

Predicting Software Vulnerabilities in the Free Open Source Software Ecosystem

Elsie Phillips, Carlos Jensen

Open Source

Open Source software differs from traditional closed source proprietary software in several ways:

- Users can run, study, modify, and redistribute the software as they wish, for any purpose
- User may also then submit those modifications back to the original developers, to be incorporated into the canonical project
- Much of the development work on open source software is done by volunteers and funded by donations

Impact of Vulnerabilities

Software Dependencies Make Open Source Software an Ecosystem

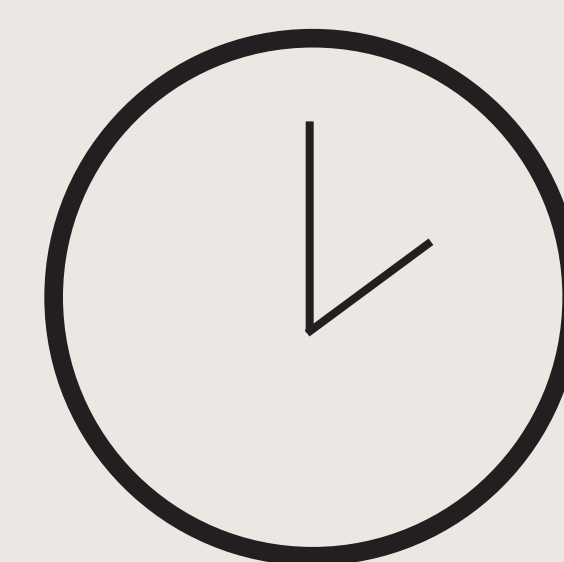
The adoption of open source software has rapidly increased over the last two decades. Many open source projects are interdependent, so the impact of a security vulnerability in one project can have a significant effect on many projects. An example of this was the “Heartbleed” vulnerability in OpenSSL in 2014. 17% of the world’s secure servers were affected, and the cost of the vulnerability was estimated at over \$500 million.



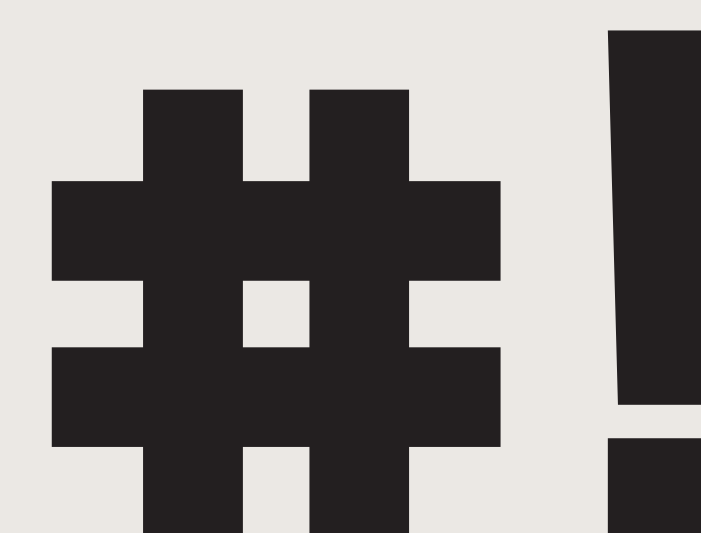
The Model

Our goal was to develop a model to predict vulnerabilities in highly used open source software projects.

For our purposes, “highly used” refers to projects in the Debian Popularity Contest’s top 150. We also only looked at projects with a public git repository. This resulted in a list of 85 open source projects.



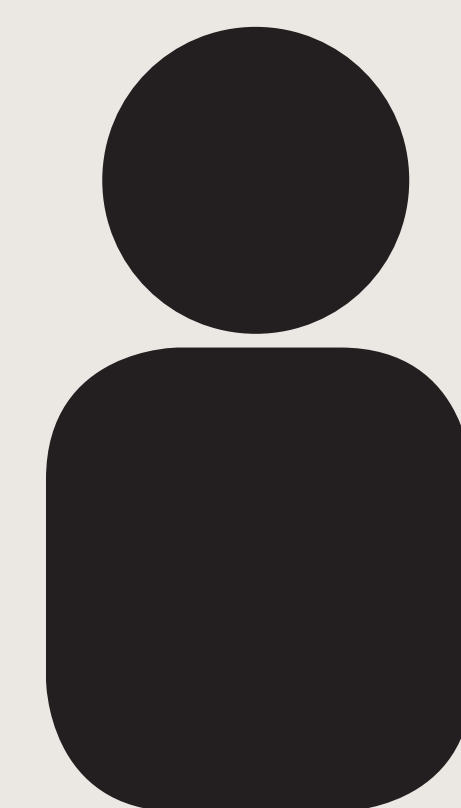
Age of the project



Lines of Code



Constructive Cost Model




Total Contributors

Active Contributors (Jan 2015-Dec 2015)

We hypothesized that such a model could be constructed using measures of development effort available to projects, and the complexity of projects.

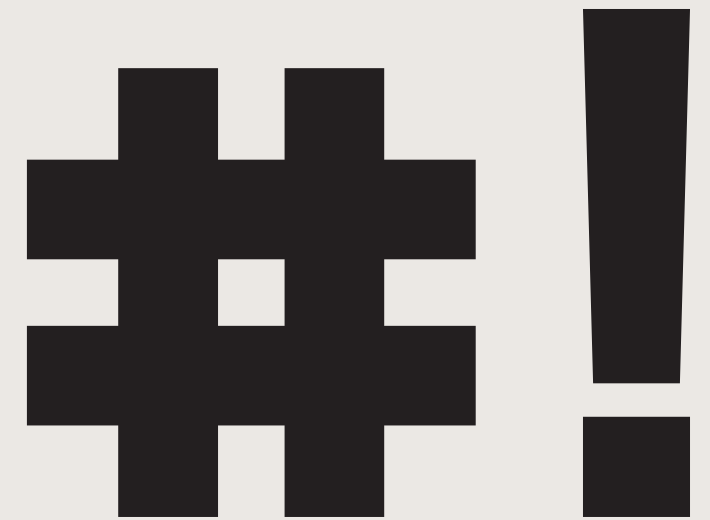
Results

Using data from the Common Vulnerabilities and Exploits database, we discovered the following relationships:


$$= -3.5$$

developer years = total number of contributors x age of the project

For every 100 developer years added to a project, predicted vulnerabilities decreases by 3.5 a year.


$$= 1.67$$

Lines of Code

Every 100,000 lines of code a project has increases the predicted number of predicted vulnerabilities by 1.67 a year.