The Orbiten Free Software Survey

1st Edition, May 2000

Over 12,000 authors, 25 million lines of code analysed

The Free Software (or Open Source) "Community" is much talked about, though little hard data on this community and its activities is available. Here, for the first time, Orbiten Research (see CONTEXT) provides a body of empirical data and analysis to explain what this community actually is.

Simple facts, such as the number of developers contributing to free software projects, the number of such projects and their size have been until now unknown. The Orbiten Free Software Survey discovers these facts, and aims with them to provide a foundation for empirical research on the free software community.

Building on the release of CODD over a year ago, the Survey will measure and track over time several aspects of the free software economy including: the concentration (or diversity) of contributions and contributors; the degree of intersection between projects and sharing of code; the participation of developers in different projects; volatility of changes to the code base and the developer base.

There will also be some basic statistics and data gained during the survey process - such as total size of free software available, amount of free software being released and/or modified each month, compendium of developers.

Hopefully the survey will be regular, prompt and gradually more comprehensive, providing an important source of information for academic researchers, free software users and developers alike.

Rishab Aiyer Ghosh & Vipul Ved Prakash
May 7, 2000

FINDINGS

The primary findings of OFSS01 were basic: the number of developers authoring projects included in the survey (12706), the size of the free software code base (1.04 Gigabytes, or roughly 25 mil lines), the number of identifiable free software projects (3149). Given the total lack of data on the free software economy, rough indicators as to its size (limited by the initial scope of the survey) are, we believe, a good start.

Secondary findings relate to the degree of contribution to the code base by

individual authors, defined for the purposes of this survey as the smallest identifiable grouping claiming credit for development of a software project. Unsurprisingly, the Free Software Foundation came out well ahead of anyone else by far, credited with 11% (124 Mb) of the entire surveyed code base and involved in 17% (546) of all identifiable projects. However, as with some other well-known (and highly ranked in the survey) Unix authors, such as Sun Microsystems and the Regents of the University of California, the FSF's position in our charts stems largely from the lack of credit given to individual programmers. A list of the top few contributors sorted by code and involvement in projects is given below (see DATA).

Further findings relate to the distribution of authors among projects, and code base contribution. The top 1271 authors, 10% of the total, accounted for 72.3% of the total code base. The top 10 authors alone (0.08% of the total) are credited for 19.8% of the code base. Free software development may be distributed, but it is most certainly very top heavy.

What goes for lines of code written goes for involvement in projects too. Only the top 25 authors (0.19% of the total) were credited with participation in more than 25 projects. The top 250 authors were credited with participation in over 5 projects, and the vast majority (over 77%) of authors were only involved in a single project. Our conclusion: Free software development is less a bazaar of several developers involved in several projects, more a collation of projects developed single-mindedly by a large number of authors.

DATA

Number of identifiable authors	12706
Uncredited/unidentifiable authors	790
% of code base uncredited	8.37%
Size of code base	+1116500467 Bytes or 1067 Mb.
Number of identifiable projects	3149

Table 1: Top 10 authors ranked by contribution of code

Author	% of total
free software foundation, inc	11.231
sun microsystems, inc	1.848
the regents of the university of california	1.359
gordon matzigkeit	1.216
paul houle	1.042
thomas g. lane	0.782
the massachusetts institute of technology	0.762

ulrich drepper	0.559
lyle johnson	0.528
peter miller	0.525
more...	

