How researchers study open source and what we've found so far

Megan Squire @MeganSquire0

Elon University & FLOSSmole / FLOSSdata / FLOSShub / FLOSSpapers

April 6, 2016

What are YOUR unanswered questions about FLOSS?

Free, Libre, and Open Source Software

Ours include ...

... how software is made

... how software can be better

... how software can be more efficient

... how software can be higher quality

... how software can be lower cost

... how software can be more fun

... how groups work

... how decisions are made

... how developers talk to each other

... and more

To help with this, I collect, clean, and store data.

FLOSSmole

FLOSSdata

FLOSShub

FLOSSpapers

Sometimes I also analyze it too, but I'm not as good at that.

What is the "data"?

(1) Surveys or case studies

2002, [link]
2013, [link]

Do you earn money from FLOSS, either directly or indirectly?		
Answer	**Count**	**Percentage**
No (1)	641	29.36%
Yes, directly: I am paid for developing FLOSS. (2)	375	17.18%
Yes, directly: I am paid for supporting FLOSS. (3)	113	5.18%
Yes, directly: I am paid for administrating FLOSS. (4)	115	5.27%
Yes, directly: Other reasons (5)	105	4.81%
Yes, indirectly: I got my job because of my previous FLOSS experience (6)	141	6.46%
Yes, indirectly: my job description does not include FLOSS development but I also develop FLOSS in my work (7)	186	8.52%
Yes, indirectly: Other reasons (9)	146	6.69%
No answer	30	1.37%
Not completed or Not displayed	331	15.16%

2013, [link]

Surveys can help us answer questions about motivations.

"Why do you develop open source?"

"Do you feel valued?"

Some of the most oft-cited facts
about FLOSS
are based on survey data.

What percentage of FLOSS developers identify as female?

Some of the most venerable "laws" in software development were based on case studies where n=1.

Conway's law, Brooks' law.

But surveys also have issues.

(2) Artifact-based research

(2) Artifact-based research

Due to my computing background, I'm strongly biased towards artifact-based research.

Artifact: Source code

```
#ifndef    __MATH_EMU_DOUBLE_H__
#define    __MATH_EMU_DOUBLE_H__

#if _FP_W_TYPE_SIZE < 32
#error "Here's a nickel kid.  Go buy yourself a real computer."
#endif
```

2005, [link]

Maintainability of the kernels of open-source operating systems: A comparison of Linux to FreeBSD, NetBSD, and OpenBSD

Liguo Yu,[a,1] Stephen R. Schach,[a,*] Kai Chen,[a] Gillian Z. Heller,[b] Jeff Offutt[c]

2006, [link]

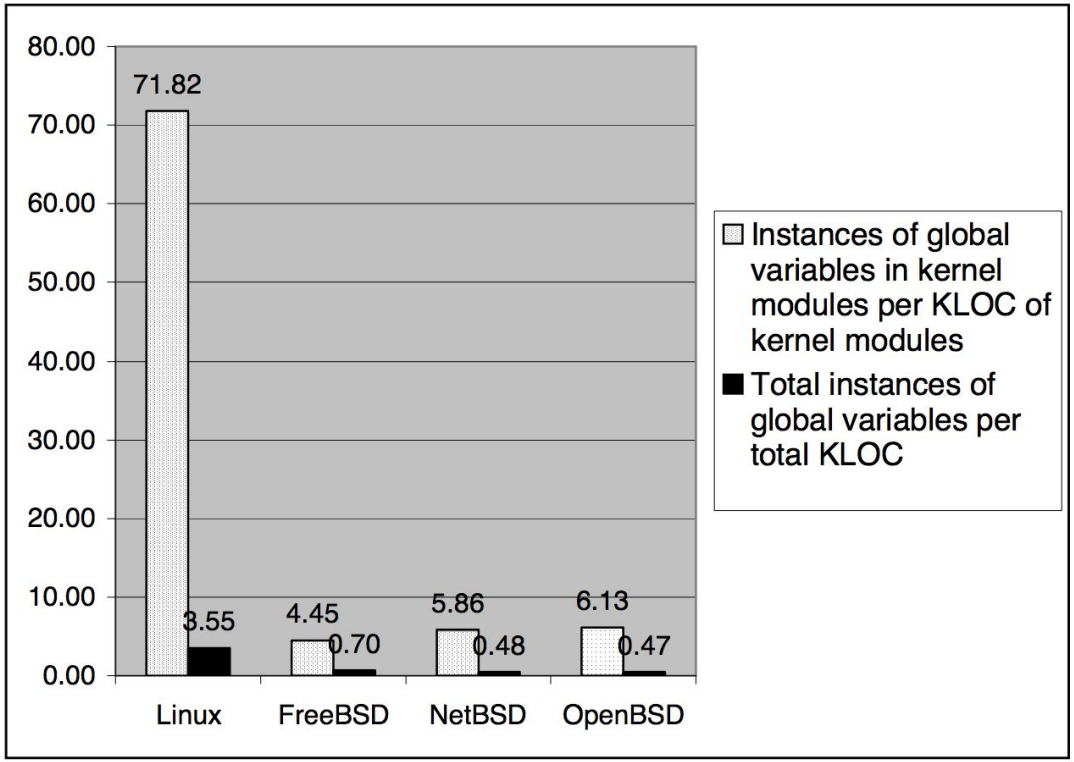


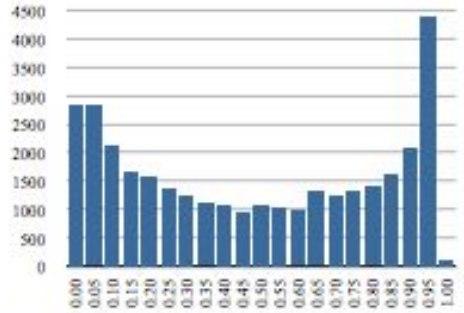
Figure 6: Instances of global variables per KLOC.

2006, [link]

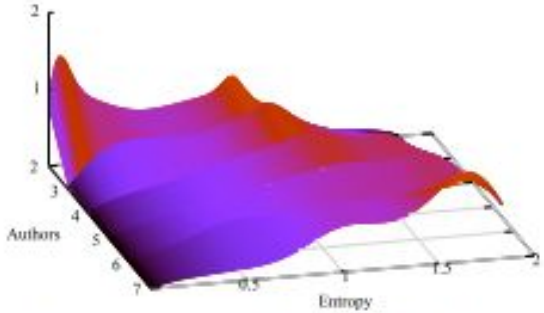
Author Entropy: A Metric for Characterization of Software Authorship Patterns

Quinn C. Taylor, James E. Stevenson, Daniel P. Delorey, and Charles D. Knutson
SEQuOIA Lab, Brigham Young University
2236 TMCB
Provo, Utah 84602
{qtaylor, jstevenson}@byu.net, {pierce, knutson}@cs.byu.edu

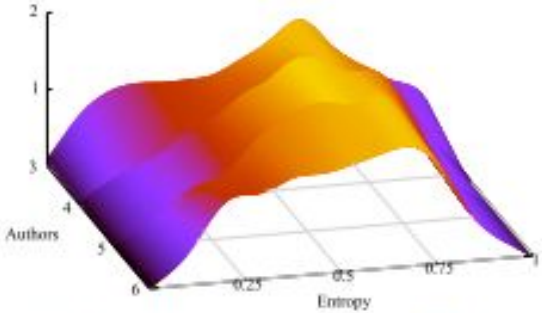
2008, [link]



(a) Entropy distribution, 2 authors.



(b) Entropy distributions, 2–7 authors.



(c) Normalized entropy, 3–6 authors.

Author entropy is a summary statistic that characterizes contribution patterns in source code. Entropy is easy to calculate, and can be calculated for different levels of granularity (e.g., lines, methods, files, modules). While author entropy does not directly imply a level of code quality, it can be used in conjunction with other software metrics to identify potential areas of concern within the source code of a project.

