Poland	136,555	21	Poland	136,798
22	Finland	132,489	22	Finland	134,024
23	New Zealand	119,515	23	New Zealand	119,515
24	Hong Kong	116,515	24	Hong Kong	117,749
25	Argentina	116,492	25	Argentina	117,419
26	Czech Republic	114,713	26	Czech Republic	114,713
27	Romania	111,843	27	Romania	111,843
28	Singapore	106,009	28	Singapore	106,009
29	Israel	98,492	29	Israel	98,653
30	Mexico	81,956	30	Mexico	81,956

4 Results: The World-Wide Allocation of OSS Activities

In this section we present the results of our data mining and assignment procedure. Before we focus on the origin of active developers and the world-wide allocation of OSS-activities, we have a quick look at the differences between active and non-active developers. It turns out that on average only 19.88% of the registered developers were active in 2006 (for the 1-quantile, 2-quantile, and 3-quantile the share of active per all developers is given by 12.5%, 18.68%, and 23.53%). These facts supports the idea that being a registered OSS developer and being an active developer is not the same thing. Clearly it is of more interest to know where the developers live who are indeed active. In addition, focusing on the number of developers (including the non active ones) might be misleading. Especially if such data is used for country-specific research or policy advice, or for cross-country studies analyzing the impact of country-specific factors on OSS etc. Therefore we are interested in the OSS *activities* and *active developers* respectively.

First we look at the share of activities that come from different regions. Therefore we analyze the number of active developers (per capita) and of activity level—i.e. number of posted messages weighted by the number of inhabitants—that can be assigned to different regions. Figure 4. presents the

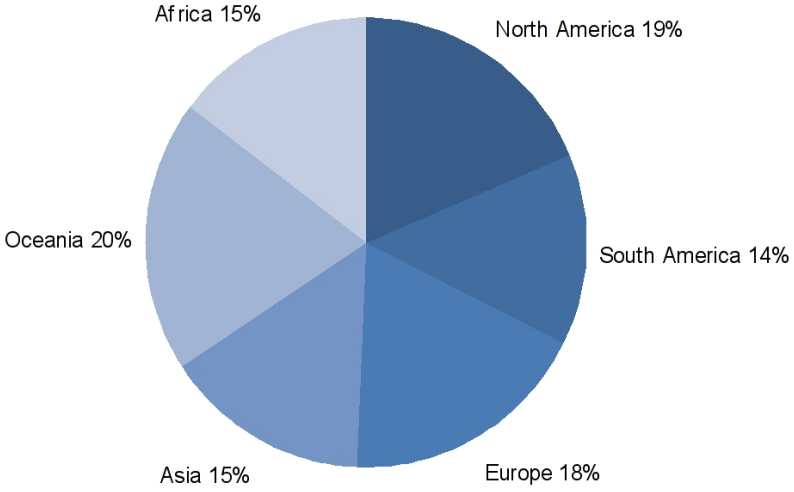
Figure 4: Activity Level (Messages per Inhabitants)



allocation of activity levels over the regions of the world. The allocation of active developers (per inhabitants) does not differ much from the shown allocation of

activity level. The reason for this fact is that the average activity level of an active developers is quite evenly spread over the regions as one can see in figure 5. Thus our first result is that although the number of active developers as well as activity levels is unequally allocated over the different regions, the average active developer does not differ that much (in terms of posted messages).

Figure 5: Average Activity of Active Developers



The fact that the activity of the average active developer does not differ much still holds when it comes to a comparison of countries rather than regions. To see this, compare figure 6 with figure 7. The two figures depict the worldwide allocation of active OSS developers and of the OSS activity-levels. The two maps look quite similar.

Furthermore, the figures 6 and 7 show that OSS activities differ over the world. There are some countries with a high degree of activity and thus a high number of active developers, opposed by a large number of countries with virtually zero active OSS developers. This fact is also expressed by figure 8. Here we illustrate the unequal distribution with a quantile plot of the activity level per country.