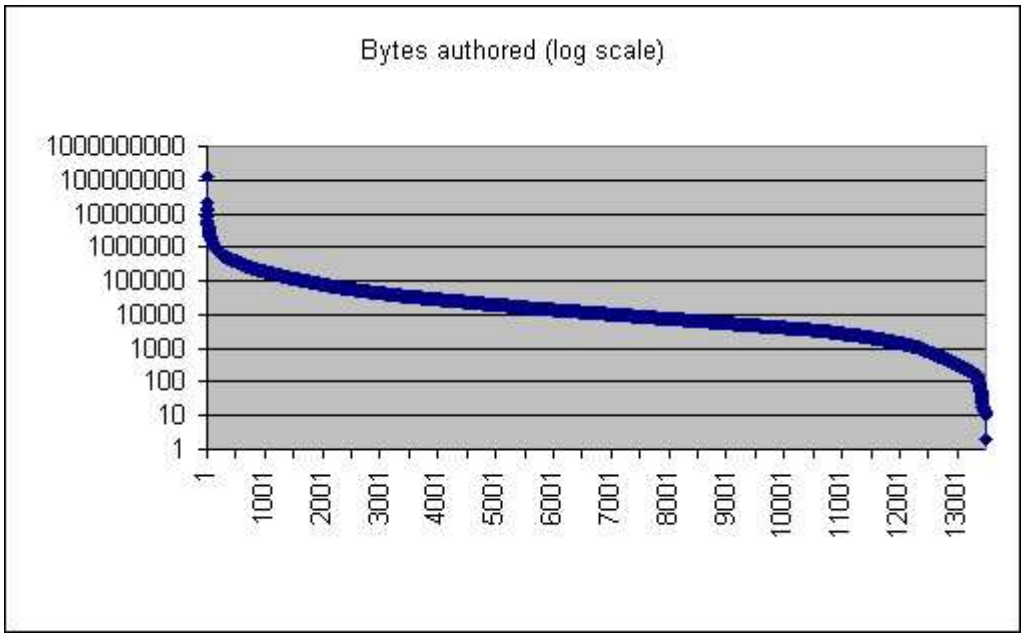
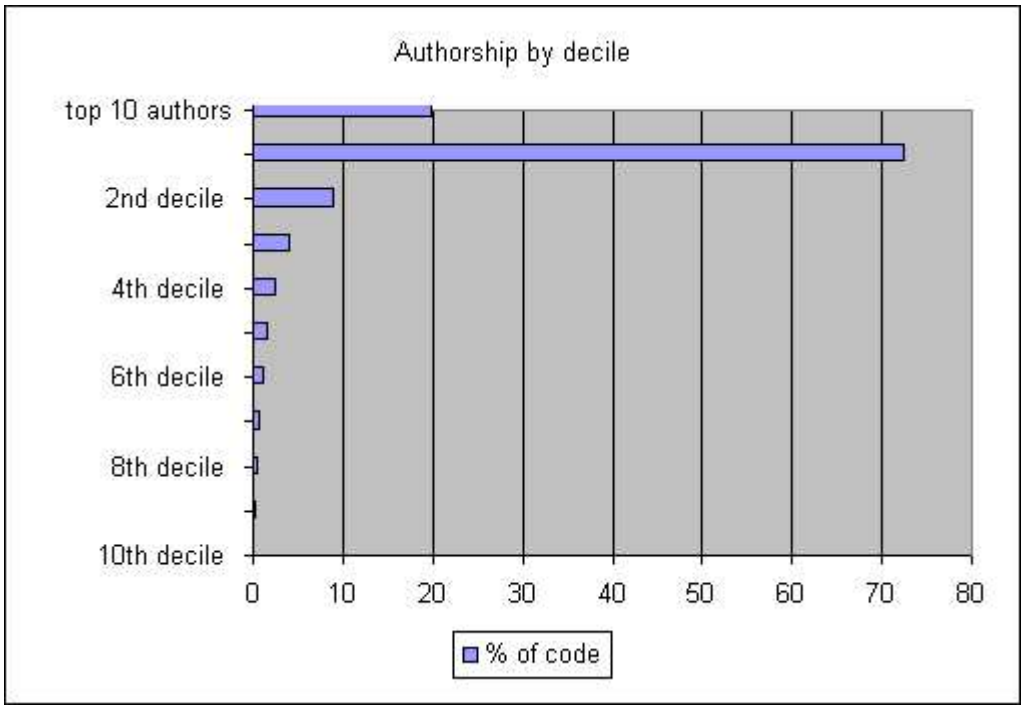


Table 2: Author contribution by decile

Authors	% of total
top 10 authors	19.854
top decile (1271)	72.320
2nd decile	8.928
3rd decile	4.062
4th decile	2.384
5th decile	1.515
6th decile	1.008
7th decile	0.672
8th decile	0.440
9th decile	0.239
10th decile	0.060



