

Sampling from the Debian GNU/Linux Distribution: Software Reuse in Open Source Software Development

HICSS 2007, Hawaii

Authors: Sebastian Spaeth, Matthias Stuermer, Stefan Haefliger, Georg von Krogh



Research Project on Software Reuse

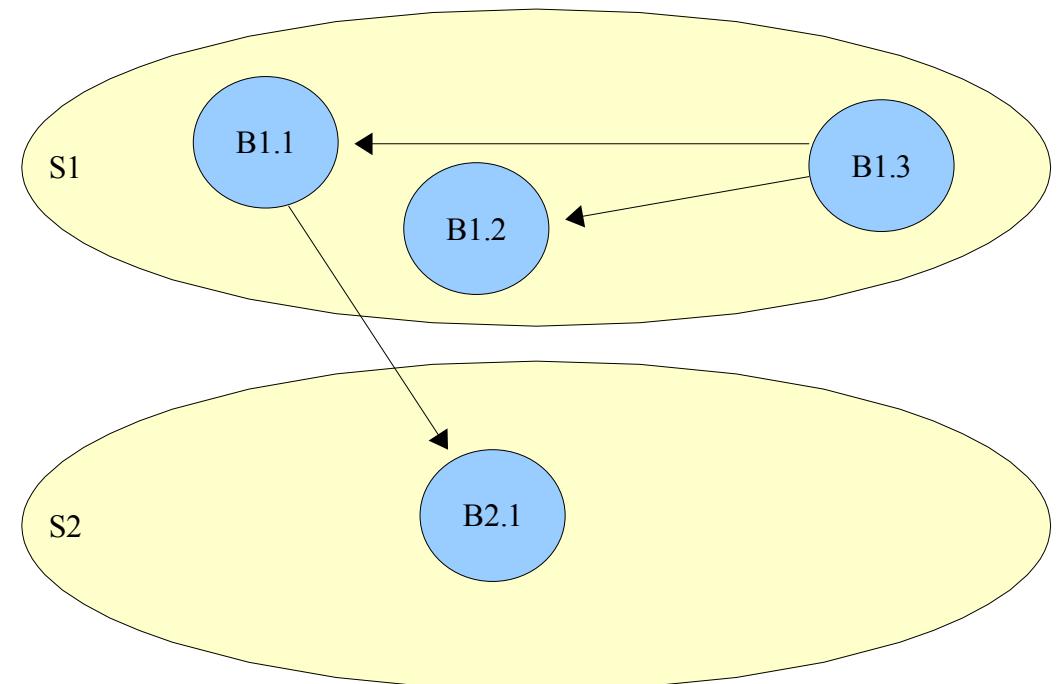
- **Motivation:**
 - Software reuse lowers development costs.
 - Software reuse is difficult to achieve.
 - Software reuse is abundant in OSS development.
- **Research question:** What are the characteristics of software components that are reused more often than others?
- **Sampling:** What are the advantages of sampling from Debian GNU/Linux?