So far we have seen that only about every fifth registered developer was active in 2006, and while the average activity of active developers do not differ much both the number of active developers (per capita) as well as the level of activity per country differs strongly over the world. In addition, OSS seem to be a phenomena of the developed world: in 2006 85% of the active developers lived in one of the OECD countries hand have together posted 88% of all messages. Therefore one might guess that OSS is simply a rich countries' phenomena, thus

Figure 6: World Map of Active OSS Developers

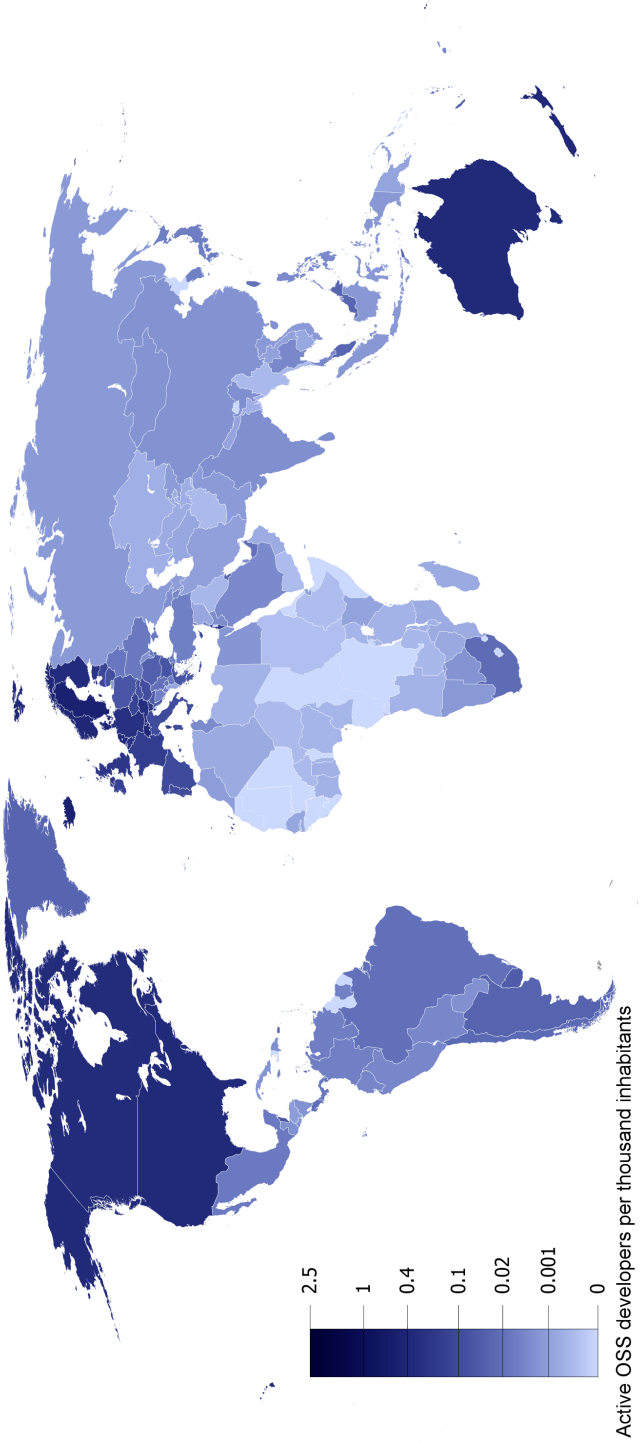


Figure 7: World Map of OSS Activity-Levels