2008, [link]

Sentiment Analysis of Commit Comments in GitHub: An Empirical Study

Emitza Guzman, David Azócar, Yang Li
Technische Universität München
Faculty of Informatics
Garching, Germany
emitza.guzman@mytum.de, dazocar@gmail.com, liya@in.tum.de

2014, [link]

days tends to a more negative emotion. A Wilcoxon rank sum test of Monday against each of the other days confirmed that commit comments were more negative on Monday than on Sunday, Tuesday, and Wednesday (p-value<=0.015). No

2014, [link]

We performed a Wilcoxon rank sum test which confirmed that commit comments from projects written in Java are more negative than projects implemented in six other languages (p-value<=0.002), namely C, C++, JavaScript, PHP, Python and Ruby. Statistical tests on the emotion scores of

2014, [link]

Artifact: Licenses

A Large Scale Study of License Usage on GitHub

Christopher Vendome
The College of William and Mary,
Williamsburg, VA, 23185
cvendome@cs.wm.edu

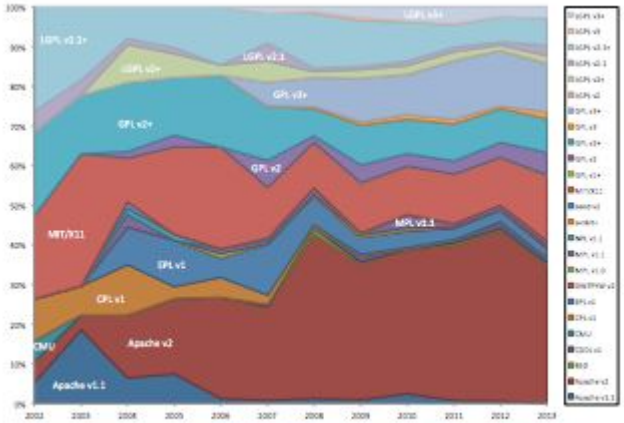
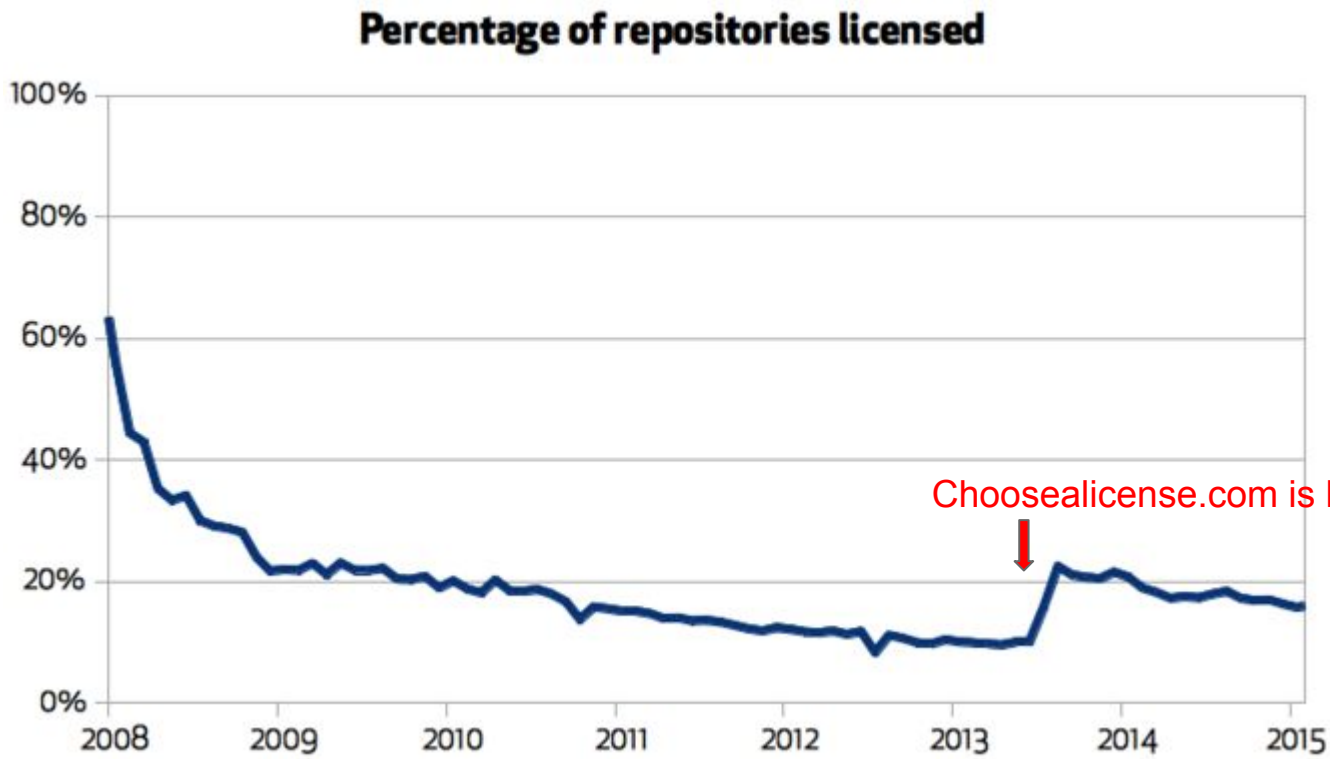


Fig. 1. Relative License Usage between 2002 and 2012.

2015, [link]

Open source license usage on GitHub.com

March 9, 2015 benbalter Watercooler



2015 [link]

Data Sets: The Circle of Life in Ruby Hosting, 2003-2015

Megan Squire
Elon University
Elon, NC 27244 USA
msquire@elon.edu

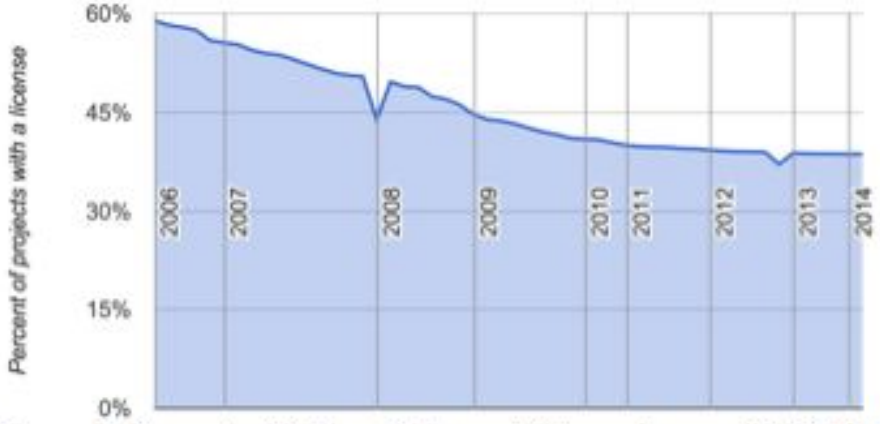


Figure 4. Percent of RF projects specifying a license, 2006-2013

Artifact: Email

Subject: Re: [PATCH 0/5] fuse: handle release synchronously (v4)
From: Linus Torvalds (torv...@linux-foundation.org)
Date: Oct 16, 2014 6:54:16 am
List: org.kernel.vger.linux-kernel

On Thu, Oct 16, 2014 at 3:43 PM, Miklos Szeredi <mik...@szeredi.hu> wrote:

> One idea is to change ->flush() so it's responsible for fput()-ing the file. That way we could take control of the actual refcount decrement. There are only 20 flush instances in the tree, so it wouldn't be a huge change.

Since that *still* wouldn't fix the problem with the whole "count elevated by other things" issue, I really don't want to hear about these random broken hacks that are fundamentally broken crap.

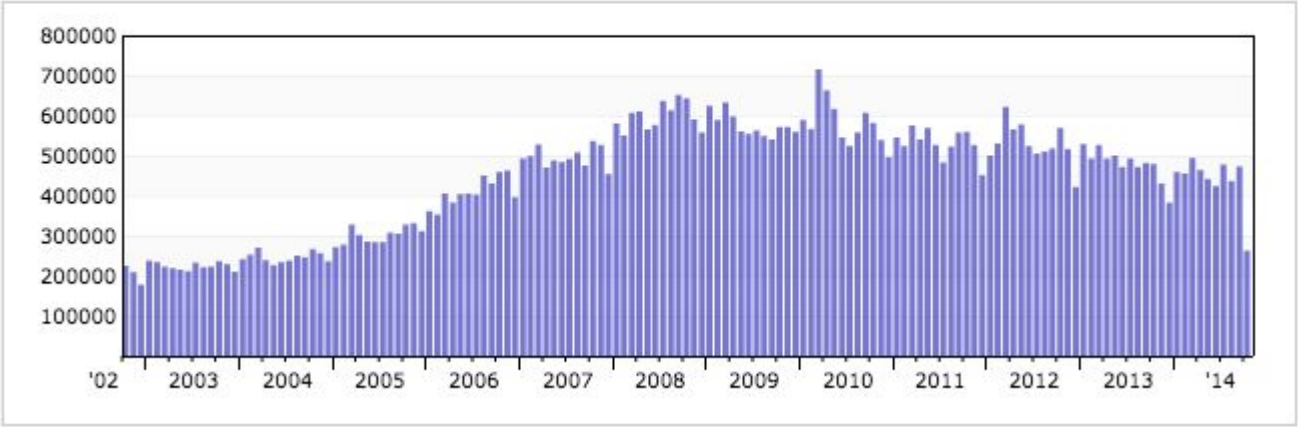
Really. Stop cc'ing me with "let's implement this hack that cannot work in general". I'm not interested. There's a reason we don't do this. We don't make up random hacks that we know cannot work in the general case.

