

An Exploratory Analysis of SourceForge.net Project Statistics

Tom Hayden, Nathan Oostendorp, Zongyun Lai, Mouly Kumaraswamy

April 8, 2008

1 Description of the Data

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. The site has 173,883 projects and 1,824,476 users as of April 5th. We took our data from the site's January 2008 database dump. We exported the database into three csv files - projects, data items and monthly download sum.

- The Projects table has the details of individual projects. The columns in this file are group id, group name, license (e.g. GPL, MIT), members. Each row represents a project hosted in Sourceforge.
- The data items table captures tool activity for projects. Its columns are group id, user id, tool type, time stamp. Group id points to the project in which the activity took place and user id points to the user who did the activity. The tools present in the site are forums, artifact, artifact message, screen shot, frs file, news and task.
- Next we took the monthly download count for all projects. This table had the columns - group id, sum, time stamp. The sum column has the total number of downloads from the project (group id) in a month (time stamp)

We decided to concentrate on the total number of downloads each project received as our primary statistic. We think this is a good metric of the consumer-side demand for software and useful as a measure of the popularity of a project.

2 Data Characteristics

We found several interesting properties of the data, the most striking of which is the different distributions of data between our primary metric of project consumption (downloads served) and our many variables based on participation (bugs logged, messages posted, etc). We found the downloads have an approximately log-normal distribution, while most participation functions have exponential “long tail” distributions. Figure A and B show histograms of the downloads by project and the number of posts in project forums. Most other participation variables show similar distributions to Figure B.

We also looked at the use of the various participation statistics in a time-series to see if use over time had changed substantially. In Figure 2 we see monthly tool use trends across projects.

It is interesting to note that the two most popular tools (Discussion Forums and Bug Trackers) can be created by both project members, and project non-members while the other tools (File Releases, Project Tasks, and News Releases) can only be contributed by project members. When

