Survey Paper - Identifying Known Vulnerabilies in OSS libraries

Ronaldo Canesqui
Southern Adventist University
ronaldocanesqui@southern.edu

***Abstract*—The use of open-source software libraries is experiencing exponential growth in the last decade. Modern software development practices fully embrace this model. Unfortunately, the advantages brought by reuse also brings the risk of vulnerabilities. The number of vulnerabilities found in open-source libraries is increasing faster than the number of available packages. How to leverage the power of open-source libraries and still manage the risk of vulnerabilities they carry is a problem that either remain ignored by many software engineers or present considerable challenges to them. Studies on dependencies management practices found that security criterion is not considered by the majority of developers and one-third of them are not aware of libraries update. This paper compiles the most recent studies on identifying known vulnerabilities on dependencies are presented, along with analytical data about vulnerability lifespan and current dependencies management practices.**

I. Introduction

The use of open-source libraries is a reality in the software industry. In 2018, Synopsys[1] found open-source code in more than 96% of its audits[20]. Open-source software (OSS) usage has increased exponentially during the last decade. The most popular source for libraries in the Java ecosystem, Maven Central Repository, growth 542.17% from 2010 to 2016 [10]. Another popular open-source repository for the Javascript ecosystem, npm growth from 0 packages at its creation in 2010 to 1,000,000 packages in 2019[2]. The figure 1 extracted from [6] shows the number of packages evolution on other popular ecosystems. How to leverage the power of open-source libraries and still manage the risk of vulnerabilities they carry is a problem that either remain ignored by many software engineers or present considerable challenges to them.

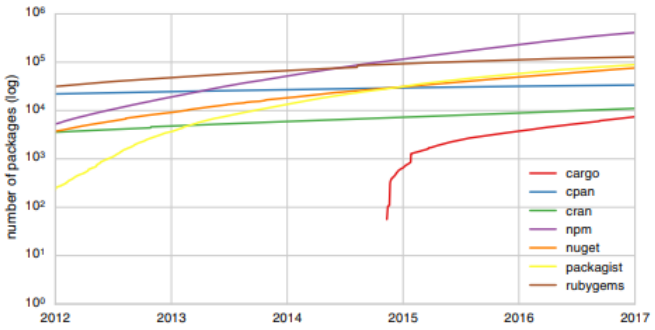

Fig. 1. Evolution of the number of packages.

A 2014 study from Sonatype determined that over 6% of the download requests from the Maven Central Repository were for component versions that included known vulnerabilities. In their review of over 1,500 applications, each of them had an average of 24 severe or critical flaws inherited from their components [3]. A white paper produced by Contrast Security stated that over 25% of all libraries downloaded from Maven Central Repository has a vulnerability. Only one vulnerable version of the Java GWT package was downloaded 17,666,703 times [21]. Among the 10 most popular npm packages, 6 present 1 or more vulnerabilities [4].

In a study on security vulnerabilities impact [5] the researchers found that out of 610,097 available packages (2017 data) 133,602 packages directly depend on a vulnerable package and 72,470 packages had at least one release that relies on a vulnerable package. In the same study, they also found a crescent number or vulnerabilities in the npm repository. Figure 2 shows the evolution of the number of vulnerabilities and the impact on dependent packages. Comparing 2014 and 2017 data, there is a clear increase on the impact that a vulnerable package (dotted lines) has over dependent packages (straight lines).

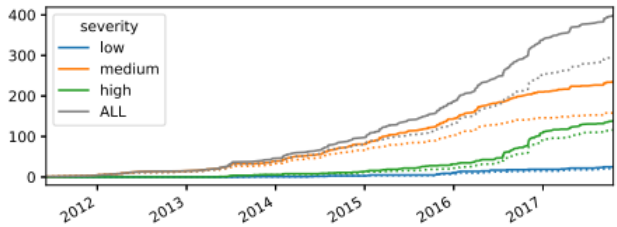

Fig. 2. Evolution of the number of discovered vulnerabilities (straight lines) and corresponding distinct packages (doted lines) per severity.

