Survey of Operating Systems Used by World Universities

Chris Sobczak

April 2, 2021

1 Introduction

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 Methodologies

I have used a dataset of all known domains registered by universities around the world (Hipo 2021). I have taken a simple random sample of these schools and then extracted all of the corresponding subdomains.

For example, Simon Fraser University (SFU) has `sfu.ca` registered with the Canadian Internet Registration Authority (CIRA) and they have thousands of subdomains under `sfu.ca` such as `mail.sfu.ca`, `canvas.sfu.ca` and `go.sfu.ca`. A lot of the subdomains do not contain relevant web content or are otherwise unused, so to filter out these subdomains before trying to identify the operating systems being used to run the surver at that address, I have used the `host` DNS lookup tool and other methods for checking if there are actual services running at that address. More about the tools I used in section 3 Tools.

2.1 Sampling Frame

2.2 Sample Size Selection

$e = 0.03$ and $\alpha = 0.05$ Lohr 2019

We want to estimate the proportion of university servers run an open source operating system using a 95% confidence interval with a margin of error of 0.03.

Lohr 2019 page 47: "In 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 fpc adjustment since the sample size is reasonable compared to the population size.

3 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

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.