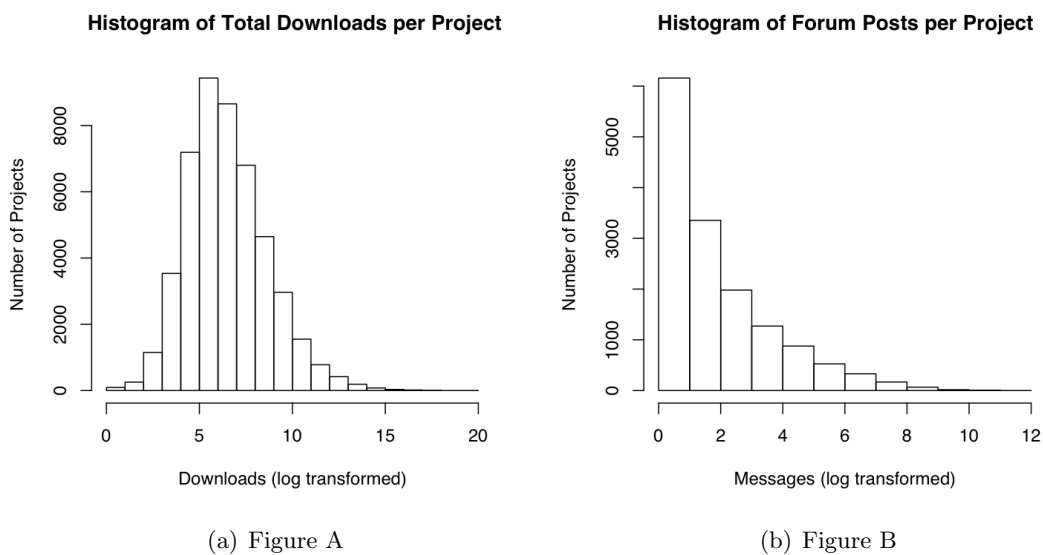


Figure 1: Histogram of Downloads

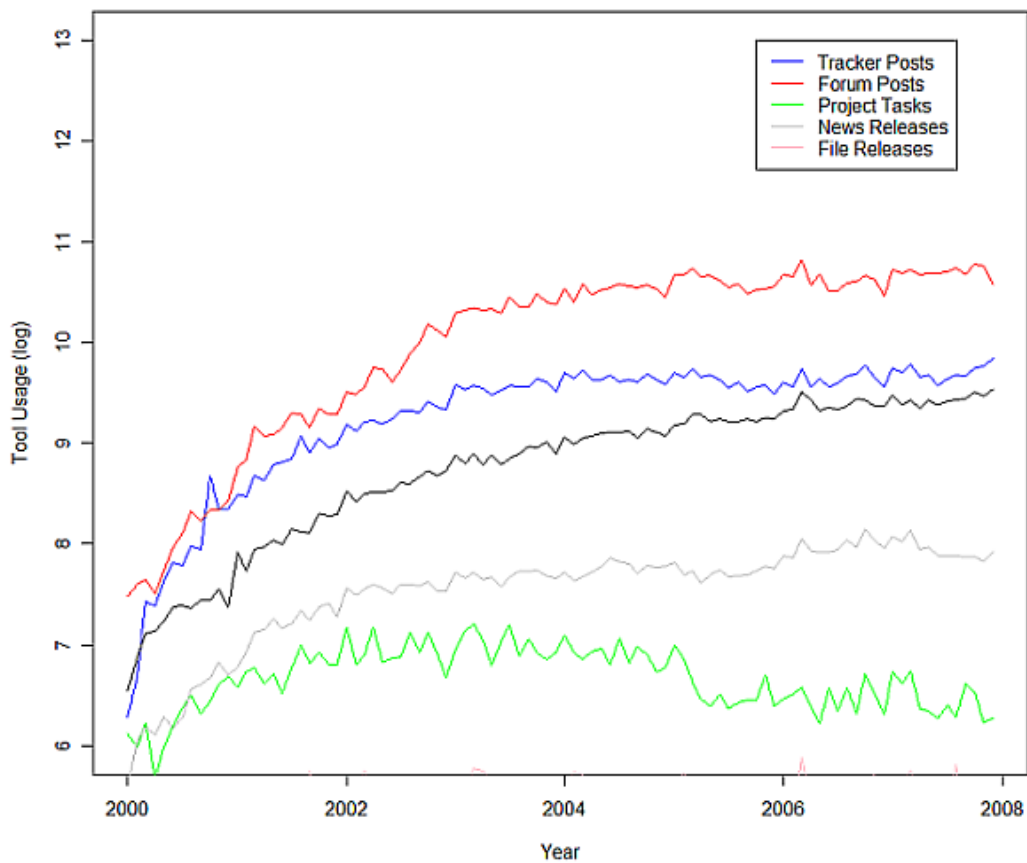


Figure 2: Monthly Tool Use Trends

fit to linear models, all tools show growth in usage except for Project Tasks, in which use declined over time.

3 Hypotheses

While the metrics of participation and metrics of consumption have different distribution, 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.

4 Data Analysis

4.1 H1. Tool Usage v. Downloads

To find the if the number of downloads for a project was dependent on tool usage in the project, we created a stratified sample of the projects. We selected a random sample of 700 projects binned by log of total downloads. Next we ran multiple regression on this sample to find the correlation between the downloads and the tool usage. 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.

4.2 H2. Project Membership v. Downloads

We found a significant relationship between project membership and downloads. For this we ran a regression of project membership against downloads on a logarithmic scale. In this relationship each additional member seems to be an indicator of increased project consumption, but with an R-squared value 0.064 the impact is very slight. This leads us to accept H2, that project membership and downloads are related, however only slightly.

4.3 H3. User Participation v. Downloads

We looked at the number of users who participated in a project over its lifetime versus the total number of downloads in a linear regression model. Participating users were counted by taking the unique user id count across all tool use on the project, and included both member and non-member participations. We found a significant relationship between the number of users participating in a project and the number of downloads. The strength of this relationship is very close to the test for project membership, with an R-squared value 0.065. This significant relationship leads us to accept H3, but again with the reservation that the effect is slight.