During this paper, the most recent studies on identifying known vulnerabilities on dependencies are presented, along with analytical data about vulnerability lifespan and current dependencies management practices. The rest of this paper will cover studies on the vulnerability lifespan II, current dependencies management practices III, proposed solutions on

[1]https://www.blackducksoftware.com/
[2]https://snyk.io/blog/npm-passes-the-1-millionth-package-milestone-what-can-we-learn/ accessed 10/10/2019
[3]Report published January 02, 2015 at http://goo.gl/i8J1Zq.
[4]https://snyk.io/blog/npm-passes-the-1-millionth-package-milestone-what-can-we-learn/ accessed 10/10/2019

how to identify known vulnerabilities and evaluate updatability IV, overview of the main challenges in the area V and conclusion VI.

II. Vulnerability Lifespan

To analyze the vulnerability lifespan the study [5] used a 700 security vulnerabilities report made available by Snyk.io [5] and retrieved the list of corresponding releases from the open-source discovery service libraries.io [13]. Based on the list of releases, they identified which ones were affected by the vulnerability. Based on the relationship between package, vulnerability, first release data, vulnerability discovery date, it was possible to trace a vulnerability timeframe.

How long does the package remain vulnerable Figure 3 shows Kaplan-Meier estimator curve [9] for the event "vulnerability is fixed". The data presented considers the date of the affected release and the date that the fix was available. After 10 months, there is a probability higher than 80% that a high severity vulnerability is still not fixed.

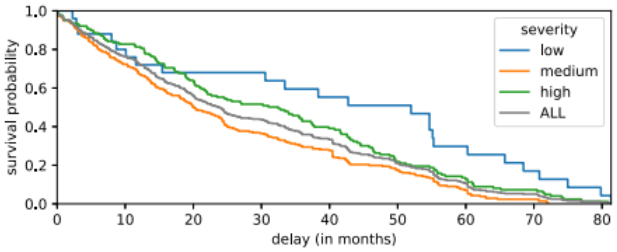

Fig. 3. Survival probability for event "vulnerability is fixed" w.r.t. the date of first affected release.

When a vulnerability is discovered Figure 4 shows that most severities are discovered in old packages. 75% of all vulnerabilities are found in libraries older than 13 months. Even not highlighted in the original study, the shorter wave in high severity vulnerabilities may suggest the higher priority in which they are handled, especially when compared to the low severity graph that shows smother curves. Most severities are found in packages older than 28 months.

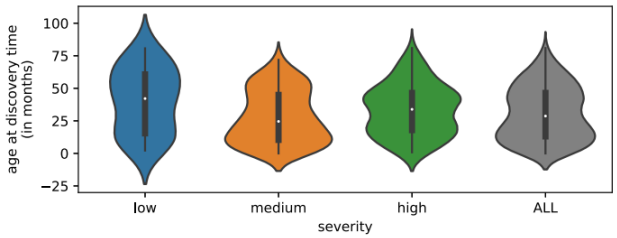

Fig. 4. Violin plots of packages age at discovery time by vulnerability severity.

When a vulnerability is fixed Most of the vulnerabilities are fixed between the discovery date and the public announcement. Figure 5 shows that there is a probability of 50% of a fix becomes available in the first month after discovery. And 80% of all vulnerabilities are fixed between 12 and 13 months. After 20 months of the discovery all high severity vulnerabilities were fixed. Some medium severity vulnerabilities, according to the graph, will take more than 40 months to be fixed.

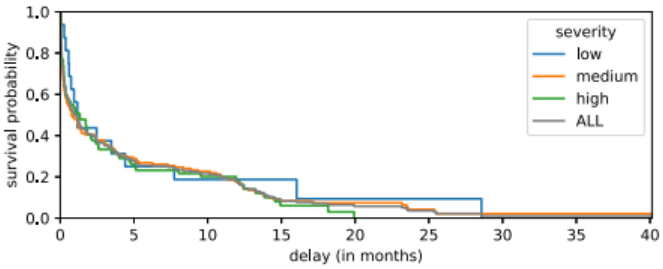

Fig. 5. Survival probability for event "vulnerability is fixes" w.r.t. vulnerability discovery time.