Linus

MarkLogic

Searching **8,836 lists** and **70,917,405 messages**. First list started in **November 1992**. There are **2,701 active lists**, recently accumulating **18,549 messages per day**. You can browse recent emails.

Traffic (messages per month):



GMANE

marc.info
debian lists
apache lists
etc.

Geographic Location of Developers at SourceForge*

Gregorio Robles
grex@gsyc.escet.urjc.es

Jesus M. Gonzalez-Barahona
jgb@gsyc.escet.urjc.es

Grupo de Sistemas y Comunicaciones
Universidad Rey Juan Carlos
Mostoles, Spain

2006 [link]

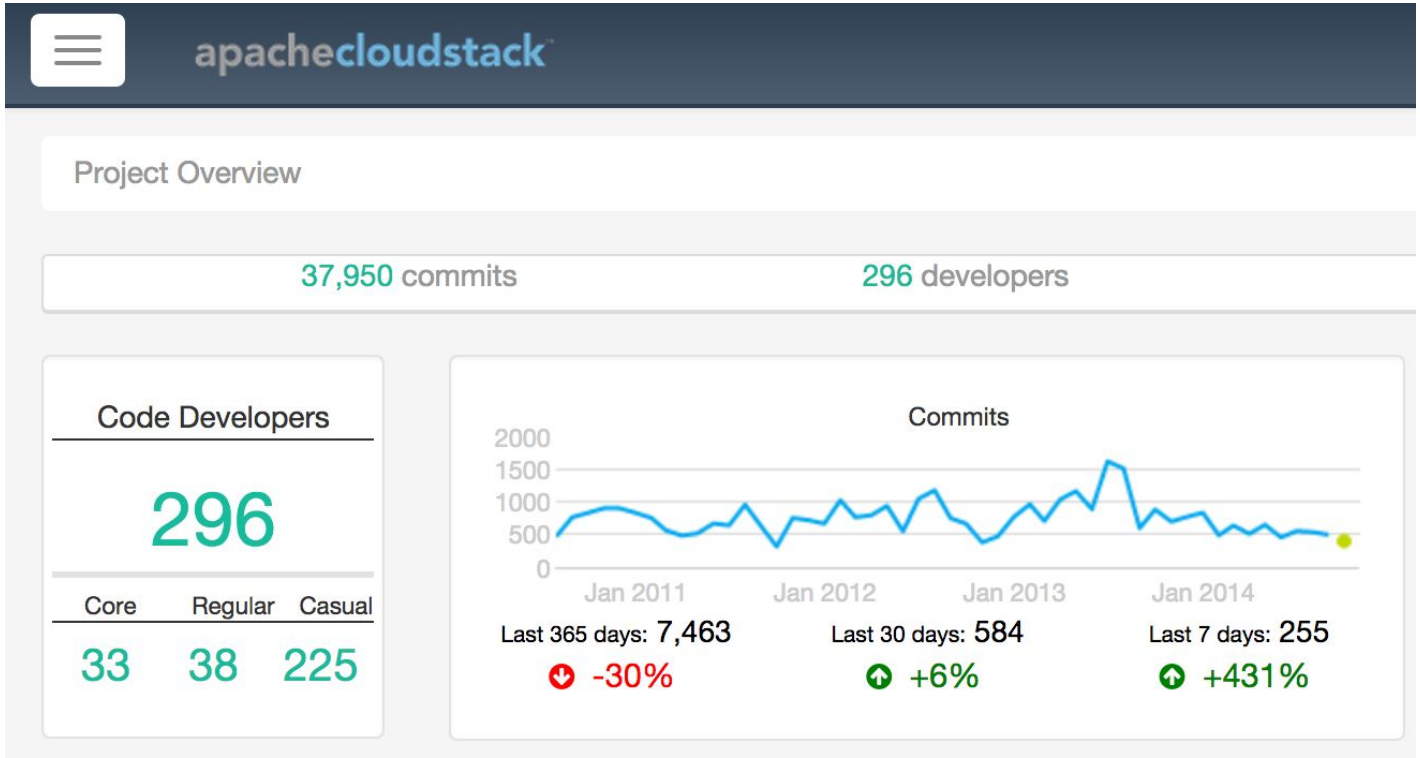
The final goal of the methodology described in this section is to estimate, as accurately as possible, the geographical distribution of the users in the database, using the domain in their e-mail address and the time zone as the base for the analysis.

Domain	Number
hotmail.com	63784
yahoo.com	40180
gmail.com	14191
aol.com	6275
gmx.net	4128
msn.com	3688
163.com	2013
ntlworld.com	1998
rr.com	1981
rediffmail.com	1881

Rank	Country	Developers
1.	United States	425620
2.	Germany	95800
3.	United Kingdom	60768
4.	Canada	49109
5.	France	44587
6.	China	36517
7.	Australia	31812
8.	Italy	30763
9.	Netherlands	29335
10.	Sweden	23867

Bitergia

OpenSource Software Development analytics and metrics



current [link]

The Diffusion of Pastebin Tools to Enhance Communication in FLOSS Mailing Lists

Megan Squire, Amber K. Smith

Elon University, Elon, NC, USA
{msquire,asmith90}@elon.edu

2015 [link]

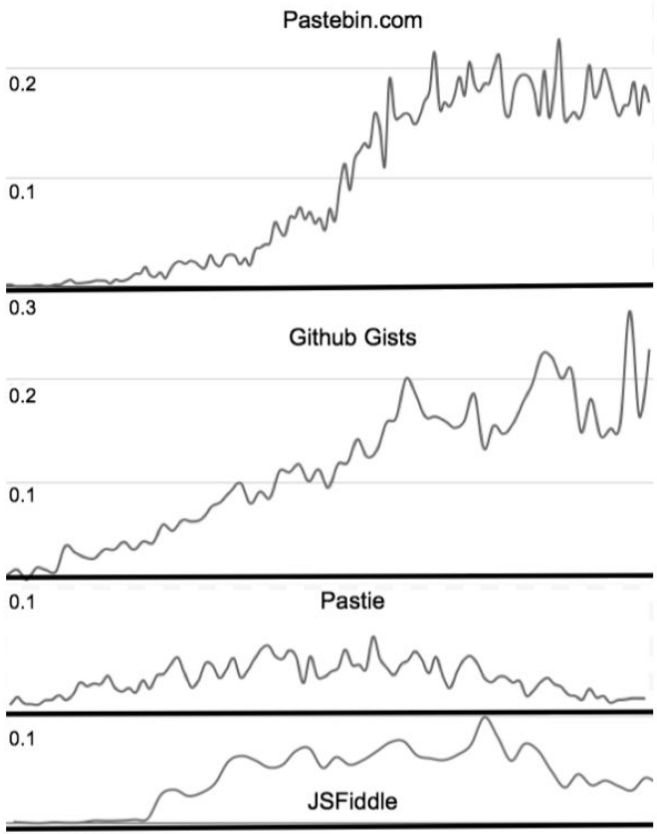


Figure 3. Comparison of pastebin tool references as a percentage of all mail sent, May 2003 - Feb 2014

2015 [link]

What can OSS mailing lists tell us?
A preliminary psychometric text analysis of the Apache developer mailing list

Peter C. Rigby *
Software Engineering Group
University of Victoria, B.C., Canada
pcr@uvic.ca

Ahmed E. Hassan
Dept. of Electrical and Computer Engineering
University of Victoria, B.C., Canada
ahmed@ece.uvic.ca

2007, [link]

Subject	Neuro	Extro	Open	Agree	Consc
Num. of Clusters	1	3	3	2	3
A	8.10^{1}	$2.15^{1,2}$	22.47^{1}	-9.20^{2}	$-4.07^{2,3}$
B	8.79^{1}	2.61^{2}	19.30^{1}	-10.46^{1}	-4.24^{2}
C	7.35^{1}	5.95^{3}	32.84^{2}	-10.58^{1}	-4.34^{2}
D	7.86^{1}	0.53^{1}	44.18^{3}	-11.12^{1}	-5.36^{1}
≤ 30	7.07^{1}	6.41^{3}	32.87^{2}	-11.15^{1}	-3.42^{3}
> 30	7.33^{1}	7.61^{3}	32.85^{2}	-10.87^{1}	-4.24^{2}

Table 2. Composite measure of Personality. Values are relative, not absolute. Post Hoc Tukey's HSD Test ($\alpha = 0.05$). Superscript represents Tukey cluster number.

2007, [link]

Personality. Using the LIWC dimensions that are correlated with the big five personality traits, we created a baseline personality score for the entire mailing list (excluding the top four committers) and compared the top four committers to the baseline and to each other. The two developers that were responsible for two major Apache releases had similar personalities. Their personalities were different from the baseline and other developers on the traits of extroversion and openness. We plan to run further comparisons with other projects as well as examine the effect of a developer's role on his or her personality.