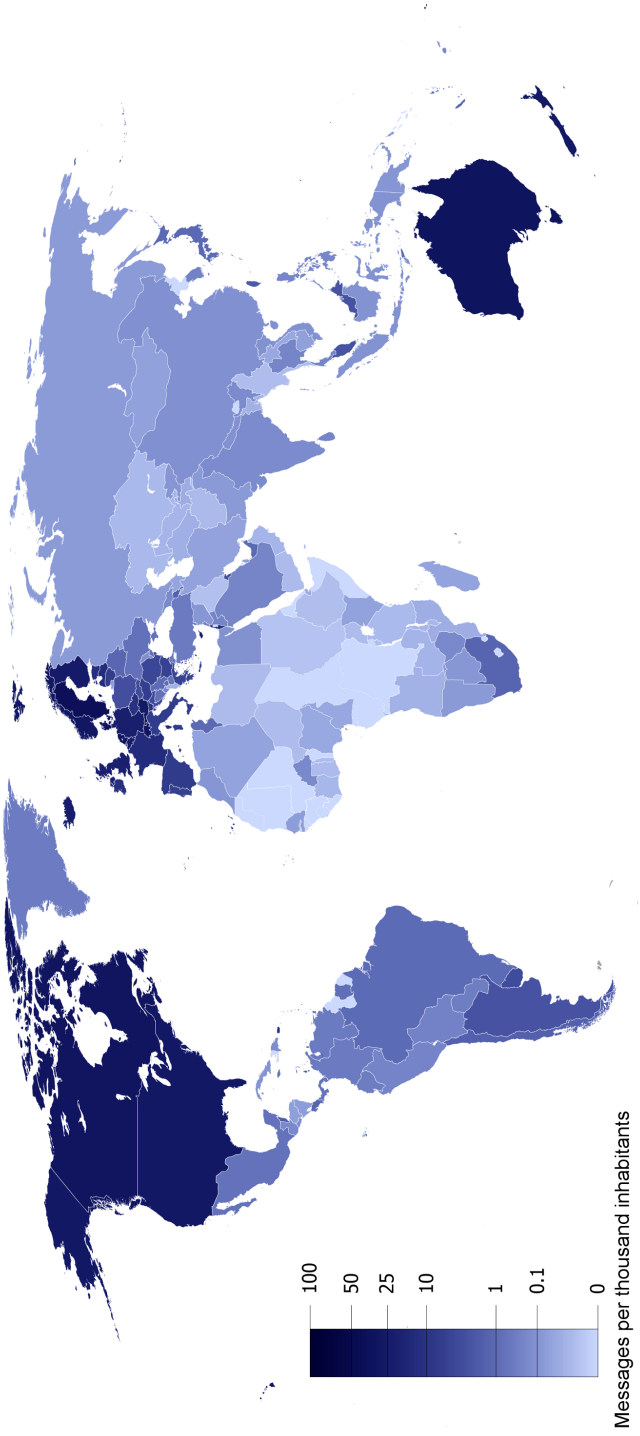
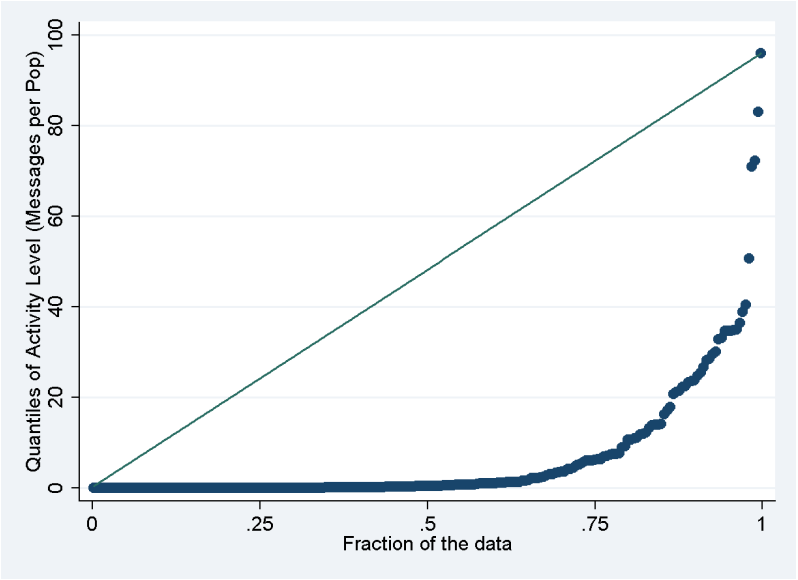


Figure 8: Distribution of Activity Levels (Quantile Plot)



correlated with GDP. To explore this, we weight our results by GDP per capita, i.e. we divide the number of active developers by the country's GDP per capita (purchasing power parity) for 2006, data source is World Bank (2007). Figure 9 shows the resulting world map of GDP-adjusted per-country number of active developers.[3] Compared to the two previous world maps the picture changes, but still the allocation is very unequal. Thus the phenomena of OSS cannot solely be explained by GDP per capita.