When a vulnerability is fixed in a dependent package The analysis of the impact on dependent packages is important due to the fast increase of dependencies. Figure 6 shows a more step curve when compared to the number of available packages (figure 1). Figure 7 shows that after 20 months, 100% of high severity vulnerabilities are fixed while there are 40% of the dependencies vulnerable at the same time. The slowness on updating dependencies found in this study is supported by similar findings in different studies covering a wide range of ecosystems such as SmallTalk[18], Pharo[8], Java [19], Apache products[2], Windows ecosystem [12], and Javascript ecosystem [11].

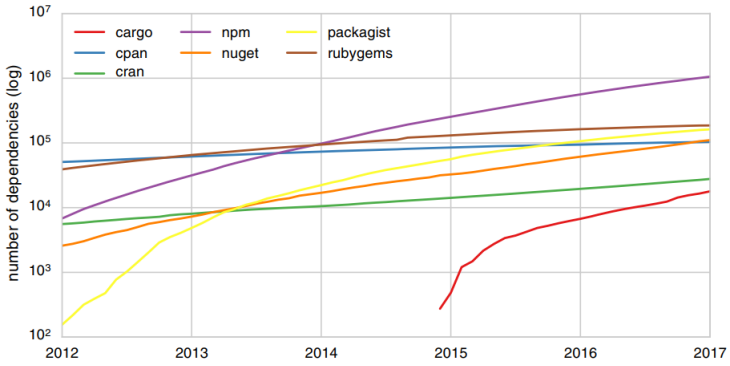

Fig. 6. Evolution of the number of dependencies (considering for each point in time the latest available release of each package).

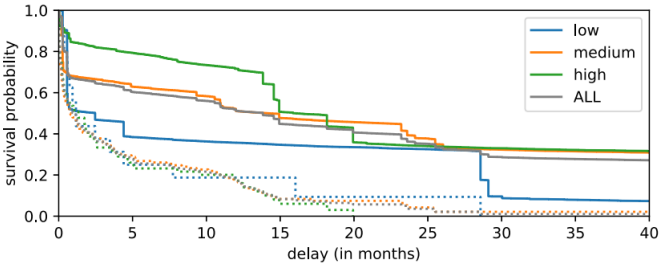

Fig. 7. Survival probability for event "package is fixed" w.r.t. vulnerability discovery time. Dependent packages are shown as straight lines and upstream packages as dotted lines.

[5]https://snyk.io

III. Current Dependencies Management Practices

To understand the vulnerabilities lifespan it is required to understand current dependencies management practices specially on updating and selecting dependencies. A study [7] conducted with 203 app developers from Google Play clarified the following research questions: *How frequently do developers update their apps/libs and what is their main motivation for updates? What are possible reasons to not update dependencies and what solutions could app developers think of?* Based on the application maintainers' answers, the study concludes that 78% of them do not have a fixed schedule for app updates (Figure 8) and only 33% of them mention a library update as a reason to update their app (Figure 9). When the maintainers were questioned why do you update your app's libraries, 96.47% of them answered bug fixing and 57.65% mentioned security (figure 10). Regarding library selection criteria, only 26.58% answered security, even thought the answer update frequency is related and received 35.16% of the votes (figure 11). When asked reasons why your app would include outdated libraries?, 57.03% answered that the library was still working, 50% answered to prevent incompatibilities and 32.81% were unaware of updates (figure 12).

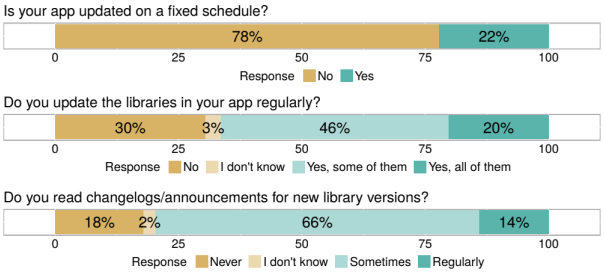

Fig. 8. Anwsers for questions regarding app/library release frequency.