4.4 H4. Software Licensing and Downloads

We found no significant relationship between license and number of downloads, when comparing download distributions against project license type. Using an anova, we found an F value of 0.0406 which resulted in a P value very close to 1. This leads us to reject H4.

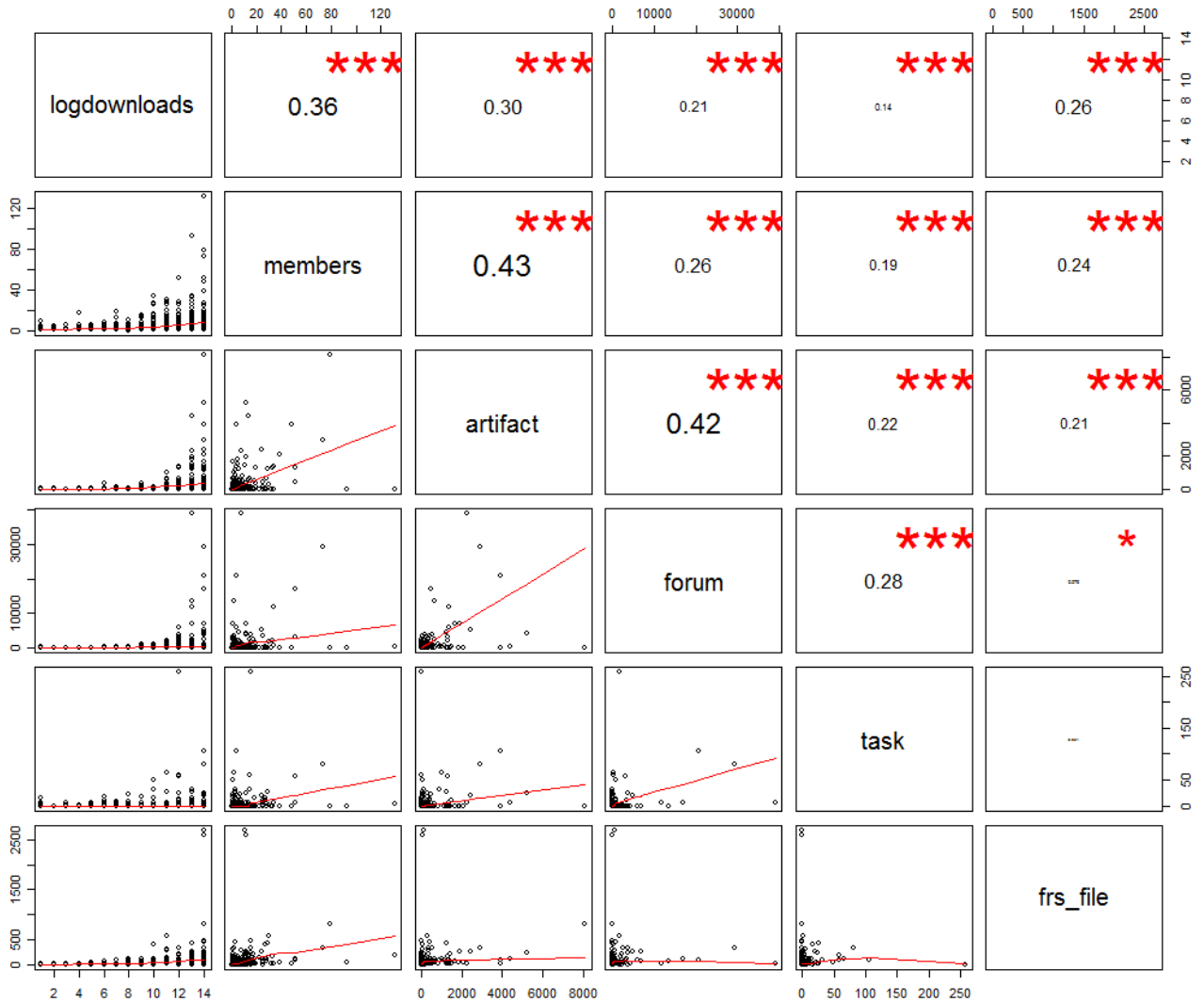


Figure 3: Visualization of a Correlation Matrix. On top the (absolute) value of the correlation plus the result of the cor.test as stars. On bottom, the bivariate scatterplots, with a fitted line

