Empirical Studies

Overview

- General introduction of empirical studies in SE
- Threats to Validity
- Paper presentation and discussion
  - Secure Coding Practices in Java: Challenges and Vulnerabilities [2]

Empirical Research [1]

- Research using empirical evidence. It is a way of gaining knowledge by means of direct or indirect observation or experience.
- Empirical evidence can be analyzed quantitatively or qualitatively
- Researchers answer empirical questions, which should be clearly defined and answerable with the evidence collected

Empirical Studies in SE

- To understand how developers build or maintain software by observing various software artifacts or monitoring software runtime behaviors
- Can be conducted with manual inspection or automatic tools
- May achieve various research goals:
  - identify software change patterns
  - reveal relations between symptoms and root causes
  - shed light on new technique design and impl.

Characteristics of Empirical Studies

- Cool algorithm design or intensive programming is NOT always required
  - Sometimes only manual inspection and eyeball checking are done
- "Interesting Research Questions" is the key contribution
  - Questions haven't been asked or answered nicely
  - Questions whose answers can provide actionable advice to tool builders or users

Threats to Validity

Man prefers to believe what he prefers to be true.

-- Francis Bacon

Threats to Validity

- Is the investigator's conclusion correct?
- Try to identify the factors which make your conclusion incorrect

External & Internal Validity

- External validity
  - The degree to which the results of an empirical investigation can be generalized to and across individuals, settings, and times
    - "Is the conclusion generalizable?"
- Internal validity
  - The degree to which a causal conclusion based on a study is warranted
    - "Is the experiment done correctly?"

Threats to External Validity

- Aptitude
  - If a medicine is effective for sample patients, will it also be effective for non-volunteers or all other people?
- Situation
  - time, location, scope and extent of measurement

Threats to External Validity

- Pre-test effects
  - The cause-effect relationship can be found when pre-tests are carried out
- Post-test effects
  - The cause-effect relationship can only be found when post-tests are carried out
- …

Examples

- The empirical study is performed within a single company with particular processes, constraints, resources, and tools
- The empirical study in done on operating system software/open source projects

Threats to Internal Validity

- Confounding
  - Changes in the observation may be related to multiple variables
- Selection bias
  - Samples should be chosen without bias
- Instrument change
  - The measurement may affect the result
- John Henry effect
  - John Henry was a worker who outperformed a machine under an experimental setting because he was aware that his performance was compared with that of a machine.

Examples

- The execution time reading may significantly affect the measured execution time
- The causal-effect relationship between bugs and bad variable names may be affected by factors like complexity of functionality, maturity of developers, and types of bugs

Importance of Threat Identification

- Help researchers decide how to propose research questions and do experiments in a plausible way
- Help people understand limitation of the research
- It is OK that you can't avoid all threats. However, you should try your best to make your results representative and meaningful

Secure Coding Practices in Java: Challenges and Vulnerabilities [2]

Problem Statement

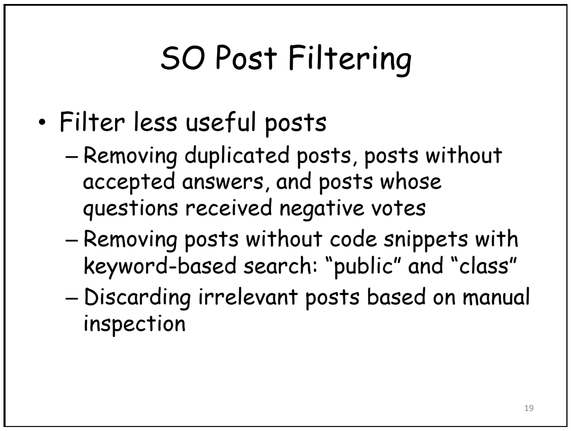
- Security software libraries facilitate secure coding
- Misusing these libraries can cost a lot of debugging effort of developers, or cause security vulnerabilities in software
- What are the biggest challenges and vulnerabilities in secure coding practice?

Research Questions

- What are the common concerns on Java secure coding?
- What are the common programming challenges?
- What are the common security vulnerabilities?