We also analyze the impact of another important factor: the internet. Having access to the internet is obviously a precondition for OSS as the collaborative way of OSS development is organized via internet. People meet on virtual platforms where they discuss and decide tasks. Software is up- and downloaded etc. With respect to the database we use for our analysis it is very clear: without having access to the internet there is simply no way to become a registered developer at SourceForge. Therefore we have to take this into account. We use data from the International Telecommunication Union (2006), as this source does provide data about the number of internet user, i.e. about the 'internet-population' of a country. We compute the number of active developers per internet user as well as activity per internet user. Figure 10 presents the results

[3]The patterned areas are countries whose GDP per capita in 2006 is not available.

Figure 9: World Map of Active OSS Developers weighted by GDP per capita

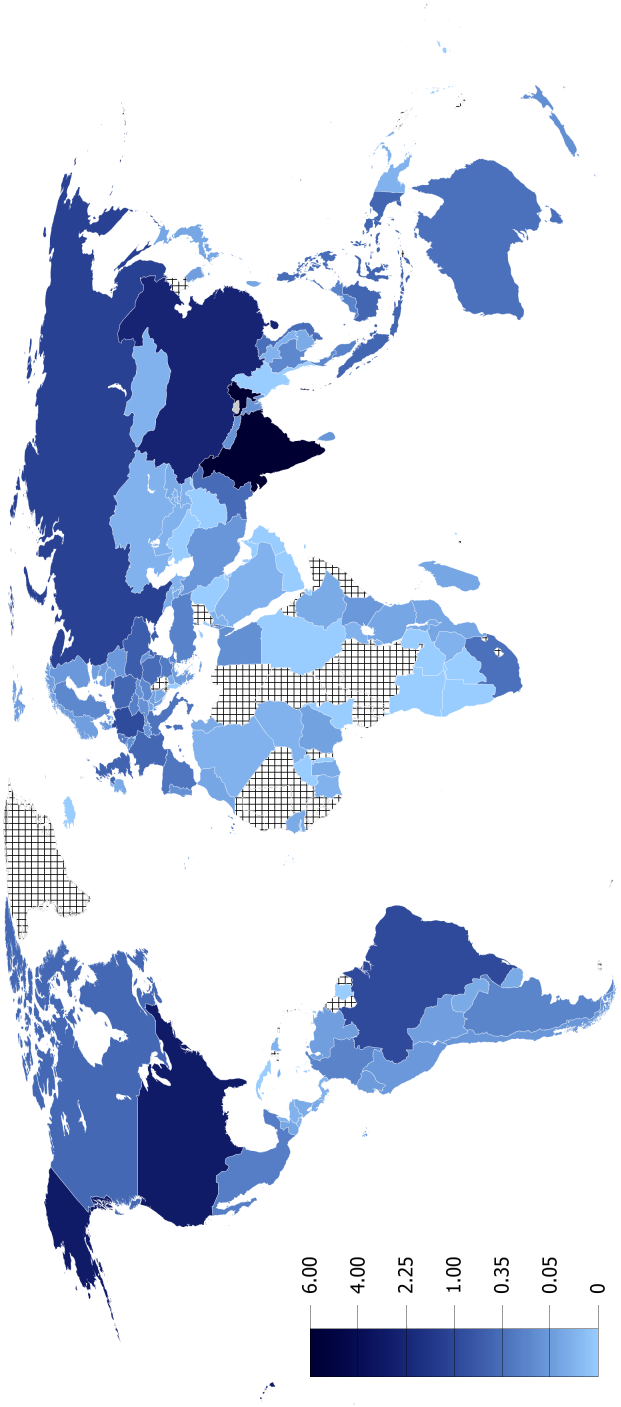
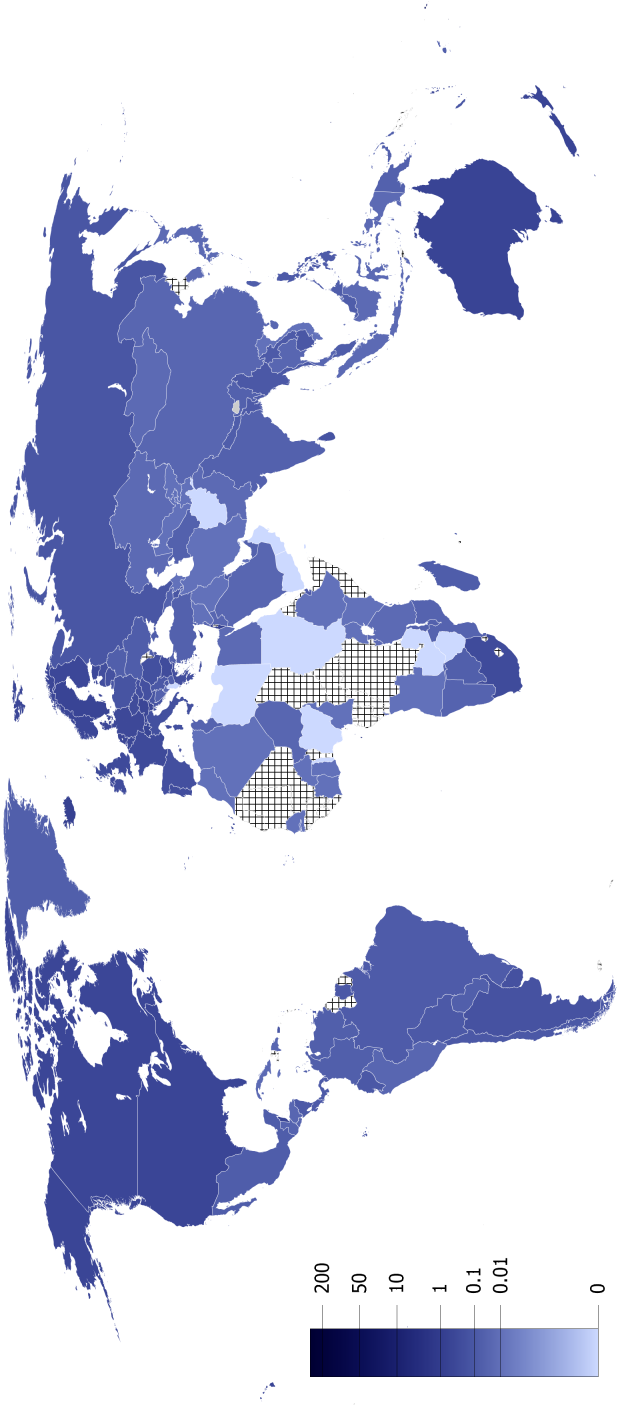


