A Curated Dataset of Security Defects in Scientific Software Projects

Hot Topics in the Science of Security (HotSoS) 2020
September 22nd, 2020

Justin Murphy
Tennessee Technological University
Cookeville, TN
jdmurphy43@tntech.edu

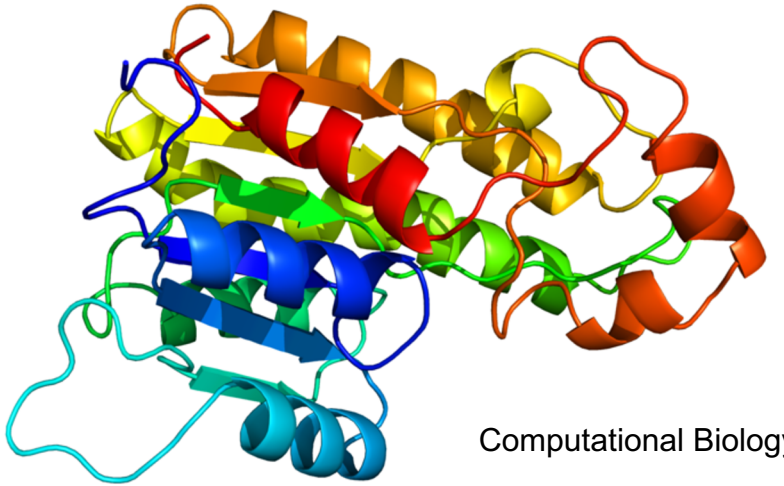
Shazibul Islam Shamim
Tennessee Technological University
Cookeville, TN
mshamim42@tntech.edu

Elias T. Brady
Tennessee Technological University
Cookeville, TN
etbrady42@tntech.edu

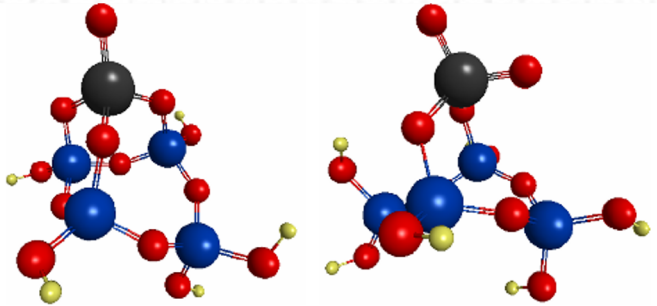
Akond Rahman
Tennessee Technological University
Cookeville, TN
arahman@tntech.edu