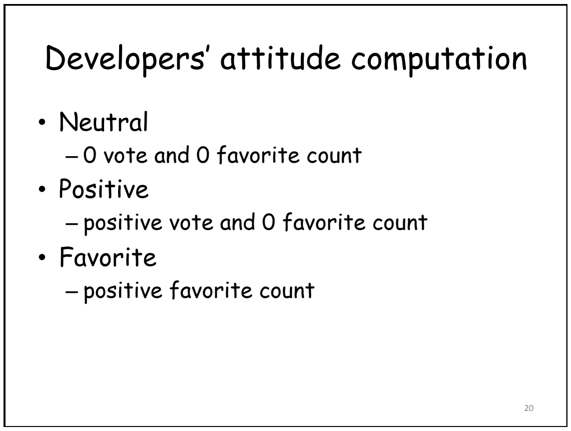
Methodology

- 22,195 StackOverflow (SO) posts containing keywords "Java" and "security"
- Mainly focus on 503 posts for manual inspection after filtering the posts
  - Initially classify posts based on the software libraries under discussion
  - Further refine the classification based on the security concerns, e.g., cryptography, authentication



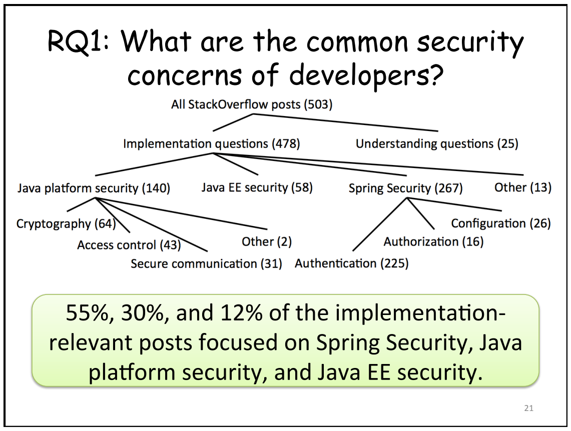
SO Post Filtering

- Filter less useful posts
  - Removing duplicated posts, posts without accepted answers, and posts whose questions received negative votes
  - Removing posts without code snippets with keyword-based search: "public" and "class"
  - Discarding irrelevant posts based on manual inspection



Developers' attitude computation

- Neutral
  - 0 vote and 0 favorite count
- Positive
  - positive vote and 0 favorite count
- Favorite
  - positive favorite count



RQ1: What are the common security concerns of developers?

All StackOverflow posts (503)

Implementation questions (478)

Understanding questions (25)

Java platform security (140)

Java EE security (58)

Spring Security (267)

Other (13)

Cryptography (64)

Access control (43)

Secure communication (31)

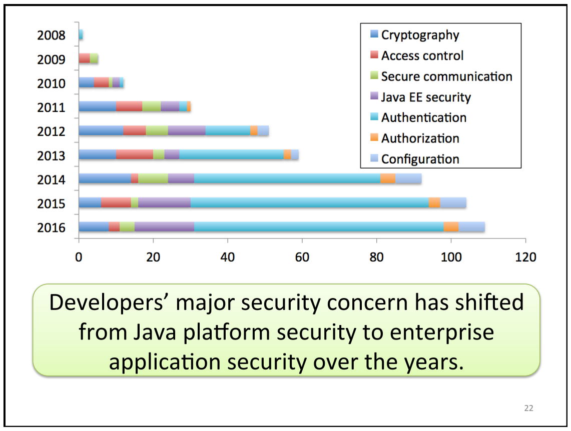
Other (2)

Authentication (225)

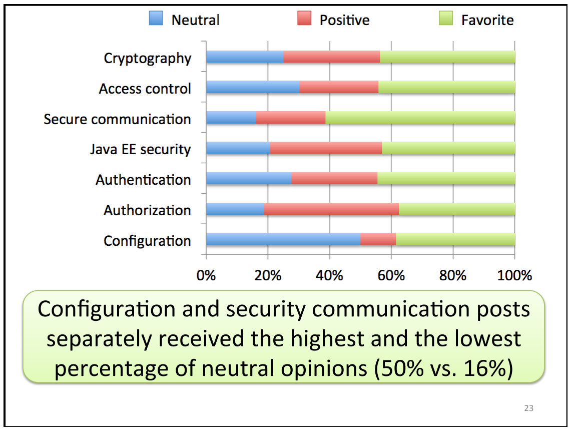
Authorization (16)