2007, [link]

Personality. Using the LIWC dimensions that are correlated with the big five personality traits, we created a baseline personality score for the entire mailing list (excluding the top four committers) and compared the top four committers to the baseline and to each other. The two developers that were responsible for two major Apache releases had similar personalities. Their personalities were different from the baseline and other developers on the traits of extroversion and openness. We plan to run further comparisons with other projects as well as examine the effect of a developer's role on his or her personality.

2007, [link]

Exploring the role of outside organizations in Free / Open Source Software projects

Darren Forrest, Carlos Jensen, Nitin Mohan, and Jennifer Davidson

School of EECS, Oregon State University, Corvallis OR 97331, USA

2012, [link]

data sources: bug reports, code commits

Table 1. GCC code contribution and bug reporting (top 20 domains)

Unique code contributors		Unique bug reporters	
redhat.com	150	gnu.org	174
gnu.org	104	redhat.com	61
ibm.com	70	ibm.com	55
adacore.com	55	debian.org	46
codesourcery.com	47	sourceforge.net	35
google.com	38	mit.edu	27
apple.com	30	acm.org	26
suse.com	29	intel.com	24
gnat.com	23	hp.com	19
intel.com	17	mpg.de	17
amd.com	14	cmu.edu	16
arm.com	14	berkeley.edu	16
sourceforge.net	12	apple.com	15
debian.org	10	nasa.gov	15
inria.fr	9	utexas.edu	14
ispras.ru	9	cern.ch	14
st.com	8	stanford.edu	13
acm.org	7	suse.com	13
hp.com	7	gentoo.org	13
kpitcummins.com	6	kth.se	12

Table 2. Linux Kernel code contribution and bug reporting (top 20 domains)

Unique code contributors		Unique bug reporters	
ibm.com	721	ibm.com	115
intel.com	571	osdl.org	112
fujitsu.com	478	intel.com	47
redhat.com	409	gentoo.org	36
kernel.org	367	redhat.com	32
google.com	228	sourceforge.net	30
ti.com	209	debian.org	26
sgi.com	203	suse.com	22
linutronix.de	187	hp.com	18
novell.com	145	kernel.org	13
suse.com	132	bigfoot.com	12
amd.com	130	linux.com	12
freescale.com	125	mit.edu	11
nokia.com	104	hut.fi	10
hp.com	96	ubuntu.com	9
atheros.com	89	amd.com	9
samsung.com	88	fujitsu.com	9
infradead.org	83	cornell.edu	8
mvista.com	81	ieee.org	8
oracle.com	78	tudelft.nl	7

2012, [link]

We found that for these large projects, corporate developers dominate in terms of code contributions. This has important implications for project governance and our understanding of FOSS demographics. Large projects may not be accurately portrayed as grass-roots volunteer efforts.

2012, [link]

The data suggests there exist two distinct communities within projects. While these communities may interact with each other through other means (e.g. mailing lists), there is a community of coders and a community of bug reporters. While this is not unexpected, it is unexpected to see that the most prolific code contributors seem not to interact with the bug reporters—we tracked any participation in bug reporting, not just the reporting of new bugs. This disconnect can in the long-term lead to alienation and declining participation of non-technical contributors.

2012, [link]

Artifact: IRC archives

<kurtis> Guys I'm looking at https://code.djangoproject.com/ticket/7591 and it says that it is very easy
<joshuajonah> kurtis, it's not as easy as it seems
<joshuajonah> the length issue is the problem
<KBme> can anyone tell me how to get the app name and model name in the admin template?
<j00bar> joshuajonah: that's what she said
<joshuajonah> :d

2015 [link]

ubuntu	irc	11/18/2004	Hoodster: you'll figure it out - it was designed for your **grandma** ;)
ubuntu	irc	1/17/2005	there's no beating ubuntu install though, excellent... and the layout of the gui is nice... all that's left is to improve hardware detection a little bit and ubuntu just may become a word **grandma** knows
linux-kernel	ML	1/14/2002	If it screws up, and **Aunt Tillie** shelled out for support (which of course, she did being the 'needing support' type)
linus-kernal	ML	1/14/2002	Yes, and yes. **Aunt Tillie** is running Linux because someone installed a distribution for her.

Building data sets for automatic detection & study of...

1. Profanity & profanity obfuscation
    a. strong
    b. mild
2. General Insults
    a. personal
    b. code-based
    c. both
3. Gender-based language & attitudes
    a. your mom / maternal insults
    b. that's what she said
    c. gender stereotyping
        i. Aunt Tillie / grandma
        ii. wives/girlfriends don't let you code!

2015, to appear

I'm Not Chatting, I'm Innovating!

Locating Lead Users in

Open Source Software Communities

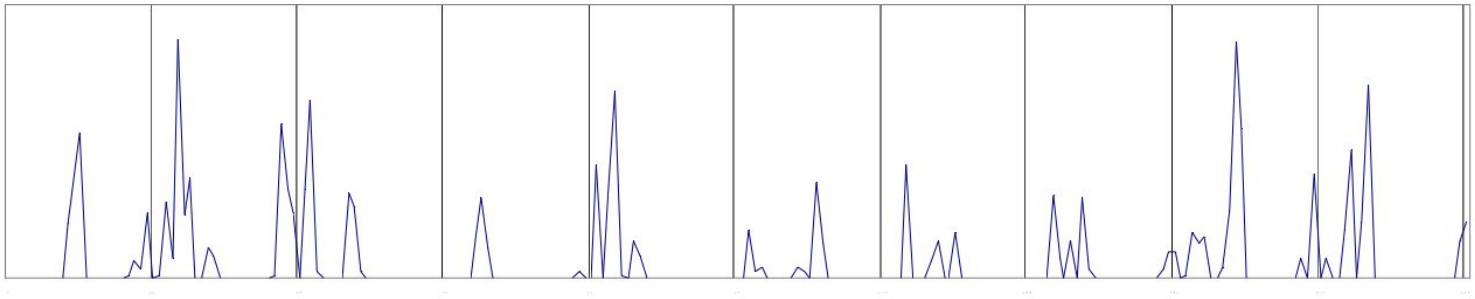


Chart 1: Exemplar non-accumulated centrality data for one participant over ten consecutive days

Geoff Breach, 2008, [link]

Digesting Virtual "Geek" Culture: The Summarization of Technical Internet Relay Chats

Liang Zhou and Eduard Hovy
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{liangz, hovy} @isi.edu

Benjamin Reed wrote a wireless Ethernet **driver that used /proc** as its interface. But he was a little uncomfortable … asked if there were any conventions he should follow. He added, "and finally, what's up with **sysctl**? …"

Linus Torvalds replied with: "the thing to do is to create a …[program code]. The **/proc/drivers/** directory is already there, so you'd basically do something like … [program code]." For the **sysctl** question, he added "sysctl is deprecated. ..."

Marcin Dalecki flamed Linus: "Are you just blind to the never-ending format/compatibility/… problems the whole idea behind **/proc** induces inherently? …[example]"

Figure 3. An original Kernel Traffic digest.

Benjamin Reed wrote a wireless Ethernet **driver that used /proc** as its interface. But he was a little uncomfortable … asked if there were any conventions he should follow. He added, "and finally, what's up with **sysctl**? …"
Linus Torvalds replied with: "the thing to do is to create a …[program code]. The **/proc/drivers/** directory is already there, so you'd basically do something like … [program code]." For the **sysctl** question, he added "sysctl is deprecated. ..."
Marcin Dalecki flamed Linus: "Are you just blind to the never-ending format/compatibility/… problems the whole idea behind **/proc** induces inherently? …[example]"

Figure 3. An original Kernel Traffic digest.

[0|0] *Benjamin Reed*: "I wrote an … driver … /proc …"
[0|1] *Benjamin Reed*: "… /proc/ guideline …"
[0|2] *Benjamin Reed*: "… syscyl …"
[1|0] *Linus Torvalds* responds to [0|0, 0|1, 0|2]: "the thing to do is …" "sysctl is deprecated … "

Figure 5. A short example from Baseline 2.

Artifact: Bug reports

Bug 191744 - lang/python27: With THREADS option: devel/pth: pthread.h conflicts with system pthread.

Status: Issue Resolved FIXED

Product: Ports Tree
Component: Individual Port(s)
Version: Latest
Hardware: i386 Any

Importance: --- Affects Many People
Assigned To: Marcus von Appen

URL: https://phabric.freebsd.org/D488
Keywords:

Duplicates: ~~190470~~ ~~190483~~ ~~190496~~ ~~190569~~ ~~190588~~ ~~190719~~ ~~191033~~ ~~191061~~ ~~191062~~ ~~191612~~ ~~191761~~ (view as bug list)
Depends on: ~~191888~~
Blocks:

Show dependency tree / graph

Reported: 2014-07-08 18:07 UTC by mikhail.rokhin
Modified: 2014-08-03 07:53 UTC (History)
CC List: 3 users (show)

See Also: ~~175390~~ ~~189844~~ ~~187979~~

What Topics do Firefox and Chrome Contributors Discuss?