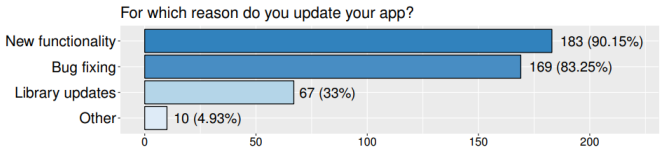

Fig. 9. Answers for questions:For which reason do you update your app?

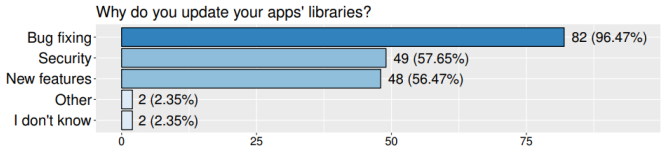

Fig. 10. Answers for questions: Why do you update your apps' libraries?

The study concludes that the main reasons to keep a library outdated are incompatibility prevention, unaware of updates and too much effort. Another study on this field [10] reaches the same conclusion regarding why maintainers keep outdated dependencies. In the following sections, studies that proposed ways to deal with these problems are presented along with their conclusions.

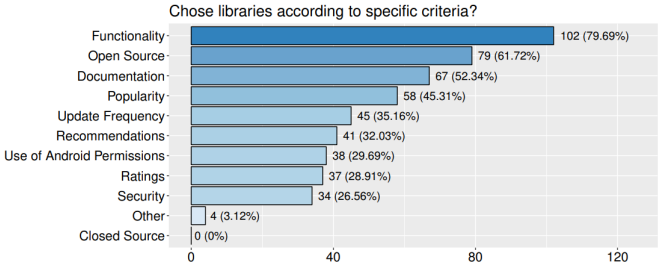

Fig. 11. Answers for questions: Chose libraries according to specific criteria?

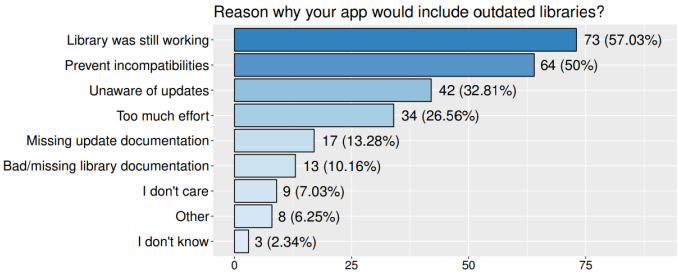

Fig. 12. Answers for questions: Reason why your app would include outdated libraries?

IV. Identifying known Vulnerabilities and Analyzing Updatability

Various researchers have worked to solve the problem of know vulnerabilities in open-source packages and their dependencies. In this section, the most recent works will be presented along with their conclusions. This is not a comprehensive list but it offers a similar variety of methods and complexity of other works not covered here.

Predicting vulnerabilities: Although the focus of this work is on papers that detect vulnerabilities using know vulnerabilities databases, it is worth mentioning that there is a branch of proposed solutions that operates identifying vulnerabilities based on a set of characteristics. Part of these solutions uses machine learning to identify vulnerabilities. One of these works [15] mapped the CVE[6] entries to the commit that generated the vulnerability and trained an SVM-based [7] model using the commit's metadata. According to the authors, this method reduced the number of false positives by 99% when compared to Flawfinder[8]. It was capable of detecting 53 of the 219 known vulnerabilities used in the study and only producing 36 false positives. The proposed solution analyses each commit and tries to predict the presence of vulnerabilities.

Measuring dependency freshness: The study [4] proposes a method to measure the dependency freshness. The authors used information from the projects' *pom.xml* file to determine the libraries' version and used the Maven Central Repository to determine the release history for the library. Based on this information, the authors were able to classify each used component in a risk profile with 4 categories: low, moderate,

[6]Common Vulnerabilites and Exposures - https://cve.mitre.org/
[7]Support Vector Machine
[8]https://https://dwheeler.com/flawfinder/

high and very high. To determine the relationship between their risk profile and the security vulnerabilities the researchers matched the dependencies list and tried to match with the CVE vulnerability database. They were able to find a correlation between their freshness index and the number of vulnerabilities. The median variance of the rating, the researchers were able to classify the systems according to dependency freshness:

- **Stable** Systems with a stable dependency freshness rating. The system dependencies see little to no updates.
- **Improving** Systems with an increasing dependency freshness rating. Dependencies are updated faster than they are released.
- **Declining** Systems with a decreasing dependency freshness rating. Dependencies are updated slower than they are released.

