Natural Language Insights from Code Reviews that Missed a Vulnerability

Nuthan Munaiah, Benjamin S. Meyers, Cecilia O. Alm, Andrew Meneely, Pradeep K. Murukannaiah, Emily Prud'hommeaux, Josephine Wolff, and Yang Yu

Background

- Developers inevitably make mistakes.
- Practices such as code reviews prevent mistakes from becoming vulnerabilities.
- Sometimes, vulnerabilities are missed and exploited in the wild.

Code Review

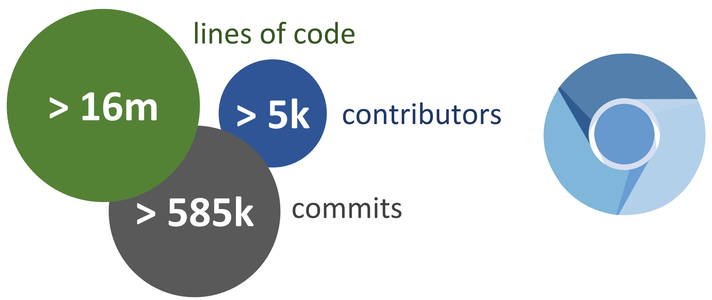
- Conversation between developers to uncover flaws.
- Natural language in code reviews contain a wealth of information: criticisms, suggestions, questions, speculation, etc.

Research Goal

Characterize the linguistic features that contribute to the likelihood that a code review has missed a vulnerability

Study Subject

The Chromium Project by Google



R·I·T | B. Thomas Golisano *College of* Computing and Information Sciences

Data Set

436,191 Bugs

1,462 Vulnerabilities

2008-16

788,437 Code Reviews

877	695,530	92,030
Fixed Vulnerability	Neutral	Missed Vulnerability

Research Questions

[Feedback Quality] Do linguistic measures of inquisitiveness, sentiment, and syntactic complexity in code reviews contribute to the likelihood that a code review has missed a vulnerability?

[Lexical Classifier] Can the words used differentiate code reviews that have missed a vulnerability?

Metrics

Feedback Quality

- ***Inquisitiveness:*** Number of questions per sentence
- ***Sentiment:*** Positivity or Negativity of messages
- ***Complexity:*** Yngve, Frazier and Propositional Density

Lexical Classifier

- *Term Frequency-Inverse Document Frequency (TF-IDF)*

Feedback Quality

- Non-parametric Mann-Whitney Wilcoxon test

Table 1: Results from Mann-Whitney Wilcoxon Test for Association between Metrics and that a Code Review Missed a Vulnerability

Metric	p-value	$Mean_{neutral}$	$Mean_{missed}$
Inquisitiveness	3.28e-12	0.1785	0.1711
Negativity	< 2.2e-16	0.3707	0.4091
Positivity	< 2.2e-16	0.0625	0.0783
Yngve	< 2.2e-16	0.0498	0.0442
Frazier	0.0031	0.8568	0.8548
P-Density	1.77e-124	0.2634	0.2708

- Code reviews that missed a vulnerability tend
  - to have ***lower*** inquisitiveness and Yngve
  - to have ***higher*** sentiment, Frazier, and p-density

Lexical Classifier

- Naïve Bayes classifier with best features

Table 2: Effectiveness of the best Naïve Bayes classifier

Precision	Recall	F_1
14%	73%	23%

- Words ***can*** differentiate code reviews that missed a vulnerability from the rest

References

N. Munaiah, B. S. Meyers, C. O. Alm, A. Meneely, P. K. Murukannaiah, E. Prud'hommeaux, J. Wolff, and Y. Yu. Natural Language Insights from Code Reviews that Missed a Vulnerability: A Large Scale Study of Chromium. In Engineering Secure Software and Systems: 9th International Symposium, ESSoS 2017, Bonn, Germany, July 3–5, 2017. Proceedings. Springer International Publishing, (to appear).