Mario Luca Bernardi
Dept. of Engineering,
University of Sannio, Italy
mlbernar@unisannio.it

Carmine Sementa
Dept. of Engineering,
University of Sannio, Italy
csementa@unisannio.it

Quirino Zagarese
Dept. of Engineering,
University of Sannio, Italy
quirino.zagarese@unisannio.it

Damiano Distante
Fac. of Economics, Unitelma
Sapienza Univ., Italy
distante@unitelma.it

Massimiliano Di Penta
Dept. of Engineering,
University of Sannio, Italy
dipenta@unisannio.it

Table 4: Overlapping topics in the same semester.

Semester	Overlapping terms in the topic
2008 S2	movi; youtub; stop; video; player; game; plai; flash; sound
2008 S2	left; width; height; border; px
2008 S2	cpu; task; usag; slow; hang
2008 S2	hit; shift; tag; keyboard; focu
2008 S2	usernam; account; login; email; authent
2009 S1	youtub; video; player; plai; flash
2009 S1	width; background; bottom; size; posit
2009 S2	left; mous; bottom; posit; screen
2010 S1	left; width; resiz; bottom; height; size; border; posit; px
2010 S1	restart; visit; login; comput; websit; hang

Artifact: Other communication media

Tweets Tweets & replies Photos & videos

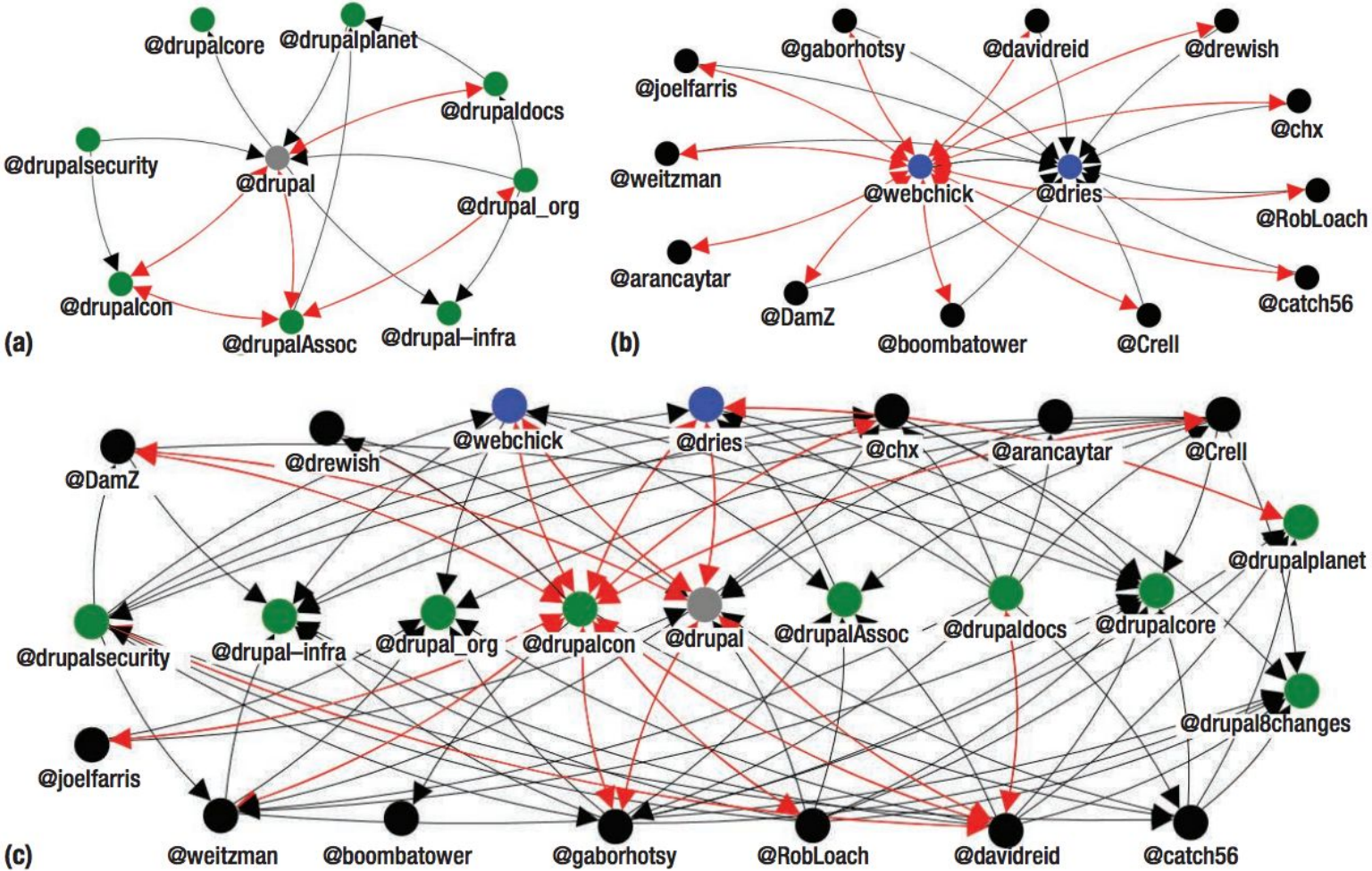
Apache Spark @ApacheSpark · Sep 11
Spark 1.1 is now out! With 171 contributors, it's our largest release yet.
spark.apache.org/releases/spark…
113 52

Apache Spark @ApacheSpark · Jul 18
Spark Summit videos are now up! See them at spark-summit.org/2014/agenda
58 58

Apache Spark @ApacheSpark · May 30
Spark 1.0 is now out! Huge congrats to everyone who contributed.
blogs.apache.org/foundation/ent…
spark.apache.org/releases/spark…
184 83

Apache Spark @ApacheSpark · Apr 9
Spark 0.9.1 released - spark.apache.org/releases/spark…
41 16

Wang, et al., 2014 [link]


@drupalcore
@drupalplanet
@drupaldocs
@drupalsecurity
@drupal
@drupal_org
@drupalcon
@drupalAssoc
@drupal–infra
(a)
@gaborhotsy
@davidreid
@drewish
@joelfarris
@chx
@weitzman
@webchick
@dries
@RobLoach
@arancaytar
@catch56
@DamZ
@boombatower
@Crell
(b)
@webchick
@dries
@chx
@arancaytar
@Crell
@DamZ
@drewish
@drupalplanet
@drupalsecurity
@drupal–infra
@drupal_org
@drupalcon
@drupal
@drupalAssoc
@drupaldocs
@drupalcore
@drupal8changes
@joelfarris
@weitzman
@boombatower
@gaborhotsy
@RobLoach
@davidreid
@catch56
(c)

Apache-Affiliated Twitter Screen Names: A Dataset

Megan Squire
Dept. of Computing Sciences
Elon University
Elon, NC, USA
msquire@elon.edu

2013, [link]

Artifact: Board Meeting Reports

Project Roles in the Apache Software Foundation: A Dataset

Megan Squire
Dept. of Computing Sciences
Elon University
Elon, NC, USA
msquire@elon.edu

2013 [link]

PMC changes:

 - Currently 16 PMC members.
 - No new PMC members added in the last 3 months
 - Last PMC addition was Raju Bairishetti on Wed July 19 2015

Committer base changes:

 - Currently 20 committers.
 - New committers :
   + Deepak Kumar Barr on Nov 07 2015
   + Pranav Kumar Agarwal on Oct 31 2015
   + Amruth S on Oct 23 2015
   + Sushil Mohanty on Oct 23 2015
 - Last committer addition was Deepak Kumar Barr at Sat Nov 07 2015

How has the community developed since the last report?

We have 2 new committers: Dylan Millikin and Ted Wilmes. Ted Wilmes has also joined the PPMC.

Artifact: Project Metadata

A Lightweight SQL Database for Cloud Infrastructure and Web Applications

Overview | Code | Bugs | Blueprints | Translations | Answers

Registered 2008-05-12 by Drizzle Developers

The Drizzle project is building a database optimized for Cloud and Net applications. It is being designed for massive concurrency on modern multi-cpu/core architecture. The code is originally derived from MySQL.