The metrics presented have great potential in quantifying the dependency freshness. The method has at least one serious limitation: only direct dependencies can be measured, which can generate wrong classifications, especially when a direct dependency is updated but a transient dependency exists and is outdated.

Library updatability The study [4] was able to correlated the library freshness with number of vulnerabilities. The finding that old libraries carry more vulnerabilities is also supported by other works [5], [20]. The study [7] proposes an automated method to evaluate the updatability of dependencies. Their method consists of three steps: determine the API robustness, determine the library usage and determine the library updatability. To determine the API robustness, the researchers extracted the public API from multiple versions of the same library, resulting in a library version/API pair list. According to the authors, this is a more fine-grained approach compared to [1]. The authors then proceed with the library usage, where they inspect the bytecode looking for a call to the API. Finally, they matched the library version/API pair list with the library usage to produce the library updatability, which informs what is the latest version the API can be updated to without causing incompatibilities.

The authors test their approach analyzing 98 libraries and 1,246,118 apps from Google Play. The results show that in 85.6% of the cases the identified library can be upgraded by at least one version (Upgrade1+) and on 48.2% of the cases the library can be updated to the most current version simply by replacing the old library, without any code change. Figure 13 shows the libraries' updatability from their study.

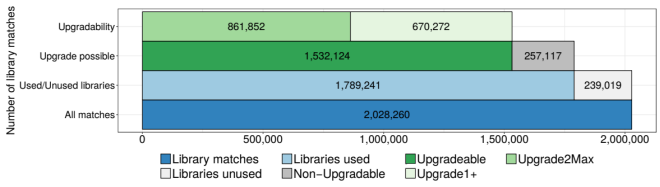

Fig. 13. Library updatability of current apps on Google Play.

Vulnerability Alert Service (VAS) The Vulnerability Alert Service (VAS) is proposed by [3]. Their focus is to increase the awareness about known vulnerabilities by making it part of the software quality process. The overall process is illustrated in figure 14. First, the target project has its dependencies extracted and recognized. Then, list of dependencies is analyzed by the matching task, which tries to match the dependency with an entry from the vulnerability disclosures, in this study a CVE entry. Upon a successful match, an alert is produced which is consumed by a human operator. The authors extended the OWASP Dependency Check Tool to extract the dependencies list from Maven *pom.xml* files and used this tool as the vulnerability checker described in the diagram. The conclusion of this work shows a false positive rate of over 70%, but the authors state that the rate is smaller when the solution was actually deployed. The researchers affirm that the high false-positive can be caused by situations like the MySQL-connector jar which is flagged with the vulnerabilities of the MySQL database server. The evaluation with the human operators shown that despite the false positive rate, an alert system was considered useful. Even the authors were satisfied with the operators' evaluation, since their solution looks for dependencies in *pom.xml* files, only direct dependencies can be identified.