Facts about Debian GNU/Linux

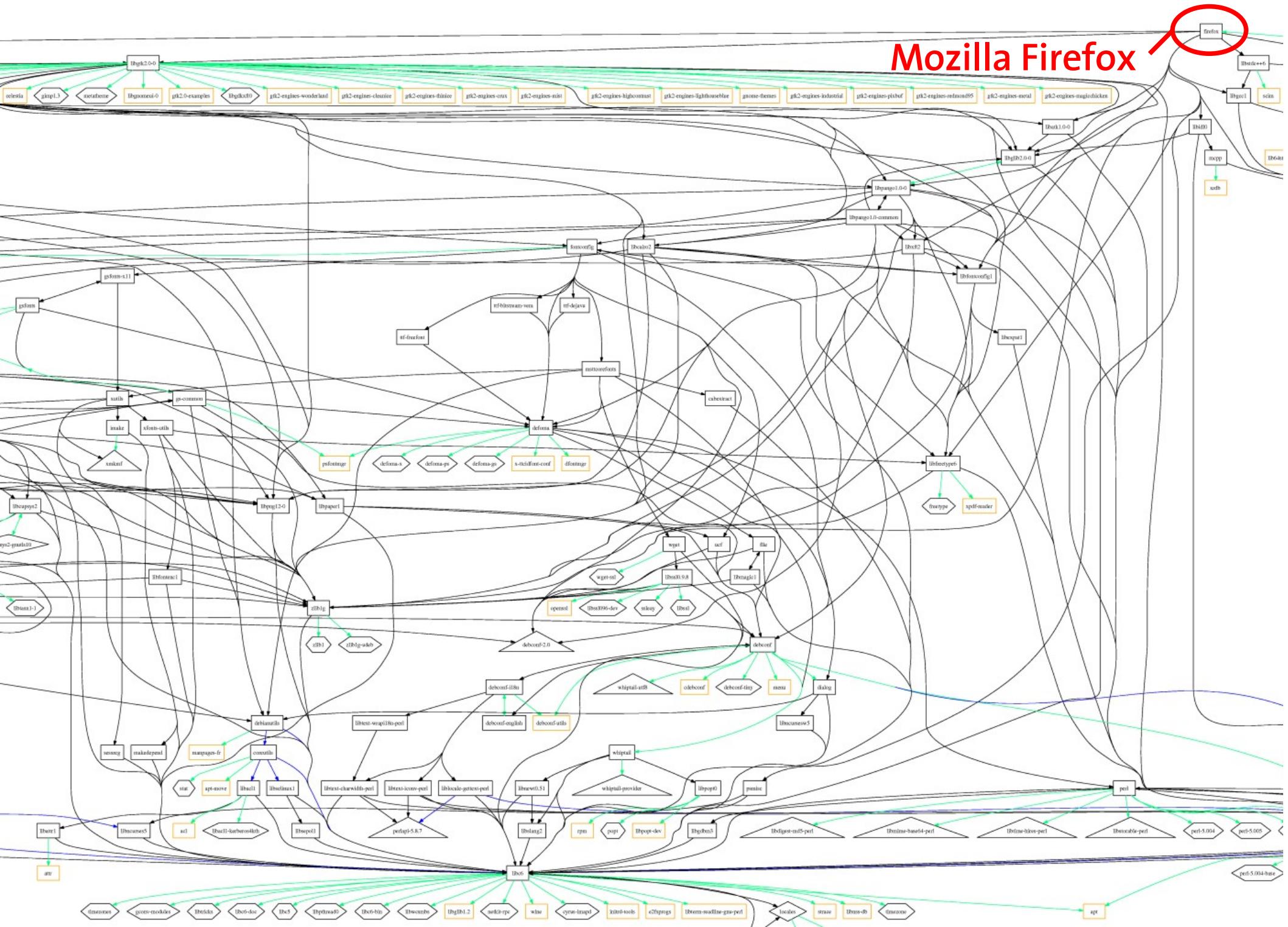
- Founded 1993 by Ian Murdock (his wife Debra -> Debian)
- Community-controlled Linux distribution
- ≈20,000 ready compiled software packages
- Categorized in sections such as mail, text or libs
- Information on packages: name, version, maintainer, license...
- Dependency information

The Debian package system: Binaries, sources and dependencies

- Binary packages (B) are compiled from source packages (S)
- Dependencies are among binary packages



Mozilla Firefox



Advantages of sampling Debian vs. SourceForge

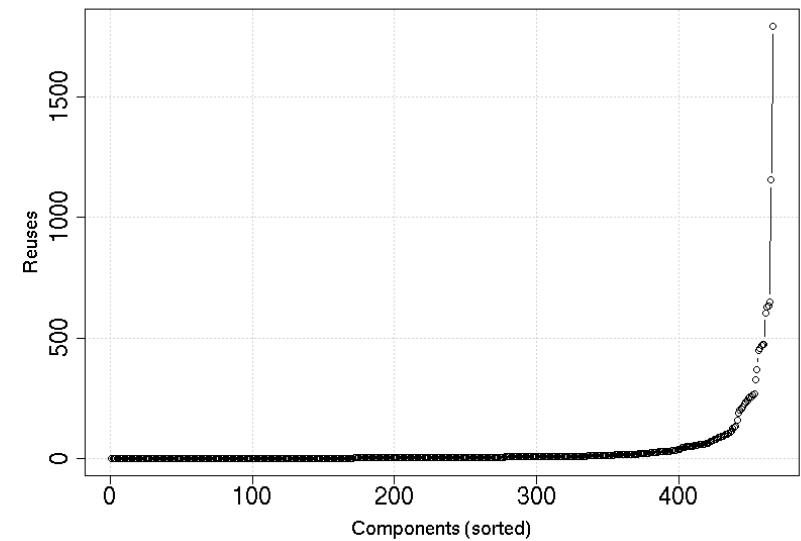
- 1. SourceForge.net excludes systematically OSS projects.**
 - other collaboration platforms (Tigris etc.)
 - individually hosted projects (Mozilla, Apache, GNU, GNOME)
- 2. Debian maintainers doing peer review**
- 3. Software actually in use**
- 4. Packages form an integrated environment**

