Estimating Types in Binaries using Predictive Modeling

Omer Katz Ran El-Yaniv Eran Yahav

Problem: Calls to virtual functions break the control flow of programs and hinder reverse engineering (RE) efforts

Goal: Statically determine most likely targets of each virtual function call

Solution: Determine types of objects used in virtual function calls based on how they are used

1st stage – Binary Analysis and Tracelet Extraction

- **Input:**
  - Many possible targets

- **Output:**
  - Focus on most likely targets

- **Steps:**
  1. **Input:** Call vfunc
  2. **1st stage – Binary Analysis and Tracelet Extraction**
     - **Input:**
       - Many possible targets
     - **2nd stage – Train a model for each type**
       - **Input:**
         - Use objects for which a type can be statically determined (e.g. recently initialized)
       - **N-gram based Statistical Language Models**
         - Computes sentence probability in a language $P_{\text{English}}(\text{I am giving a talk}) > P_{\text{English}}(\text{A talk giving I am})$
         - Sentences of length $n$
         - Uses Prediction by Partial Match implementation
           - Originally a compression algorithm
       - **Train model by parsing to $n$-grams and building probability tree**
       - **Query model by traversing path to sentence**
     - **3rd stage – Compute ranking of types for objects**
       - **Input:**
         - Use objects for which a type can’t be statically determined
       - **Output:**
         - Focus on most likely targets

- **Evaluation:**
  - Evaluated over 20 benchmarks
  - Compared to ground truth from manual RE
  - **Objective:** rank expected target highest
  - Across all benchmarks, for over 80% of calls to virtual functions, expected target ranked in top 3

- **Benchmark results:**
  - **X axis – maximum rank**
  - **Y axis – percentage of calls to virtual functions**
    - For which the expected target was ranked below the maximum rank

The research leading to these results has received funding from the European Union's - Seventh Framework Programme (FP7) under grant agreement n° 615688 – ERC-COG-PRIME.