Survey of Post-Secondary Institution Server Software

Chris Sobczak

April 13, 2021

1 Abstract

The goal of the survey is to determine the market share that open source server software occupies in the academic setting. The figure to be estimated is the proportion of university web services that run opern source software. This includes all computers tha universities run for their websites, mail servers and other digital services that provide a portal for students and members of the public to access.

2 Background

Open source software or "free and open source software" (FOSS) is a piece or package of software that is distributed with the source code available for auditting and editting. The statistical software R is a great example of a flourishing FOSS project where users are able to contribute to the project and create new packages. Another example is the OpenBSD operating system. In the case of OpenBSD and other FOSS operating systems, the public has the ability to know exactly what is going on behind the scenes. This is not to say that every user is required to read and understand the low level source code, but it is a principle that the project is transparent and the software community can audit it, creating a web of trust.

In contrast, there is no reason for a consumer to trust in the word of a company that distributes closed source software. This organization can claim whatever they want about the security and privacy of their software but without having the right to audit the source code, there is no reason to believe any of those claims. If the only way a software package is considered *secure* is by the fact that no one know's how it works except the producer, then as soon as the code becomes available due to a leak, it can all be considered compromised as we have seen often from Microsoft (Willett 2021). FOSS is secure through industry standard security schemes, cryptography and by the knowledge of the thousands of programmers who have access to it (compared to the relatively smaller number of people who work on a single companies development team).

Finally, maybe the more important reason for institutions, FOSS is also (most of the time) free as in no cost (in addition to "free as in freedom"). Universities and other public institutions spend millions of dollars each year on software licenses for their students, staff and web services and other infrastructure. Increasing costs for post-secondary education and healthcare administration costs can be attributed to some extent to these rising software licensing prices. The increasing costs make higher education less excessible and harm the most vulnerable populations who rely on affordable healthcare. A very obvious way to reduce costs is be replacing these proprietary packages with equivalent and more secure free and open source software (Fitzgerald and Kenny 2004).

In addition to the problamatic cost of closed source, proprietary software, these costly, licensed software packages are most vulnerable to security threats. Security should be a high priority of all organizations but especially at public institutions where all of the software that staff and students interact with poses a threat to those individuals. Providing one for profit company exclusive access to your staff and students, in the age of survailliance capitalism is quite reckless (Yeung 2018).

So this study, establishes a base-line estimate of the proportion of free and open source servers being used by universities around the world. The questions that are answered include: Are there specific countries whose universities use more open source servers than others? What kinds of services are associated with certain operating systems?

3 Methods

This survey was conducted using cluster sampling. The primary sampling unit (psu) being the name of an institution, and the primary unit of interest being the associated domain(s). Each of the schools highest level registered domain names are probed to extract all their associated subdomains, defining the secondary sampling units (ssus). A census was then taken of these ssus to determine the software run at each domain name and calculate to proportion of the school's software was open source.

To provide an example, if Simon Fraser University (SFU) was draw into our sample, the unit of interest would be their registered domain name `sfu.ca`. Using the tool `findomain`, all of the subdomains under `sfu.ca` are collected into a file for probing. Some example subdomains are `mailgate.sfu.ca`, `imapnew.sfu.ca`, and `canvas.sfu.ca` for the schools email services and canvas portal. This is only a very short list of some of the possible subdomains as it is common to have thousands of subdomains associated to the same top name. Details on the tools I used in section 5 Tools.

3.1 Survey Design

Using cluster sampling, an SRS is taken of all the schools in the sampling frame (Hipo 2021). This sampling frame is an open source dataset of most of the universities in the world. Inevitably this list will have omitted some schools, but the set contains 9693 schools which is slightly on the low end of estimates, which may introduce bias in the conclusions of this study.

3.2 Sample Size Selection

The goal of this study is to estimate the proportion of university servers running an open source operating system using a 95% confidence interval with a margin of error of 0.03. Therefore, using Lohr 2019, "...surveys in which one of the main responses of interest is a proportion, it is often easiest to use that response in setting the sample size. For large populations, $S^2 \approx p(1-p)$, which attains its maximal value when $p = 1/2$. So using $n_0 = 1.96^2/(4e^2)$ will result in a 95% CI with width at most $2e$."

$$n_0 = \frac{z_{\alpha/2}^2 S^2}{e^2} = \frac{1.96^2(\frac{1}{2})(1-\frac{1}{2})}{e^2} \approx 1067$$

No need to use the finite population correction adjustment since the sample size is reasonable compared to the population size and the full dataset can be collected with a reasonable amount of resources.