Figure 10: World Map of Active OSS Developers per Thousand Internet Users



for active developers per internet user[4] (the results for activity per internet user are again quite similar). Compared to the active developers per GDP (figure 9), or the numbers of figure 6 and 7, the OSS-active share of the internet-population is more equally allocated over the world. Nevertheless there remain differences in the share of OSS developers related to the internet-population of countries. This indicates that internet usage alone can not explain differences in world-wide OSS activities.

5 Summary and Outlook

Data about the supply-side of the software industry typically ignore the—for the most part non-paid—OSS activities. Reliable data that distinguish between registered OSS developers, active OSS developers, and OSS activity level can help to correct here. A more complete picture is of importance for both, policy markers and businesses. Furthermore, such country data is a valuable stock for cross-country studies and further research on the supply-side of OSS.

Analyzing the data about developer's IP address, email address and indicated time-zone from the SourceForge Research Data Archive, enables us to geographically identify 94% of all registered OSS developers in 2006. With the information about the number of posted messages we have a good proxy for activity of each developer. Based on this we analyze the world-wide allocation of OSS activities. Geographic origin seems to matter as the allocation of active OSS developers (and of OSS activities) is unequal. This still holds if one weights the absolut numbers by population or GDP per capita. And even if we relate it to the 'internet-population' i.e. the number of internet users, countries still differ.

As the worldwide allocation of OSS activities is not solely related to GDP or number of internet users, the question arises which further factors have an impact on OSS. Particularly, cultural and institutional aspects are potential candidates for factors that shape OSS activities. Analyzing this should also help to get a better understanding of what OSS is about. Our data are a good basis for such research, as we have country-specific informations about the number of OSS developers , the number of active OSS developers, and the level of OSS activity. In von Engelhardt & Freytag (2010) we therefore undertake such an analysis and examine the impact of country-specific cultural and institutional factors on the number developers, active developers and activity.

[4] The patterned areas are those countries with lack of data regarding the number of internet users.

References

Au, Y. A., Carpenter, D., Chen, X. & Clark, J. G. (2009), 'Virtual organizational learning in open source software development projects.', *Information & Management* **46**(1), 9 – 15.

den Besten, M., Dalle, J.-M. & Galia, F. (2008), 'The allocation of collaborative efforts in open-source software', *Information Economics and Policy* **20**(4), 316 – 322.