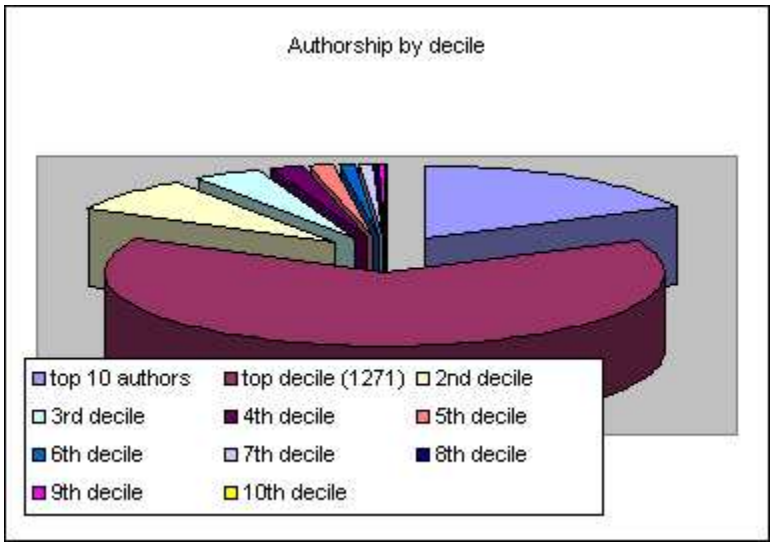


Table 3: Top 10 authors ranked by participation in projects

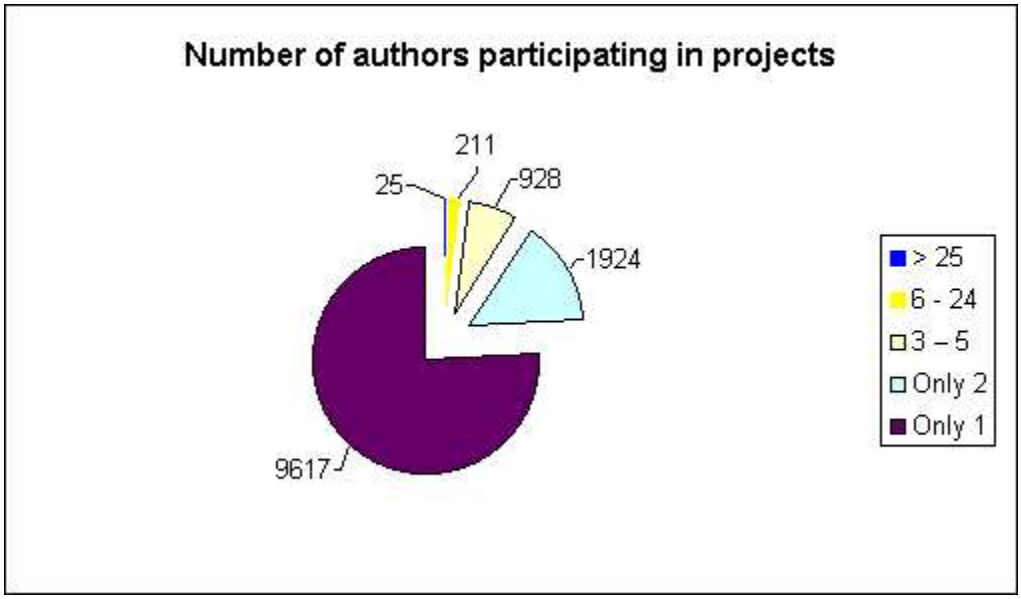
Author	Projects
free software foundation, inc	546
gordon matzigkeit	267
the regents of the university of california	156
ulrich drepper	142
roland mcgrath	99
sun microsystems, inc	66
rsa data security, inc	59

martijn pieterse	50
eric young	48
login-vern	47
more...	

Table 4: Author participation in projects

Projects	Authors
> 25	25
6 - 24	211
3 - 5	928
Only 2	1924
Only 1	9617

Note: 211 authors participated in 6 to 24 projects, etc.



SCOPE AND METHOD

The first Orbiten Free Software Survey has been prepared based on over 18 months of work in identifying, tracking and modeling interaction in the free software economy. Clearly this was not enough time, and the scope and methodology of the first survey is far from ideal.

The technical task of identifying credits in poorly documented source code was complex, especially given the vast and changing nature of the code base. Credits are often not available, they rarely follow a set format, and various heuristics have been applied and "policy" decisions made on, for example, how to divide credit among multiple listed authors. Details can be found in the documentation for CODD.

The code base itself was limited. Although far from being a complete set of all code ever released without payment on the Internet - our ideal, eventual goal - we believe we have used a fairly representative sample of software projects (released under the GNU Public Licence and its variants) developed in recent years.

The source code base for OFSS01 is:

- RedHat Linux v6.1 source rpms. [http://www.redhat.com]
- Linux kernel sources version 2.2.14.
- Munitions cryptography/security archive as on January 11, 2000 [http://munitions.vipul.net]
- Approximately 50% of source code available through Freshmeat as on January 5, 2000. Explanation: source code is not easily available for all projects on Freshmeat, at least when accessed through an automated script with simple intelligence. [http://freshmeat.net]

For each module or package analysed, source code is broken into projects identified according to the package distribution. Source code and some documentation files are scanned for authorship, credit or copyright information, from which author names are identified. Data collected includes, for each identified author, number of bytes of code authored, number and names of projects authored. From this the degree of contribution, in terms of bytes of code can be calculated for any given project. Project data is collated to form a broader picture of authorship distribution, which can be examined at several levels.

In this survey, very basic analysis has been performed. The next survey will broaden the scope of analysis to include features such as the degree of cross-participation between projects and groups of authors.

The next survey - planned for June - will also use a bigger code base. At the very least the code base will expand to include Sourceforge [http://sourceforge.net], OpenBSD [http://openbsd.org] and Perl CPAN libraries [http://cpan.org].

As the survey continues and becomes more frequent, we plan to track changes in the code base over time (including historical perspectives using older versions of, say, the Linux kernel) and monitor movement between projects and groups.

REFERENCES

1. CODD documentation, Orbiten.
2. "Cooking-pot markets" by Rishab Aiyer Ghosh, First Monday, Issue 3 Volume 3 March 1998.
3. "Identifying, tracking and measuring activity in cooking-pot networks" by Rishab Aiyer Ghosh, Orbiten.