3.3 Taking the Sample

With an appropriate sample size of $n = 1067$, using an R script to draw the sample, the selected school domains are saved in the `data.Rda` file found in the GitHub repository.
The sampling script can be found in the appendix. Within the sample, 1049 schools have only one registered domain, 16 schools have two registered domains and 2 have three domains.

$$N = 9693,\ n = 1067,\ m_0 = 2539008$$

The following 18 schools that were drawn in the sample have more than one domain name registered: Augusta University, University of Manchester, Universidad del País Vasco, Chinju National University of Education, Royal Holloway and Bedford New College, Northeastern University, University of the Pacific, Kwangju University, Kwangwoon University, University of Massachusetts at Lowell, St. Mary's University, University of Essex, Chonnam National University, Hanshin University, Savannah College of Art and Design, University of Technology Sydney, Universitat Pompeu Fabra, and Kyungil University.

These 16 schools with two domains and 2 schools with three domains accounts for the total 1087 ($1067 + (16 \text{ duplicates}) + (2 \times 2 \text{ more duplicates}) = 1087$) domains in the `domains` file of the GitHub repository. This file contained a few duplicates and subdomains for the school, so some of the domains were repeated and some were just included within the schools `subdomains/` file, reducing it to 1081 files in `subdomains/`.

All domains were separated onto their own line of the `probing/domains` file and processed with `findomain`. When extracting the subdomains, the `gen-subdomains` script only processed 1079 domains, identifying an inconsistency. Three of these domains that were not processed were `aloma.edu`, `student.uts.edu.au`, and `www.clcmn.edu`. In the case of `aloma.edu`, this is just a typo in the sampling frame for the Alamo Colleges' domain, which naturally is corrected to `alamo.edu`. The next missing domain `student.uts.edu.au`, for the University of Technology Sydney in Australia which is just a subdomain for their website `uts.edu.au`, identified as a duplicate domain. Finally, `www.clcmn.edu` for Central Lakes College-Brainerd is another subdomain for their college that was already extracted from `clcmn.edu`, another duplicated domain.

In the github repository, the file `undup-domains` is the audited file containing all of the highest level domains for the sample. From this file, the domains belonging to the same school are concatinated so that results can be organized by psu.

4 Results

The full results dataset of proportion of servers at each psu is available at the **github names . . . .**

4.1 Discussion and Conclusion

Any anticpiated shortcomings in your study desgin and their impact on your conclusions about the questions of interest.

- Non-responses
- Shortcomings in the sampling frame (may not actually contain 'all' schools in the world)

5 Tools

For taking the raw json file and selecting the sample, I used R and the `rjson` package. All project source files can be found at this GitHub repository. The R script outputs the base second and third level subdomains into a file with one domain per line, for which I run `findomain -t`. This tool takes a domain and searches various databases and tests the domain for subdomains associated with it. I take this and output all the subdomains into a file for each school, and then test if the service is up.

Documentation for the tools can be found in the Appendix

5.1 Extracting Info From SSUs

```
curl -I
```

References

Fitzgerald, B. and T. Kenny (2004). "Developing an information systems infrastructure with open source software". In: *IEEE Software* 21.1, pp. 50–55. DOI: `10.1109/MS.2004.1259216`.

Hipo (2021). *university-domains-list*. `https://github.com/Hipo/university-domains-list`.

Lohr, Sharon L (2019). *Sampling: Design and Analysis*. eng. 2nd ed. Chapman & Hall/CRC Texts in Statistical Science Series. Milton: CRC Press. ISBN: 0367273462.

Willett, Marcus (2021). “Lessons of the SolarWinds Hack”. In: *Survival* 63.2, pp. 7–26. DOI: 10.1080/00396338.2021.1906001. eprint: https://doi.org/10.1080/00396338.2021.1906001. URL: https://doi.org/10.1080/00396338.2021.1906001.

Yeung, Karen (2018). “Five fears about mass predictive personalization in an age of surveillance capitalism”. eng. In: *International data privacy law* 8.3, pp. 258–269. ISSN: 2044-3994.

6 Appendix