Project information

Part of:
Drizzle Umbrella Project

Maintainer:
Drizzle Developers

Driver:
Drizzle Developers

Development focus:
7.2 series

lp:drizzle
Browse the code

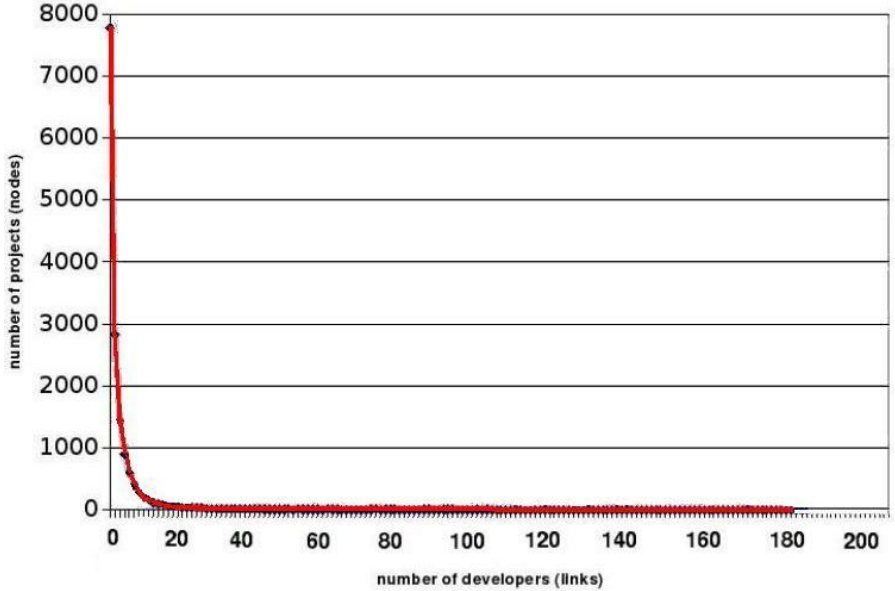
Programming Languages:
C++

Licences:
Simplified BSD Licence, GNU GPL v2
()

Do the Rich Get Richer?

The Impact of Power Laws on Open Source Development

O'Reilly Open Source Convention
July 26–30, 2004

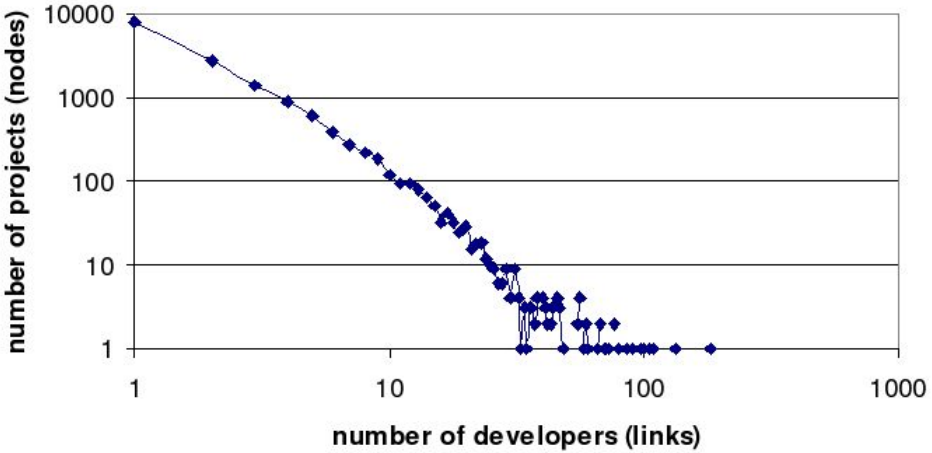


Data Source: Sourceforge.net

Data: Projects & Developers

Method: Social network analysis

Finding: OSS shows many characteristics of a scale-free network, but so far has stopped short of exhibiting winner-take-all behavior.



Artifact: Mixed data

Community, Joining, and Specialization in Open Source Software Innovation: A Case Study*

Georg von Krogh*,
Sebastian Spaeth*,
Karim R. Lakhani**

2003 [link]

"We gathered data from four different sources.

- "Firstly, we conducted **thirteen telephone interviews** in two rounds with eight Freenet developers

"We gathered data from four different sources.

- "Firstly, we conducted **thirteen telephone interviews** in two rounds with eight Freenet developers
- "Secondly, we collected the project's public **email** conversations stored in the projects' **mailing lists** which is archived on Freenet's website

"We gathered data from four different sources.

- "Firstly, we conducted **thirteen telephone interviews** in two rounds with eight Freenet developers
- "Secondly, we collected the project's public **email** conversations stored in the projects' **mailing lists** which is archived on Freenet's website
- "The third source of data included the **history of changes to the software code** available via the project's software repository within the **CVS**

"We gathered data from four different sources.

- "Firstly, we conducted **thirteen telephone interviews** in two rounds with eight Freenet developers
- "Secondly, we collected the project's public **email** conversations stored in the projects' **mailing lists** which is archived on Freenet's website
- "The third source of data included the **history of changes to the software code** available via the project's software repository within the **CVS**
- "Fourthly, in order obtain contextual understanding of the project we collected **publicly available documents** related to open source in general and to the project in particular..."

participants in the development e-mail list. We define a *joining script* as the level and type of activity a joiner goes through to become a member of the developer community. And therefore, joining scripts represents a cost to any would-be developer in the project.

One question,
multiple approaches

Gender Differences in Early Free and Open Source Software Joining Process

Victor Kuechler, Claire Gilbertson, and Carlos Jensen

School of Electrical Engineering and Computer Science
Oregon State University, Corvallis, OR, USA
kuechlej@onid.orst.edu, claire.gilbertson@gmail.com,
cjensen@eecs.oregonstate.edu

2012, [link]

ing statistics of female FOSS participants. New participants often experience their first interaction on a FOSS project's mailing list. We explored six FOSS projects – Buildroot, Busybox, Jaws, Parrot, uClibc, and Yum. We found a declining rate of female participation from the 8.27% of subscribers, to 6.63% of posters, and finally the often reported code contributor rate of 1.5%. We found a disproportionate attrition rate among women along every step of the FOSS joining process.

2012, [link]

ing statistics of female FOSS participants. New participants often experience their first interaction on a FOSS project's mailing list. We explored six FOSS projects – Buildroot, Busybox, Jaws, Parrot, uClibc, and Yum. We found a declining rate of female participation from the 8.27% of subscribers, to 6.63% of posters, and finally the often reported code contributor rate of 1.5%. We found a disproportionate attrition rate among women along every step of the FOSS joining process.

2012, [link]

Gendered Patterns of Politeness in Free/Libre Open Source Software Development

Eunyoung Moon
School of Information
University of Texas at Austin
eymoon@utexas.edu

2013, [link]

In this paper, we have shown that "practical" politeness during FLOSS development is linked to involving and sustaining women FLOSS developers and non-developers' participation. However, this

2013, [link]

Gender and Tenure Diversity in GitHub Teams

Bogdan Vasilescu[†§*], Daryl Posnett[†], Baishakhi Ray[†], Mark G.J. van den Brand[§], Alexander Serebrenik[§], Premkumar Devanbu[†], Vladimir Filkov[†*]
[†]University of California, Davis and [§]Eindhoven University of Technology
*vasilescu@ucdavis.edu, filkov@cs.ucdavis.edu

(1) We collected a data set comprising thousands of projects from GITHUB, capturing or inferring the contributions, gender, and tenure of each recorded participant. We also ran a survey on GITHUB, to get a sense of how participants assess and value team composition and diversity.

(2) We use statistical modeling to analyze the relationship of gender and tenure diversity to productivity, when controlling for team size and other confounds. We find that *both gender and tenure diversity have a significant, positive effect on productivity*, gender across all team sizes and tenure for teams larger than 10. Together, these two explain 1–2.5% of the data variance, depending on team size.

(3) Models of turnover, or team change over time, in those teams reveal a negligible effect of gender diversity. Tenure has a large negative effect on turnover, while tenure diversity has a small, positive effect of turnover.

2015 [link]

Gender bias in open source: Pull request acceptance of women versus men

Human-Computer Interaction | Social Computing | Programming Languages | Software Engineering

Josh Terrell[1], Andrew Kofink[2], Justin Middleton[2], Clarissa Rainear[2], Emerson Murphy-Hill[2], Chris Parnin[2]

February 9, 2016

Preprint-2016 [link]

Abstract

Biases against women in the workplace have been documented in a variety of studies. This paper presents the largest study to date on gender bias, where we compare acceptance rates of contributions from men versus women in an open source software community. Surprisingly, our results show that women's contributions tend to be accepted more often than men's. However, when a woman's gender is identifiable, they are rejected more often. Our results suggest that although women on GitHub may be more competent overall, bias against them exists nonetheless.

Preprint-2016 [link]

Privacy?