Scientific Software

Astrophysics



Computational Biology



Computational Chemistry

Scientific Software in Julia

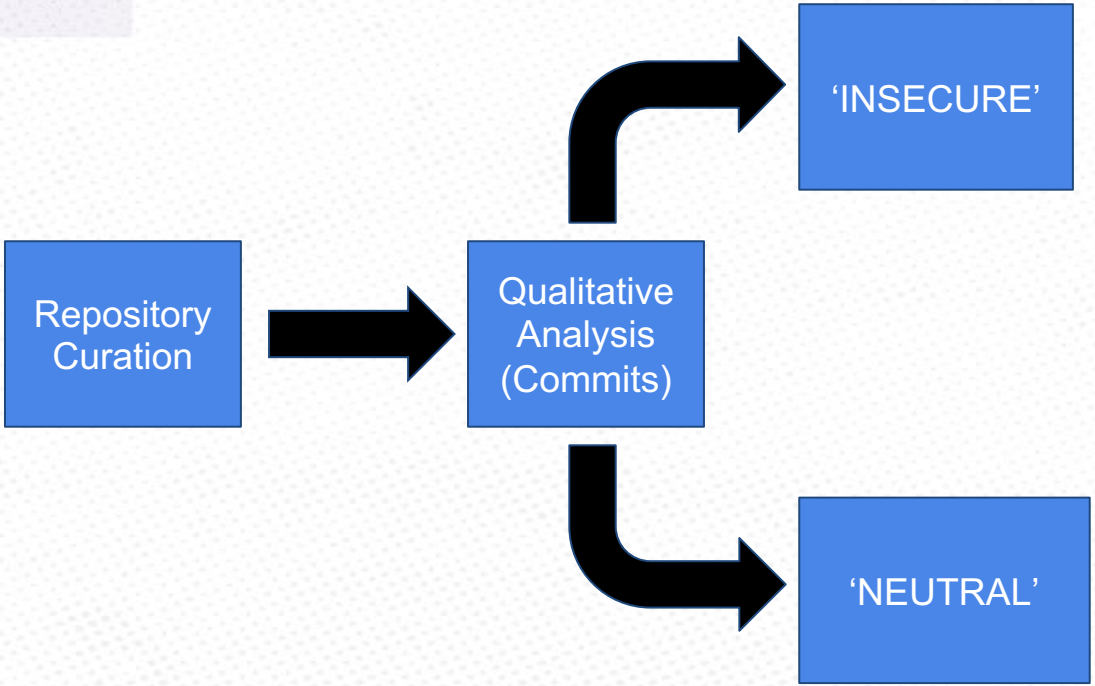
julia

CANCER RESEARCH UK

Research Question

- RQ: How frequently do security defects appear in scientific software projects?

Methodology



Nuthan Munaiah, Steven Kroh, Craig Cabrey, and Meiyappan Nagappan. 2017. Curating GitHub for engineered software projects. Empirical Software Engineering (2017), 1–35. https://doi.org/10.1007/s10664-017-9512-6

Curated Dataset (Repository Curation)

Initial Repo Count	3,405,303
Criteria-1 (1% Julia files)	3,866
Criteria-2 (Available)	3,115
Criteria-3 (Not a clone)	2,173
Criteria-4 (Commits/Month $\geq$ 2)	2,173
Criteria-5 (Contributors $\geq$ 5)	253
Criteria-6 (CI)	20
Final Repo Count	20

Curated Dataset (Qualitative Analysis)

Identifying the Characteristics of Vulnerable Code Changes: An Empirical Study

Amiangshu Bosu
Dept. of Computer Science
University of Alabama
Tuscaloosa, AL USA
asbosu@ua.edu

Jeffrey C. Carver
Dept. of Computer Science
University of Alabama
Tuscaloosa, AL USA
carver@cs.ua.edu

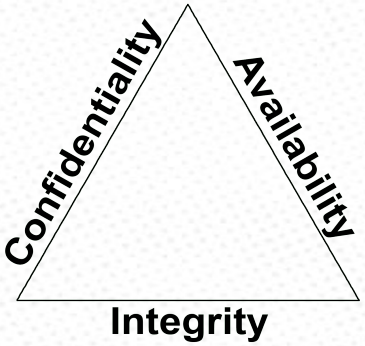
Munawar Hafiz
Dept. of Computer Science
Auburn University
Auburn, AL USA
munawar@auburn.edu

Patrick Hilley
Dept. of Math and Computer Science
Providence College
Providence, RI USA
philley@friars.providence.edu

Derek Janni
Dept. of Mathematical Sciences
Lewis & Clark College
Portland, OR USA
derekjanni@lclark.edu

22nd ACM SIGSOFT International Symposium FSE 2014

Example commit message labeled ‘INSECURE’
:
“exploit utf16 support in julia 0.3 if available for proper utf16 conversion”



Results

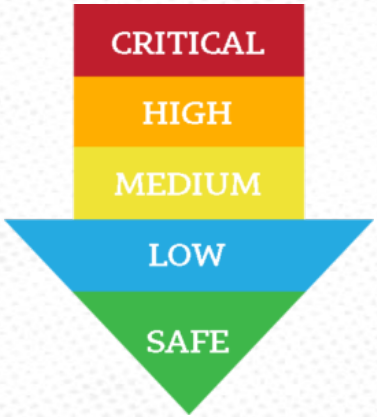
Initial Repo Count	3,405,303
Criteria-1 (1% Julia files)	3,866
Criteria-2 (Available)	3,115
Criteria-3 (Not a clone)	2,173
Criteria-4 (Commits/Month ≥ 2)	2,173
Criteria-5 (Contributors ≥ 5)	253
Criteria-6 (CI)	20
Final Repo Count	20