4.5 H5 Downloads v. Month

To test the hypothesis if the total number of the downloads from all projects in Sourceforge is independent of the month. We tabulated monthly down count (for 2007) as month vs total downloads matrix. The result of the chi test on the tabulated data was sufficiently ($p\text{-value} < 2.2e-16$) low for us to reject the null hypothesis and accept H5. The plot of number of downloads against time of the year shows that there is reduction in the number of downloads for the month of December.

5 Discussion of the Results

From the regression results, we found that the tool usage correlated with the number of downloads. So projects that make better of the Sourceforge tools are likely to have more downloads. Amongst the tools, artifact and file releases have high positive correlation with the downloads. While usage of document has a negative correlation with downloads. We think this is probably due to the fact that only old projects use the documents tool. Usage of forums was correlated only at 5% level,

which not a strong evidence for a large dataset like ours.

User Participation seemed to be linked to total number of downloads, however only very slightly. This indicates that membership and number of participating users are not a very important factor in how popular a project is. We think think this may indicate that whether a project is maintained by a large group or a small core of individuals may be slightly reflected in project consumption, so users of open source may have a weak preference for collaborative versus individual work.

6 Appendix

6.1 H1. Tool Usage v. Downloads

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

Call:
lm(formula = downloads ~ artifact + forum + artifact_message + frs_file + document)

Residuals:

Min	1Q	Median	3Q	Max
-1016200	-136316	-133260	-86012	2824752

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	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 434500 on 693 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-Squared: 0.3198, Adjusted R-squared: 0.3149
F-statistic: 65.16 on 5 and 693 DF, p-value: < 2.2e-16

6.2 H2. Project Membership v. Downloads

```
> summary(lm(log(downloads) ~ members))
```

Call:
lm(formula = log(downloads) ~ members)

Residuals:

Min	1Q	Median	3Q	Max
-37.8365	-1.4590	-0.2030	1.3028	12.9533

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.213710	0.010712	580.05	<2e-16	***
members	0.120472	0.002109	57.11	<2e-16	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.1 on 47798 degrees of freedom
Multiple R-Squared: 0.06388, Adjusted R-squared: 0.06386
F-statistic: 3262 on 1 and 47798 DF, p-value: < 2.2e-16

6.3 H3. User Participation v. Downloads

```
> summary(lm(log(downloads) ~ users))
```

Call:

```
lm(formula = log(downloads) ~ users)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.6025	-1.4649	-0.1994	1.3326	13.3277

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.4138833	0.0096764	662.84	<2e-16 ***
users	0.0069478	0.0001202	57.78	<2e-16 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.099 on 47798 degrees of freedom
Multiple R-Squared: 0.06528, Adjusted R-squared: 0.06527
F-statistic: 3338 on 1 and 47798 DF, p-value: < 2.2e-16

6.4 H4. Software Licensing and Downloads

```
> summary(lm(downloads ~ license))
```

Call:

```
lm(formula = downloads ~ license)
```

Residuals:

Min	1Q	Median	3Q	Max
-151955	-43361	-41904	-23932	385210430

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	338.50	1387383.03	2.44e-04	1.000
licenseadaptive	4838.15	1466725.08	0.003	0.997
licenseafl	2812.79	1392608.60	0.002	0.998
licenseapache	11790.69	1389858.29	0.008	0.993
licenseapache2	3836.13	1389586.98	0.003	0.998
licenseapssl	12702.71	1436077.87	0.009	0.993
licenseartistic	19609.44	1389693.41	0.014	0.989
licenseattribut	1007.75	1498544.95	0.001	0.999

licensebsd	12986.98	1387838.73	0.009	0.993
licensecatosl	-70.50	2403017.90	-2.93e-05	1.000
licensecddl	1380.21	1405285.49	0.001	0.999
licensecpal	279.50	1962055.90	1.42e-04	1.000
licensecvw	-37.00	1699190.25	-2.18e-05	1.000
licensedatagrid	-216.50	2403017.90	-9.01e-05	1.000
licenseeclipselicense	4222.82	1396480.72	0.003	0.998
licenseeducom	1025.69	1419281.06	0.001	0.999
licenseeiffel	-145.50	1962055.90	-7.42e-05	1.000
licenseeiffel2	238.28	1533809.65	1.55e-04	1.000
licenseenduser_advanced	123.50	2403017.90	5.14e-05	1.000
licenseentessa	129.50	2403017.90	5.39e-05	1.000
licensefair	1546.03	1466725.08	0.001	0.999
licenseframe	-221.50	2403017.90	-9.22e-05	1.000
licenseframeworkx	-43.07	1573144.49	-2.74e-05	1.000
licensegpl	43253.46	1387425.75	0.031	0.975
licensehistorical	-12.17	1602011.93	-7.59e-06	1.000
licenseibm	63350.90	1417878.61	0.045	0.964
licenseibmcpl	8604.86	1391000.73	0.006	0.995
licenseiosl	1537.46	1446446.80	0.001	0.999
licensejabber	7319.00	1483174.85	0.005	0.996
licensejgpl	34243.58	1387678.88	0.025	0.980
licensejit	19742.75	1389010.46	0.014	0.989
licensejpl	46124.49	1394337.38	0.033	0.974
licensejpl11	19578.86	1390590.86	0.014	0.989
licensejasalicense	11.00	1962055.90	5.61e-06	1.000
licensejauisite	-281.50	2403017.90	-1.17e-04	1.000
licensejncsa	5553.54	1439755.41	0.004	0.997
licensejnethack	27657.37	1471541.92	0.019	0.985
licensejenokia	3040.00	1602011.93	0.002	0.998
licensejenone	25622.00	1483174.85	0.017	0.986
licensejoclc	1084.50	2403017.90	4.51e-04	1.000
licensejopengroup	8479.62	1551141.38	0.005	0.996
licensejosi	3233.50	1962055.90	0.002	0.999
licensejosl	4909.38	1391422.00	0.004	0.997
licensejother	96741.78	1388916.90	0.070	0.944
licensejphp	1650.33	1410702.03	0.001	0.999
licensejphp-license	1841.88	1402221.78	0.001	0.999
licensejpsfl	151621.46	1402546.14	0.108	0.914
licensejpublic	7977.00	1389894.13	0.006	0.995
licensejpublic102	-171.00	1962055.90	-8.72e-05	1.000
licensejpublicdomain	2634.01	1390194.35	0.002	0.998
licensejpython	31089.27	1406787.93	0.022	0.982
licensejqpl	7660.74	1400533.88	0.005	0.996
licensejreal	488.50	2403017.90	2.03e-04	1.000
licensejserpl	2234.36	1451946.62	0.002	0.999
licensejerscpl	2456.25	1699190.25	0.001	0.999
licensejisissl	3464.90	1441810.74	0.002	0.998
licensejleeycat	2650.33	1602011.93	0.002	0.999
licensejusunpublic	1113.05	1431437.87	0.001	0.999

licensesybase	-258.50	2403017.90	-1.08e-04	1.000
licensevovida	141521.50	2403017.90	0.059	0.953
licensew3c	6103.85	1466725.08	0.004	0.997
licensewebsite	17795.22	1417222.64	0.013	0.990
licensewxwindows	4603.58	1439755.41	0.003	0.997
licensexnet	3206.37	1551141.38	0.002	0.998
licensezlib	37307.27	1394101.63	0.027	0.979
licensezope	7677.05	1455099.60	0.005	0.996

Residual standard error: 1962000 on 47734 degrees of freedom

Multiple R-Squared: 6.624e-05, Adjusted R-squared: -0.001295

F-statistic: 0.04865 on 65 and 47734 DF, p-value: 1

```
> anova(lm(downloads ~ license))
```

Analysis of Variance Table

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		