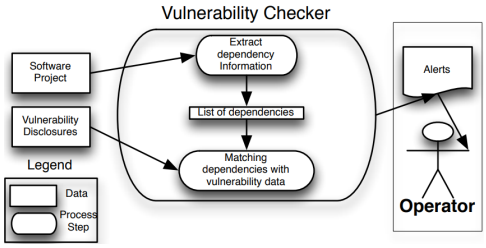

Fig. 14. The Vulnerability Alert Service process

Code-based vulnerability identification The study [17] uses a method to identify vulnerabilities that, according to the authors, is more accurate than the methods proposed so far. Their work aims to answer the questions: *does my application is vulnerable if it uses a vulnerable dependency?* The researchers claim that this is not always the case and other studies failed to answer this question. The authors state that the software that contains a certain library that has known vulnerability does not need to be subject to costly efforts if proven that the vulnerability does not present harm. This way, the fix can be applied in the subsequent releases under no additional cost. Conceptually their solution can be used on any ecosystem, but their implementation is based on the Java ecosystem and the Maven Central Repository. They used the project's *pom.xml* file to determine the dependencies. After the dependencies are determined, they remove all dependencies classified as provided and test, since those will not be deployed. Then the halted libraries[9] are identified and removed. One key aspect for this approach to work is a database vulnerability that contains the meta-data about the library, such as name and version according to Maven Central Repository, the vulnerable source code and the source code for the fix. Using this database is possible to associate the dependency identified previously

[9]Libraries that do not receive an updated in a certain period of time define by the study.

and the vulnerability, along with the source code involved in the vulnerability. This database was manually produced by the authors using CVE entries and it covers 90% of Java ecosystem vulnerabilities. The authors used static and dynamic analysis to determine if the vulnerable code is reachable. The dynamical analysis is done by instrumenting the vulnerable code and running automated and manual tests. Their solution will notify the user if the vulnerability has impact on the application, meaning, the vulnerable code can be reached and also notify about the presence of halted libraries.

Other studies from the same authors use the same method to identify known vulnerabilities but focus on different aspects of the solution. [16] introduces a simpler but very similar method to the one previously described. It was presented as a proof-of-concept and targets performance comparison between their approach, OWASP Dependency Check and two commercial tools. [14] focuses on how the impact of know vulnerabilities is inflated because other studies are counting vulnerable dependencies that are not deployed, such as libraries that are only used during tests.

According to the researchers, the code-based identification method is superior to the ones based on meta-data, since it does not suffer from problems such library artifacts renamed, named inconsistently and the fact that libraries can be extracted and repacked as a single self-contained archive. All these problems affect meta-data based identification. In the study [16] the authors claim this method is the state-of-art in the filed. This method is the officially recommended by SAP to scan projects and it is used by over 500 distinct development projects.

V. Main Challenges

Although other studies on identifying known vulnerabilities do not explicitly mention the main challenges while trying to associate a library with a known vulnerability the study [16] does. Their challenges can be observed implicitly on all studies compiled in this work. To establish this relationship, information from different sources needs to be integrated, such as vulnerabilities database, libraries repository, and source code repository. The authors state that multiple problems hinder this integration, requiring ad-hoc and technology-specific solutions. Such problems hindered the detection of known vulnerabilities. As evidence, all studies compile in this work have at least one manual step that in its absence, the work would be infeasible.

- *Non-uniform reporting of products affected by a vulnerability* - In some cases, only the CPE of the respective library is mentioned, in other cases, products or libraries using the library are also listed.
- *Vulnerability and dependency management make use of different naming schemes and nomenclatures* - CPE does not straight map to packages repositories such as Maven or npm.
- *Vulnerabilities and VCS information of the respective patch are not linked in a systematic and machine-readable fashion* - While committing a fix to the repository, developers do not have a standard way to refer to the CPE.

Improved know vulnerabilities databases and standard ways to identify the OSS libraries in the repositories and commits would increase significantly the automation level and the impact of the available solutions.

VI. Conclusion

There are several works that tackle different aspects of the known vulnerability identification problem. Some of them directly identify the vulnerability, such as the code-based method, while others offer an indication of possible vulnerabilities such as the dependency freshness metrics. There is no solution that will fit all cases yet. Even in production ready solutions, there is a dependency on specific technologies and manually built databases. This paper covered a set of different solutions that can help software development practitioners to streamline their dependency management process.

References

[1] M. Backes, S. Bugiel, and E. Derr, "Reliable third-party library detection in android and its security applications," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 356–367. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978333

[2] G. Bavota, G. Canfora, M. Di Penta, R. Oliveto, and S. Panichella, "How the apache community upgrades dependencies: an evolutionary study," *Empirical Software Engineering*, vol. 20, no. 5, pp. 1275–1317, Oct 2015. [Online]. Available: https://doi.org/10.1007/s10664-014-9325-9

[3] M. Cadariu, E. Bouwers, J. Visser, and A. van Deursen, "Tracking known security vulnerabilities in proprietary software systems," in *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, March 2015, pp. 516–519.

