An Investigation of Online Reverse Engineering Community Discussions in the Context of Ghidra

Daniel Votipka, Mary Nicole Punzalan, Seth M. Rabin, Yla Tausczik, and Michelle L. Mazurek
SOFTWARE REVERSE ENGINEERING
SOFTWARE REVERSE ENGINEERING

Looking at someone else’s code to figure out how it works!
SOFTWARE REVERSE ENGINEERING

Looking at someone else’s code to figure out how it works!

Vulnerability Discovery
SOFTWARE REVERSE ENGINEERING

Looking at someone else’s code to figure out how it works!

Vulnerability Discovery

Malware Analysis
SOFTWARE REVERSE ENGINEERING

Looking at someone else’s code to figure out how it works!

Vulnerability Discovery

Malware Analysis

30 students and RE professionals took ~39 minutes on average to reverse engineer decompiled code snippets with <150 lines

Yakdan et al. 2016
TOOL SUPPORT
TOOL SUPPORT

Many sophisticated tools
Many sophisticated tools
- Limited adoption

Hafiz and Fang 2015
Votipka et al. 2018
RESEARCH QUESTIONS
RESEARCH QUESTIONS

Identify user needs
RESEARCH QUESTIONS

Identify user needs

RQ1: What features do REs discuss most?
RESEARCH QUESTIONS

Identify user needs
RQ1: What features do REs discuss most?

Understand community dynamics
RESEARCH QUESTIONS

Identify user needs
RQ1: What features do REs discuss most?

Understand community dynamics
RQ2: How is knowledge shared and developed?
RESEARCH QUESTIONS

Identify user needs
RQ1: What features do REs discuss most?

Understand community dynamics
RQ2: How is knowledge shared and developed?

RQ3: How does the forum used impact community behavior?
RESEARCH QUESTIONS

Identify user needs
RQ1: What features do REs discuss most?

Understand community dynamics
RQ2: How is knowledge shared and developed?

RQ3: How does the forum used impact community behavior?
GHIDRA

NSA publicly released Ghidra
GHIDRA

NSA publicly released Ghidra
- Fully-featured RE framework
GHIDRA

NSA publicly released Ghidra
- Fully-featured RE framework
- Significant attention at security conferences and in the press
DATA COLLECTION AND CLEANING
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement

Twitter, StackExchange
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement - Only included direct discussions of Ghidra features
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement - Only included direct discussions of Ghidra features
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement - Only include direct discussions of Ghidra features.
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement

- Only included direct discussions of Ghidra features
DATA COLLECTION AND CLEANING

All Ghidra posts for the 6 months after announcement

- Only included direct discussions of Ghidra features
QUALITATIVE ANALYSIS

Two independent coders using existing codebooks

Features
- Codebook based on issue labels in Ghidra Github repo

Conversational Act
- Main reason for the thread
- Type of discourse act for each comment

Collective Sensemaking
- Community members share information and build knowledge together
RESEARCH QUESTIONS

Identify expressed user needs
RQ1: What features do REs discuss most?

Understand community dynamics
RQ2: How is knowledge shared and developed?

RQ3: How does the forum used impact community behavior?
FEATURE FINDINGS
FEATURE FINDINGS

Customization is most common (43%)
FEATURE FINDINGS

Customization is most common (43%)
- Scripts (25%)
FEATURE FINDINGS

Customization is most common (43%)
- Scripts (25%)
- New Architectures (12%)
FEATURE FINDINGS

Customization is most common (43%)
- Scripts (25%)
- New Architectures (12%)

Decompiler was the most discussed current tool (5%)
RESEARCH QUESTIONS

Identify expressed user needs
RQ1: What features do REs discuss most?

Understand community dynamics
RQ2: How is knowledge shared and developed?

RQ3: How does the forum used impact community behavior?
Most questions were answered.
Most questions were answered
CLUSTERED USERS

Twitter
StackExchange
Reddit
CLUSTERED USERS

Central cluster and mostly isolated threads

Twitter

StackExchange

Reddit
CLUSTERED USERS

Central cluster and mostly isolated threads

Twitter

StackExchange

Reddit

Active responder threads included more sensemaking
RECOMMENDATIONS

Tool Developers
- Consider customization early
- Answer isolated user questions

RE Community
- Good information on all forums
- Incentivize answers for less active member
SUMMARY

Feature Discussions
- Customization is most common

Community Dynamics
- Most questions are answered
- Active responders more likely to have questions answered
- Central clusters of connected users; Others isolated

Recommendations
- Consider customization early
- Answer isolated user questions
- Consider all forums

Questions:
@drvotipka
dvotipka@cs.tufts.edu
sec-professionals.cs.umd.edu