Survey of Post-Secondary Institution Server Software

Chris Sobczak

April 9, 2021

1 Abstract

The goal of the survey is to determine the market share of that linux and other open source operating systems occupy in the academic setting. So the figure of interest is the proportion of machines that universities run for thier websites, mail servers and other digital services that run an open source operating system.

2 Background

Universities spend millions of dollars each year on licenses for their students' software offerings and also for their own infrastructure. Increasing costs for post-secondary education are making it less excessible and a very obvious way to reduce costs is find free and open source alternatives to those expensive licenses. **(cite some study showing costs and how much licenses contribute to the overall cost of attendance).**

In addition to the problamatic cost of closed source, proprietary software, it is also less secure than the open source alternatives **(cite study about this or reference the OpenBSD claims etc).** Especially at public institutions, all of the software that staff and students are required to interact with should also be open source so for the sake of privacy and security, universities should not be given the exclusive access to a specific class of student.

So in this study, I would like to establish a base-line understanding of the proportion of free and open source servers being used by universities around the world. **(look at poststratification)** 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 $2e$."

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

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 `subdomains/` directory of the GitHub repository.

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 list of domains, there were 16 duplicated domains and this also accounts for the difference in number of domains that went through the `gen-domains` script and the original sampled list. Here is a list of the duplicated domains: `aku.edu`, `bashedu.ru`, `most.gov.mm`, and `uwo.ca`.

After removing the duplicated domains and generating the domains for the corrected `alamo.edu`, I ended up with a full set of 1081 lists of subdomains.

In the github repository, the file `undup-domains` is the audited file containing all of the highest level domains for the sample.

The next step will be ensuring that two related domains did not go through the process, for example making sure that we did not run `gen-domains` on `sfu.ca` and `mail.sfu.ca`, since this would likely result in duplicated information.

Using the line `cat subdomain/* | uniq -d` we can see that there were no repeated lines.

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

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.

6 Appendix