[4] J. Cox, E. Bouwers, M. van Eekelen, and J. Visser, "Measuring dependency freshness in software systems," in *Proceedings of the 37th International Conference on Software Engineering - Volume 2*, ser. ICSE '15. Piscataway, NJ, USA: IEEE Press, 2015, pp. 109–118. [Online]. Available: http://dl.acm.org/citation.cfm?id=2819009.2819027

[5] A. Decan, T. Mens, and E. Constantinou, "On the impact of security vulnerabilities in the npm package dependency network," in *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*, May 2018, pp. 181–191. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8595201

[6] A. Decan and T. M. P. Grosjean, "An empirical comparison of dependency network evolution in seven software packaging ecosystems." [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10664-017-9589-y.pdf

[7] E. Derr, S. Bugiel, S. Fahl, Y. Acar, and M. Backes, "Keep me updated: An empirical study of third-party library updatability on android," *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017. [Online]. Available: https://dl.acm.org/citation.cfm?id=3134059

[8] A. Hora, R. Robbes, N. Anquetil, A. Etien, S. Ducasse, and M. T. Valente, "How do developers react to api evolution? the pharo ecosystem case," in *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, Sep. 2015, pp. 251–260.

[9] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. [Online]. Available: http://www.jstor.org/stable/2281868

[10] G. D. O. A. Kula, R.G., "Do developers update their library dependencies?" *Empirical Software Engineering*, vol. 23, 2018. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10664-017-9521-5.pdf

[11] T. Lauinger, A. Chaabane, S. Arshad, W. Robertson, C. Wilson, and E. Kirda, "Thou shalt not depend on me: Analysing the use of outdated javascript libraries on the web." [Online]. Available: https://arxiv.org/abs/1811.00918

[12] A. Nappa, R. Johnson, L. Bilge, J. Caballero, and T. Dumitras, "The attack of the clones: A study of the impact of shared code on vulnerability patching," in *2015 IEEE Symposium on Security and Privacy*, May 2015, pp. 692–708.

[13] A. Nesbitt and B. Nickolls, "Libraries.io Open Source Repository and Dependency Metadata," Jun. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.808273

[14] I. Pashchenko, H. Plate, S. E. Ponta, A. Sabetta, and F. Massacci, "Vulnerable open source dependencies: Counting those that matter," in *Proceedings of the 12th International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Oct 2018.

[15] H. Perl, S. Dechand, M. Smith, D. Arp, F. Yamaguchi, K. Rieck, S. Fahl, and Y. Acar, "Vccfinder: Finding potential vulnerabilities in open-source projects to assist code audits," 10 2015, pp. 426–437.

[16] H. Plate, S. E. Ponta, and A. Sabetta, "Impact assessment for vulnerabilities in open-source software libraries," in *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, Sep. 2015, pp. 411–420.

[17] S. E. Ponta, H. Plate, and A. Sabetta, "Beyond metadata: Code-centric and usage-based analysis of known vulnerabilities in open-source software," in *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, Sep. 2018, pp. 449–460.

[18] R. Robbes, M. Lungu, and D. Röthlisberger, "How do developers react to api deprecation?: The case of a smalltalk ecosystem," in *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*, ser. FSE '12. New York, NY, USA: ACM, 2012, pp. 56:1–56:11. [Online]. Available: http://doi.acm.org/10.1145/2393596.2393662

[19] A. A. Sawant, R. Robbes, and A. Bacchelli, "On the reaction to deprecation of 25,357 clients of 4+1 popular java apis," in *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, Oct 2016, pp. 400–410.

[20] Synopsys, "2019 open source security and risk analysis," *Synopsys Open Source Security and Risk Analysis*, 2019.

[21] J. Williams and A. Dabirsiaghi, "The unfortunate reality of insecure libraries," *Asp. Secur. Inc*, pp. 1–26, 2012.