Comino, S., Manenti, F. M. & Parisi, M. L. (2007), 'From planning to mature: On the success of open source projects.', *Research Policy* **36**(10), 1575 – 1586.

David, P. A. & Rullani, F. (2008), 'Dynamics of innovation in an "open source" collaboration environment: lurking, laboring, and launching FLOSS projects on SourceForge.', *Industrial & Corporate Change* **17**(4), 647 – 710.

David, P. A., Waterman, A. & Arora, S. (2003), FLOSS-US. the free/libre/open source software survey for 2003, Technical report, Stanford Institute for Economic Policy Research.

Edelman, B. (2002), 'Registrations in Open ccTLDs', Online Article, http://cyber.law.harvard.edu/archived_content/people/edelman/open-cctlds/.

Eilhard, J. (2008), Loose contracts, tight control? Firms on SourceForge, CERNA Working Paper.

von Engelhardt, S. & Freytag, A. (2010), Institutions, culture, and open source, Jena Economic Research Papers in Economics 2010-010, Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics, http://www.jenecon.de.

Fershtman, C. & Gandal, N. (2008), Microstructure of collaboration: The 'social network' of open source software, CEPR Discussion Papers 6789.

Ghosh, R. A. (2006), Economic impact of open source software on innovation and the competitiveness of the information and communication technologies (ICT) sector in the EU, Study, European Commission.

Ghosh, R. A., Glott, R., Krieger, B. & Robles, G. (2002), FLOSS final report, part 4: Survey of developers, Technical report, International Institute of Infonomics, University of Maastricht.

Giuri, P., Ploner, M., Rullani, F. & Torrisi, S. (2010), 'Skills, division of labor and performance in collective inventions: evidence from open source software', *International Journal of Industrial Organization* **28**(1), 54 – 68.

Giuri, P., Rullani, F. & Torrisi, S. (2008), 'Explaining leadership in virtual teams: The case of open source software', *Information Economics and Policy* **20**(4), 305 – 315.

Gonzalez-Barahona, J. M., Robles, G., Andradas-Izquierdo, R. & Ghosh, R. A. (2008), 'Geographic origin of libre software developers.', *Information Economics and Policy* **20**(4), 356 – 363.

International Telecommunication Union (2006), ICT statistics 2006, part 4 – internet indicators: subscribers, users and broadband subscribers, ICT Statistics, ITU, http://www.itu.int/ITU-D/ICTEYE/Reports.aspx.

Johnson, J. P. (2002), 'Open source software: Private provision of a public good', *Journal of Economics & Management Strategy* **11**(4), 637–662.

von Krogh, G., Spaeth, S. & Lakhani, K. R. (2003), 'Community, joining, and specialization in open source software innovation: A case study', *Research Policy* **32**(7), 1217 – 1241.

Lancashire, D. (2001), 'Code, culture and cash: The fading altruism of open source development', *First Monday* **6**(12), http://firstmonday.org.

Lerner, J., Pathak, Parag, A. & Tirole, J. (2006), 'The dynamics of open-source contributors', *The American Economic Review* **96**(2), 114–118.

Lerner, J. & Tirole, J. (2005), 'The scope of open source licensing', *Journal of Law, Economics, and Organization* **21**(1), 20 – 56.

Madey, G. (ed.), The SourceForge research data archive (SRDA), A repository of FLOSS research data, University of Notre Dame, http://srda.cse.nd.edu.

Robles, G. & Gonzalez-Barahona, J. M. (2006), Geographic location of developers at SourceForge, *in* 'MSR '06: Proceedings of the 2006 international workshop on Mining software repositories', ACM, pp. 144–150.

Robles, G., Scheider, H., Tretkowski, I. & Weber, N. (2001), Who is doing it? A research on libre software developers, Technical report, Technische Universität Berlin.

World Bank (2007), *World Development Indicators 2007*.

Xu, J., Christley, S. & Madey, G. (2006), Application of social network analysis to the study of open source software, *in* J. Bitzer & P. J. Schröder, eds, 'The Economics of Open Source Software Development', Elsevier Press, pp. 247–270.