308/7024 = 4.4%
'INSECURE'

Cohen's Kappa = 1.0

Future Work

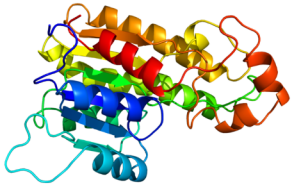
SECURITY DEFECT CATEGORIZATION & PREDICTION RESEARCH

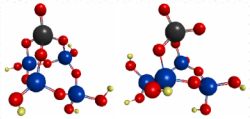


FIND UNDISCLOSED SECURITY DEFECTS

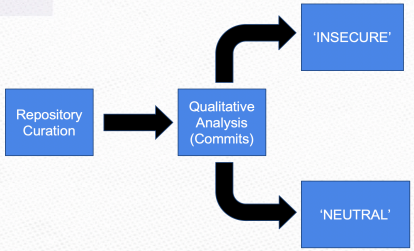
Summary

Scientific Software





Methodology



Results

Initial Repo Count	3,405,303
Criteria-1 (1% Julia files)	3,866
Criteria-2 (Available)	3,115
Criteria-3 (Not a clone)	2,173
Criteria-4 (Commits/Month ≥ 2)	2,173
Criteria-5 (Contributors ≥ 5)	253
Criteria-6 (CI)	20
Final Repo Count	20

308/7024 = 4.4%
'INSECURE'

Cohen's Kappa = 1.0

Contact Information:

Justin Murphy
jdmurphy43@tntech.edu

Shazibul Islam Shamim
mshamim42@tntech.edu

Elias Brady
etbrady42@tntech.edu

Akond Rahman
arahman@tntech.edu

References

[1] [n.d.]. The Julia Language. https://docs.julialang.org/en/v1/.

[2] Amiangshu Bosu, Jeffrey C. Carver, Munawar Hafiz, Patrick Hilley, and Derek Janni. 2014. Identifying the Characteristics of Vulnerable Code Changes: An Empirical Study (FSE 2014). Association for Computing Machinery, New York, NY, USA, 257–268. https://doi.org/10.1145/2635868.2635880

[3] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20, 1 (1960), 37–46. https://doi.org/10.1177/ 001316446002000104 arXiv:http://dx.doi.org/10.1177/001316446002000104

[4] George Thiruvathukal Jeffrey. Carver, Neil Hong. 2016. Software Engineering for Science (1st ed.). CRC Press, NY, NY, USA.

[5] Richard Landis and Gary Koch. 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33, 1 (1977), 159–174. http://www.jstor.org/ stable/2529310

[6] E. S. Mesh and J. S. Hawker. 2013. Scientific software process improvement decisions: A proposed research strategy. In 2013 5th International Workshop on Software Engineering for Computational Science and Engineering (SE-CSE). 32–39. https://doi.org/10.1109/SECSE.2013.6615097

[7] Nuthan Munaiah, Steven Kroh, Craig Cabrey, and Meiyappan Nagappan. 2017. Curating GitHub for engineered software projects. Empirical Software Engineering (2017), 1–35. https://doi.org/10.1007/s10664-017-9512-6

[8] Akond Rahman, Amritanshu Agrawal, Rahul Krishna, and Alexander Sobran. 2018. Characterizing the Influence of Continuous Integration: Empirical Results from 250+ Open Source and Proprietary Projects (SWAN 2018). ACM, New York, NY, USA, 8–14. https://doi.org/10.1145/3278142.3278149

[9] Johnny Saldaña. 2015. The coding manual for qualitative researchers. Sage