Configuration (26)

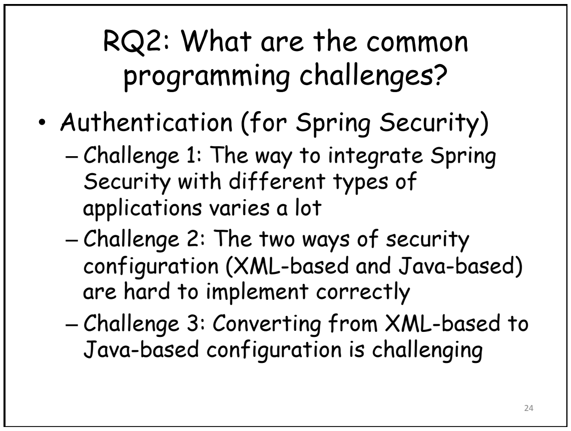
55%, 30%, and 12% of the implementation-relevant posts focused on Spring Security, Java platform security, and Java EE security.



Developers' major security concern has shifted from Java platform security to enterprise application security over the years.



Configuration and security communication posts separately received the highest and the lowest percentage of neutral opinions (50% vs. 16%)



RQ2: What are the common programming challenges?

- Authentication (for Spring Security)
  - Challenge 1: The way to integrate Spring Security with different types of applications varies a lot
  - Challenge 2: The two ways of security configuration (XML-based and Java-based) are hard to implement correctly
  - Challenge 3: Converting from XML-based to Java-based configuration is challenging

RQ2 (cont'd)

- Cryptography
  - Challenge 1: The error message did not provide sufficient useful hints about fixes
  - Challenge 2: It is difficult to implement security with multiple programming languages
  - Challenge 3: Implicit constraints on API usage cause confusion

```
//privKey should be in PKCS#8 format
byte[] privKey = ...;
PKCS8EncodedKeySpec keySpec=
  new PKCS8EncodedKeySpec(privKey);
```

25

RQ2 (cont'd)

- Java EE security
  - These posts were mainly about authentication and authorization. One challenge is the complex usage of declarative security and programmatic security, and any complicated interaction between the two

26

RQ2 (cont'd)

- Access Control
  - Challenge 1: The effect of access control varies with the program context
  - Challenge 2: The effect of access control varies with the execution environment

27

RQ2 (cont'd)

- Secure Communication
  - These posts mainly discussed the process of establishing SSL/TLS connections. This process contains so many steps that developers were tempted to accept a broken solution to simply bypass the security check

28

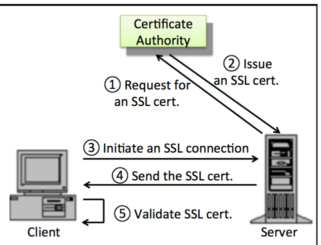
RQ3: What are the common security vulnerabilities?

- Spring Security's csrf()
  - Cross-site request forgery (CSRF) is a serious attack that tricks a web browser into executing an unwanted action in a web application for which a user is authenticated
  - Some developers took the suggestion to irresponsibly disable the default CSRF protection. Developers are unaware of the security consequences of their insecure coding

29

RQ3 (cont'd)

- SSL/TLS



9 of 11 SSL/TLS-relevant posts discussed insecure code to bypass security checks. StackOverflow contains a lot of obsolete and insecure coding practices, and secure programmers are unaware of the state-of-the-art security knowledge.

30

RQ3 (cont'd)

- Password Hashing
  - Six posts were related to hashing passwords with MD5 or SHA-1 to store the user credentials in a database
  - Three of these posts accepted vulnerable solutions as correct answers, indicating that developers were unaware of the best practice of secure programming

Reference

[1] Empirical research, https://en.wikipedia.org/wiki/Empirical_research

[2] Na Meng, Stefan Nagy, Daphne Yao, Wenjie Zhuang, and Gustavo Arango Argoty, Secure Coding Practices in Java: Challenges and Vulnerabilities, ICSE 2018