Developing an H-Index for OSS Developers

Andrea Capiluppi
Brunel University
London, United Kingdom
andrea.capiluppi@brunel.ac.uk

Alexander Serebrenik
MDSE, Eindhoven University of Technology
Eindhoven, The Netherlands
a.serebrenik@tue.nl

Ahmmad Youssef
ACE, University of East London
London, United Kingdom
ahmed.ali.youssef@hotmail.com

2012 [link]

Data sources: Gnome and several high-participation projects on Google Code (Go, for example)

dramatically among individuals in some cases. Therefore, it should be possible to formulate an index (or set of indexes) clearly conveying the information on individual developers, to be used as an external metric by hiring managers when recruiting new developers[1]. Existing external indicators (e.g.,

Table I
THE $h_{A,committer}$ INDEX PER AUTHOR IN THE "GO" PROJECT

Developer	Top Ranking in the "Go" project projects participated	top ten in	index
Dev1	9	9	9.9
Dev2	2	2	2.0
Dev3	2	2	2.0
Dev4	2	2	2.0
Dev5	3	3	3.0
Dev6	3	3	3.1
Dev7	3	3	3.1
Dev8	4	3	3.2
Dev9	3	2	2.1
Dev10	3	2	2.0

2012 [link]

Paid vs. Volunteer Work in Open Source

Dirk Riehle
Computer Science Department
Friedrich-Alexander University Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
dirk@riehle.org

Philipp Riemer
Computer Science Department
Friedrich-Alexander University Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
contact@philippriemer.de

Carsten Kolassa
Software Engineering
RWTH Aachen University
Ahornstr. 55, 52074 Aachen, Germany
carsten@kolassa.de

Michael Schmidt
Mathematics Department
Friedrich-Alexander University Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
michael.schmidt.nbg@gmail.com

2014, [link]

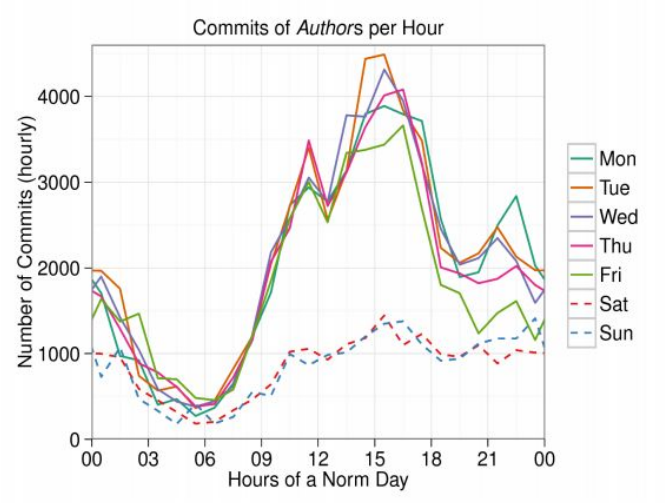


Figure 3. Number of commits by authors (when code is developed) per hour counted over all weeks 2005-2011 for the Linux Kernel

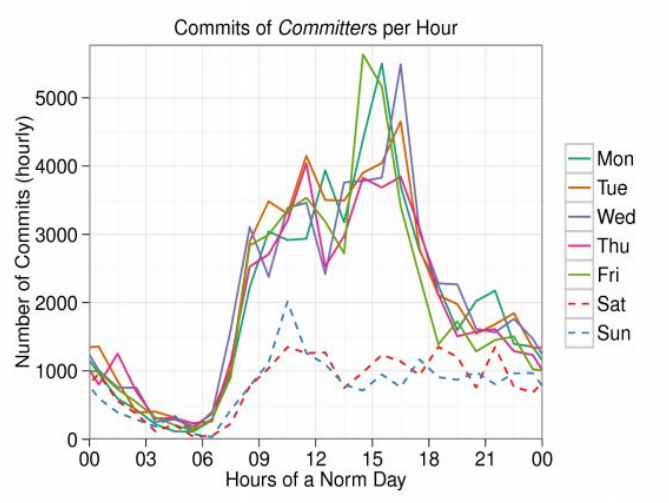


Figure 4. Number of commits by committers (when code is integrated) per hour counted over all weeks 2005-2011 for the Linux Kernel

2014, [link]

projects, a large set of more than 5,000 active open source projects, from 2000 to 2007, we find that about 50% of all contributions to projects in our sample population have been paid work. Moreover, no change in

2014, [link]

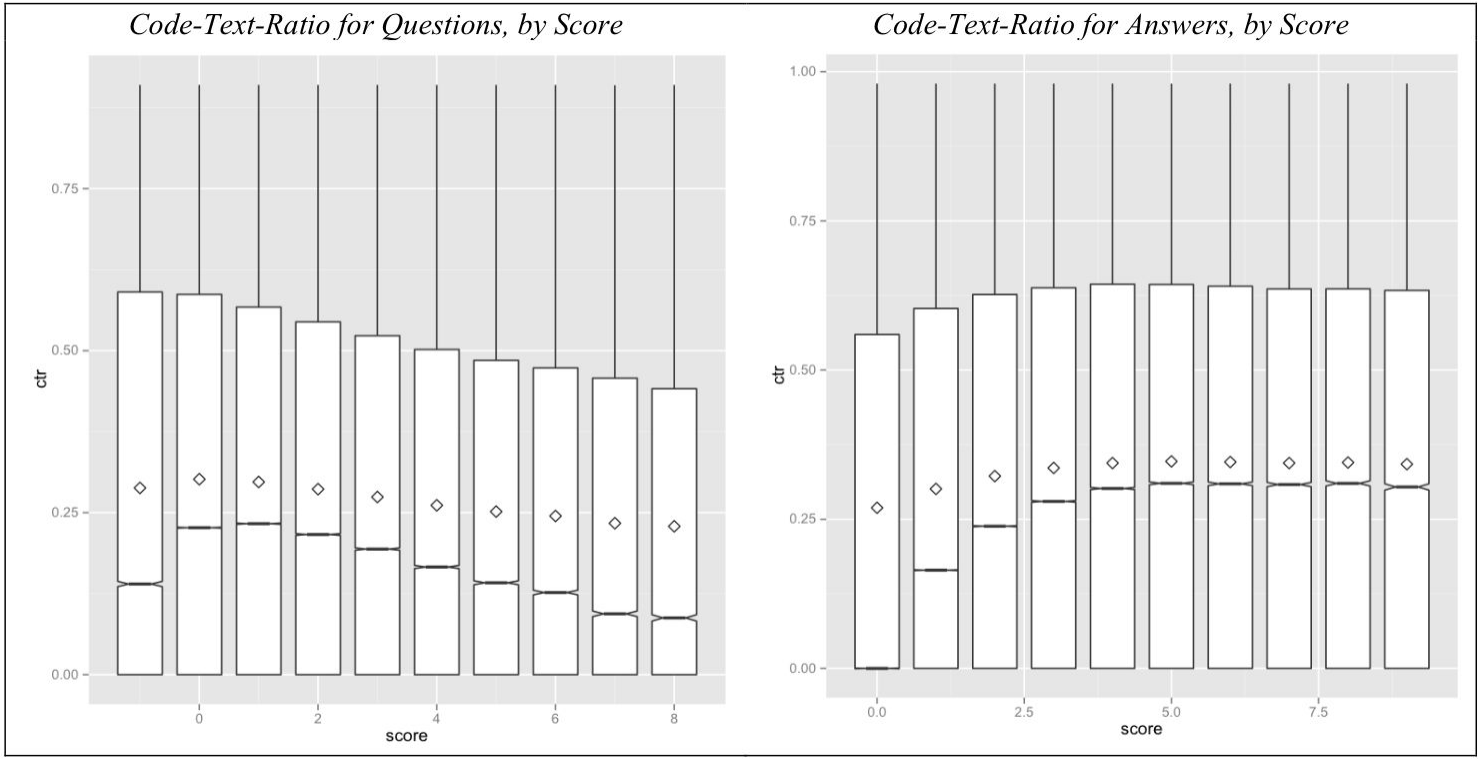
Sometimes we study FLOSS-like artifacts too.

