Problem Statement

• Stack Overflow (SO) is a rich source of information for software developers
• Anecdotes report that software developers copy and paste code snippets from SO for convenience
• Such behaviors constantly introduce community-provided code to production software, although the impact on code security is unknown

Contributions

• Identified all Android posts on SO, extracted all security-related code snippets, and analyzed their security using a robust ML-based approach
• Applied clone detection techniques to detect the extracted code snippets in Android apps

Contributions (cont’d)

• 15.4% of all 1.3M Android apps contained security-related code snippets, out of which 97.9% contain at least one insecure code snippet

The Approach

Extraction & Filtering

• A code snippet is considered security-related iff it calls one of the following APIs:
  – Cryptography: JCA, JCE
  – Security network communication: JSSE, ...
  – Public key infrastructure: X.509, ...
  – Authentication and access control, ...
  – ...

Stack Overflow Considered Harmful? The Impact of Copy&Paste on Android Application Security
Extraction & Filtering (cont’d)

• Extracted code from SO posts based on <code> tags
• Leveraged JavaBaker to infer the qualified names of invoked APIs, and then decide whether a snippet is security-related
• May have false positives

Classification

• Secure code
  – Snippets that contain up-to-date and strong algorithms (adhere to best practice)
  – Snippets that contain code that does not result in easily exploitable vulnerabilities
  – Snippets that contain code whose security depended on additional user input
• Insecure code
  – Snippets with insecure code, e.g., using outdated algorithms

Classification (cont’d)

• Rule-based data labeling for 1360 snippets, which were extracted from answers
• Leveraged the labeled data as training set to train an ML model with SVM
  – The feature vector is a tf-idf vectorizer
  – Automatically classified the rest of the snippets

Clone Detection in Android Apps

• Converted Android apk files to Java bytecode with enjarify, and then converted bytecode to WALA’s IR
• Converted code snippets from SO to JDT with PPA, and then converted JDT to WALA’s IR
• Created PDGs for each method, and detect clones based on the PDG matching

Evaluation

• Code extraction and filtering
  – 818,572 question threads tagged with “android”
  – 2,474 snippets from question posts and 1,360 distinct snippets from answer posts

Evaluation (cont’d)

• Code classification
  – Trained the classifier with 1,360 snippets, and tested the classifier with the complete set of 3,834 snippets
  – 5-fold cross validation on 1,360 snippets, obtaining 0.904 accuracy and 0.903 precision
  – Among the 3,834 snippets, 1,161 were labeled insecure, and 2,474 were labeled secure
Evaluation (cont’d)

• Clone Detection
  – 15.4% (200,372) apps included clones of the analyzed snippets
  – 15% apps included at least one insecure code snippet