

An Exploratory Analysis of SourceForge.net Project Statistics

Nate Oostendorp, Mouly Kumaraswamy, Tom Hayden, Zach
Lai

April 9, 2008

About Sourceforge.net

Sourceforge is the leading open source project hosting website. The site provides a web space and tools like code repository, bug trackers, etc for managing software projects.

- ▶ 173,886 Projects
- ▶ 1,824,476 Users

Our data was taken from the site's January 2008 database dump.

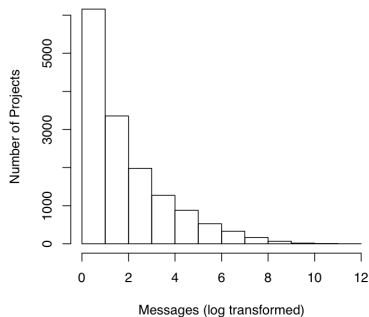
Description of Data

- ▶ **Projects table** - Details of individual projects. Each row represented a project hosted by Sourceforge.net.
- ▶ **Data items table** - Captures tool activity for projects. The tools we considered are: forums, artifact, artifact message, screen shot, frs file, news and task
- ▶ **Monthly download counts** - Captures the download sums for each project. We decided to use downloads as a metric for popularity for a project.

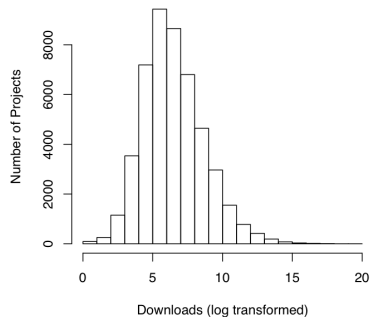
Data Characteristics

We found the downloads have an approximately log-normal distribution, while most participation functions have exponential "long tail" distributions.

Histogram of Forum Posts per Project

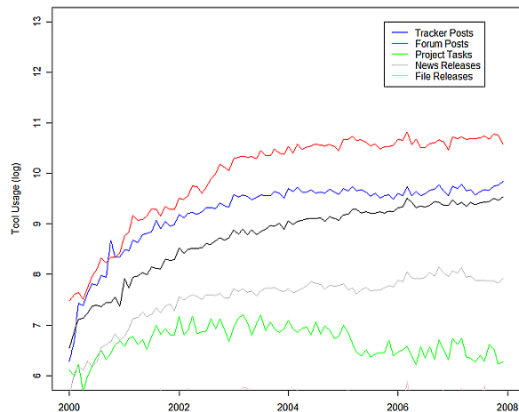


Histogram of Total Downloads per Project



Data Characteristics

We also looked at the use of the various participation statistics in a time-series to see if use over time had changed substantially.



Hypotheses

We were interested in determining if any of the participation metrics were predictors of project popularity. To test this we came up with the following hypotheses:

- ▶ H1. Downloads will increase as project tool use increases.
- ▶ H2. Downloads will increase with more project members.
- ▶ H3. Downloads will increase with more users participating.
- ▶ H4. Downloads will decrease with More restrictive licenses.
- ▶ H5. Total Monthly Downloads on the site are independent of the month observed.

H1. Tool Usage v. Downloads

Testing

- ▶ We selected a random sample of 700 projects binned by log of total downloads.
- ▶ We ran multiple regression on this sample to find the correlation between the downloads and the tool usage. From this, ran correlation between downloads and tool usage

Results

- ▶ This regression gave us an R-squared value of 0.32, which is a significant impact.
- ▶ The results of the factors of Bug Tracker, Forum Posts, and File Releases were significant, proving H1 that Tool use and Downloads are correlated

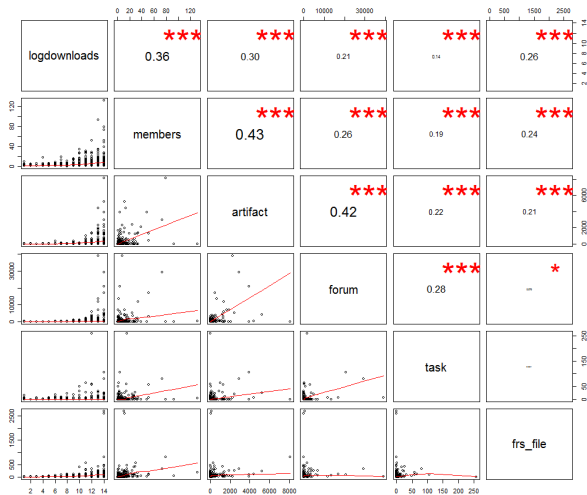
H1. Tool Usage v. Downloads

```
> summary(lm(downloads ~ artifact + forum +  
  artifact_message + frs_file + document))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	132137.305	17168.136	7.697	4.82e-14	***
artifact	561.212	80.877	6.939	9.07e-12	***
forum	16.505	8.035	2.054	0.0403	*
artifact_message	-33.086	21.931	-1.509	0.1318	
frs_file	1129.885	109.419	10.326	< 2e-16	***
document	-40913.945	6338.624	-6.455	2.04e-10	***

H1. Tool Usage v. Downloads



H2. Project Membership v. Downloads

Testing

- ▶ We looked at the relationship between project membership versus number of downloads
- ▶ Used regression of project membership against downloads on log scale

Results

- ▶ Each additional member seems to be an indicator of increased project consumption (Coefficient=0.120472 & p-value < 2.2e-16)
- ▶ R-squared value of 0.064
- ▶ Accept H2, however only slightly

H3. User Participation v. Downloads

Testing

- ▶ We looked to see if the number of users in a project over it's lifetime v. total downloads
- ▶ Used linear regression
- ▶ Participating User = Unique member and non-member participants

Results

- ▶ **Significant relationship** between users participating and downloads (Coefficient=0.0069478 & p-value < 2.2e-16)
- ▶ R-squared value of 0.065
- ▶ Accept H3, however only slightly

H4. Software Licensing and Downloads

Testing

- ▶ We looked to see if the software license has a statistically significant on downloads
- ▶ Used anova

Results

- ▶ **No significant effect** - F-value=0.0406 & p-value=1

Response: downloads

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
license	65	1.2173e+13	1.8728e+11	0.0486	1
Residuals	47734	1.8376e+17	3.8497e+12		

H5. Downloads v. Month

Testing

- ▶ We noticed a considerable dip in December and January downloads
- ▶ Therefore, we looked to see if the total number of downloads from all projects is not independent of the month
- ▶ Compared monthly download count per month v. total downloads
- ▶ Used chi test

Results

- ▶ **Reject H5:** $p\text{-value} < 2.2e-16$

Fin

Thank You!