Limitations Using Debian for Sampling

- exclusion of non-Linux software (e.g. OSS for Windows)
- not appropriate to measure project failure
- license restrictions for Debian-acceptable software

Sample and Method

- Total of 19,692 binary packages (for 32bit computers)
- Total of 8,890 source packages (for all platforms)
- 1,146 source packages contain packages marked as “libs” or “oldlibs” (categorized by Debian developers as reusable library component)
- Our sample: random sub-sample of 466 components
- Total reuse of 16,949 average reuse 36
- Log-linear regression
- Deductive research
- Reuse as dependent variable



Results: Coefficients of log-linear regression

- Multiple R²: 0.256
- * 10% ** 5% *** 1% significance level
- $\ln(Y) = \ln(\alpha) + \ln(\beta)X$

Variable	coef(std.err)
(Intercept)	-0.33(-0.3)
website	0.06(0.1)
doc	0.38(0.4)*
freshmeat	0.73(0.7)***
umbrella	0.47(0.5)***
legal_entity	0.25(0.3)*
C_prog	0.50(0.5)***
standard	-0.18(-0.2)
strict_lic	-0.50(-0.5)***
bugs	0.01(0.0)***
binmodules	0.08(0.1)***
binsizeMB	-0.04(-0.0)*
age	0.13(0.1)***

Descriptive Statistics

	Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11
1.	website	0,86	0,34											
2.	doc	0,58	0,4	0,53										
3.	freshmeat	0,54	0,5	0,35	0,37									
4.	umbrella	0,35	0,48	-0,07	-0,09	-0,1								
5.	legal_entity	0,37	0,48	-0,05	0,06	-0,07	0,32							
6.	C_prog	0,71	0,46	0	-0,1	0,06	-0,05	-0,07						
7.	standard	0,4	0,49	0,17	0,15	0,15	-0,14	0,03	0,14					
8.	strict_lic	0,28	0,45	-0,07	-0,1	0,02	0,11	-0,05	0	-0,07				
9.	bugs	9,43	37,63	0,05	0,06	0,11	0,06	0,15	-0,03	0,08	0,16			
10.	binmodules	4,1	4,62	0,08	0,14	0	0,03	0,1	-0,1	-0,05	0,09	0,24		
11.	binsizeMB	1,71	4,28	0,05	0,16	0,02	-0,03	0,14	-0,17	-0,03	0,06	0,32	0,66	
12.	age	5,83	2,69	0,05	0,1	0,12	-0,18	0,06	0,16	0,05	-0,08	0,16	0,2	0,21

n=466

Results: Hypotheses and Control Variables

#	Hypotheses/Control Variable	Result	Effect on Reuse
H1	A dedicated web page of a component has a positive effect on reuse.	∅	Not significant
H2	Published documentation of a component has a positive effect on reuse.	✓	+ 46% *
H3	The listing of a component on a platform (Freshmeat) has a positive effect on reuse.	✓	+ 107% ***
H4	The existence of an umbrella project for a component has a positive effect on reuse.	✓	+ 60% ***
H5	The existence of a legal entity for a component has a positive effect on reuse.	✓	+ 29% *
H6	The use of C as a programming language has a positive effect on reuse.	✓	+ 65% ***
H7	The implementation of a standard has a positive effect on reuse.	∅	Not significant
H8	The use of a restrictive license (GPL) has a negative effect on reuse.	✓	- 39% ***
C	Number of bugs in the Debian bug database for this component		+ 0.5% per bug ***
C	Number of binary packages per source package		+ 7.9% per package *
C	Size of binary packages of a source package		- 3.7% per MB ***
C	Age of a component		+ 13.7% per year ***

n=466 r²=0.256

Discussion

- Questions?
- Contact and weblog: www.smi.ethz.ch