“A bit of code”:
How the Stack Overflow Community Creates Quality Postings

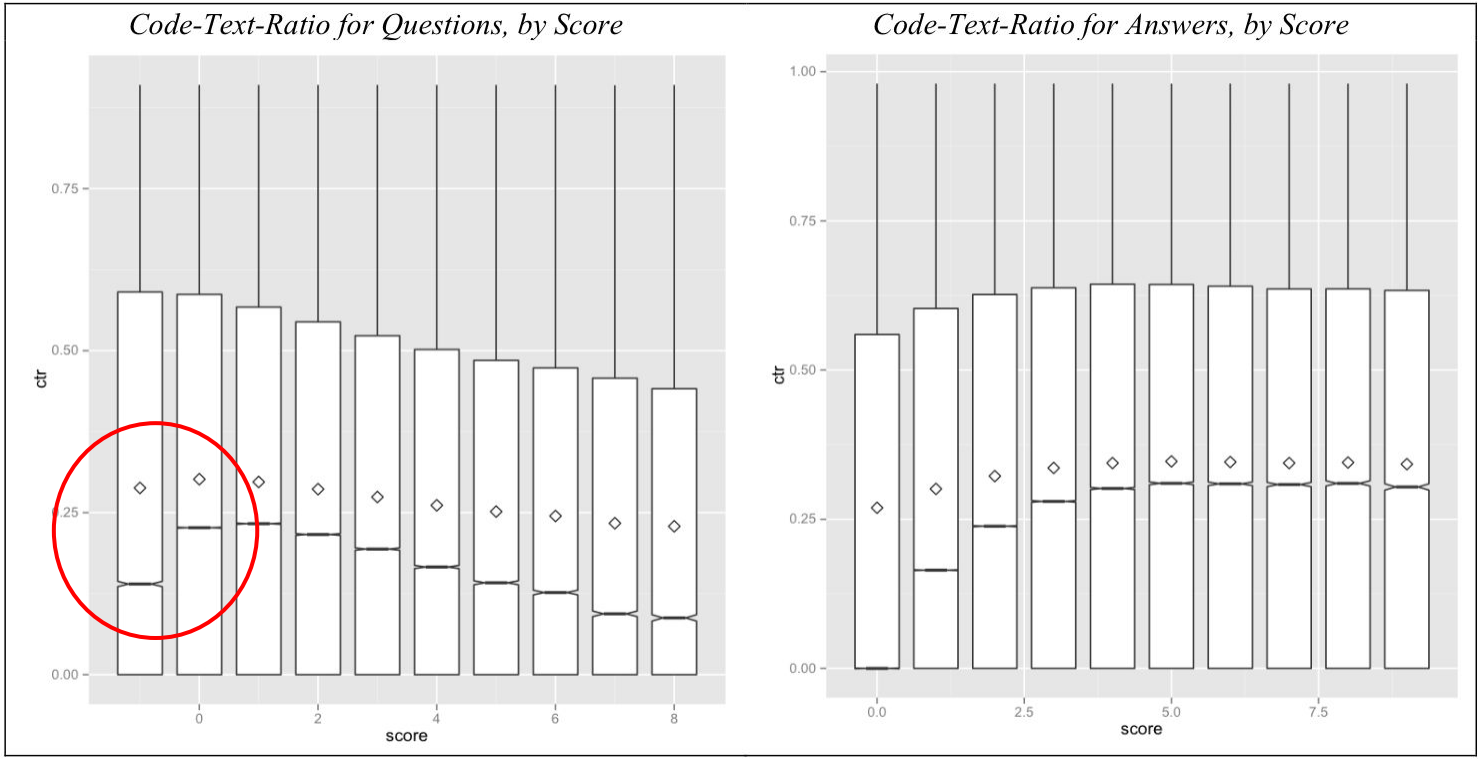
Megan Squire
Elon University
msquire@elon.edu

Christian Funkhouser
christian.funkhouser@gmail.com

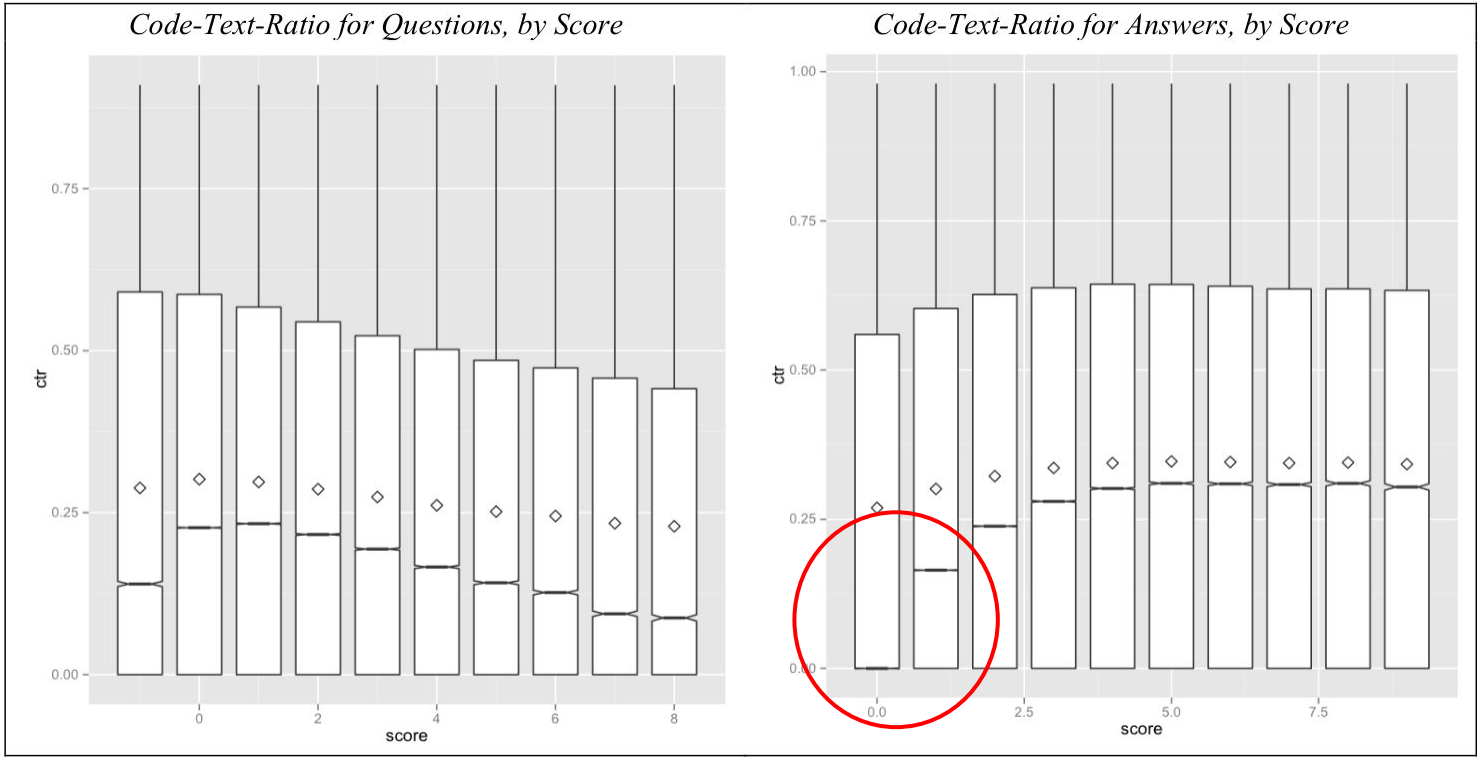
2014, [link]


Code-Text-Ratio for Questions, by Score
ctr
0.00
0.25
0.50
0.75
0
2
4
6
8
score
Code-Text-Ratio for Answers, by Score
ctr
0.00
0.25
0.50
0.75
1.00
0.0
2.5
5.0
7.5
score

2014, [link]


Code-Text-Ratio for Questions, by Score
ctr
0.75
0.50
0.25
0.00
0
2
4
6
8
score
Code-Text-Ratio for Answers, by Score
ctr
1.00
0.75
0.50
0.25
0.00
0.0
2.5
5.0
7.5
score

2014, [link]


Code-Text-Ratio for Questions, by Score
ctr
0.00
0.25
0.50
0.75
0
2
4
6
8
score
Code-Text-Ratio for Answers, by Score
ctr
0.00
0.25
0.50
0.75
1.00
0.0
2.5
5.0
7.5
score

2014, [link]

“Should we move to Stack Overflow?”

Measuring the utility of social media for developer support

2015 [link]

TABLE VI. Response Time for questions on mailing lists and forums, by project

	FS	BQ	DW	DS
Mean response time (in days)	-	7.45 days	-	28.96 days
Standard deviation	-	876 days	-	104 days
Median response time (in days and hours)	-	.75 days (18 hours)	-	.83 days (20 hours)

TABLE VII. Response Time for questions on Stack Overflow, by project

	FS	BQ	DW	DS
Mean response time (in days)	9.12 days	5.06 days	19.5 days	1.31 days
Standard Deviation	41.86 days	29.85 days	58.6 days	8.12 days
Median response time (in days and hours)	.56 days (13.44 hours)	.26 days (6.24 hours)	.63 days (15.12 hours)	.18 days (4.32 hours)

What did I miss?

What *should* we be studying?

What are we doing wrong?

More resources

- FLOSSdata flat files
- FLOSShub/biblio
- Google Scholar
- OSS Conference
- MSR Conference
- ICSE Conference
- 2013 FLOSS Survey
- What We Know and What We Do Not Know
- GH Torrent
- Stack Overflow data dumps
- MarkMail
- Gmane
- Apache Board Meeting minutes

How researchers study open source and what we've found so far

Megan Squire @MeganSquire0

Elon University & FLOSSmole / FLOSSdata / FLOSShub / FLOSSpapers

April 6, 2016