Using Github to Create a Dataset of Natural Occurring Vulnerabilities

Sofia Reis[1] and **Rui Abreu**[2]
[1]Faculty of Engineering of University of Porto, Portugal
e-mail: sofia.reis@fe.up.pt
[2]IST, University of Lisbon & INESC-ID, Portugal
e-mail: rui@computer.org

Abstract

Currently, to satisfy the potential high number of system requirements, complex software is crafted which turns its development cost-intensive and more susceptible to security vulnerabilities. According to IBM's X-Force Threat Intelligence 2017 Report, the number of vulnerabilities per year has been significantly increasing over the past years. In software security testing, performing empirical studies is challenging due to the lack of widely accepted and easy-to-use databases of real vulnerabilities as well as the fact that it requires both human effort and CPU time. Consequently, researchers tend to use databases of hand-seeded vulnerabilities, which may differ inadvertently from real vulnerabilities and thus might lead to misleading assessments of the capabilities of the tools. Although there are databases targeting security vulnerabilities test cases, only one database contains real vulnerabilities, the other ones are a mix of real and artificial or even only artificial samples. This paper explains our efforts to create a vulnerability database, *Secbench*, by mining 238 repositories from GitHub. GitHub is particularly interesting since it hosts millions of open-source projects carrying a considerable number of security vulnerabilities. More than $1M$ of commits were mined for 16 different patterns which yielded 602 security vulnerabilities. The study described in this paper provides a methodology to mining security vulnerabilities from open-source software. Our methodology has proven itself as being valuable since we were able to collect a considerable number of security vulnerabilities from a small group of repositories. However, there is still much work to do in order to improve not only the mining process but also the vulnerabilities diagnosis. All the information related to *Secbench* is available at/through `https://tqrg.github.io/secbench/`.