A Comparison of Static, Dynamic, and Hybrid Analysis for Malware Detection

LyuJiuyang, Dec 23, 2021

Info

- **Title:** A Comparison of Static, Dynamic, and Hybrid Analysis for Malware Detection.
- **Author:** Anusha Damodaran, Fabio Di Troia, Corrado Aaron Visaggio, Thomas H. Austin & Mark Stamp.
- **Journal:** J. Comput. Virol. Hacking Tech. 2017.

Introduction

Difference between Static and Dynamic detection methods: if needed to execute the software.

Signature Based Detection

Core: A database of signatures of malware.

Pros: simple, relatively fast and effective

Cons: need an up-to-date database, can be simply evade.

Behavior Based Detection

Classify a software while focusing on the actions performed by the malware during execution.

Statistical Based Detection

based on statistical properties derived from program features.

Hidden Markov Models

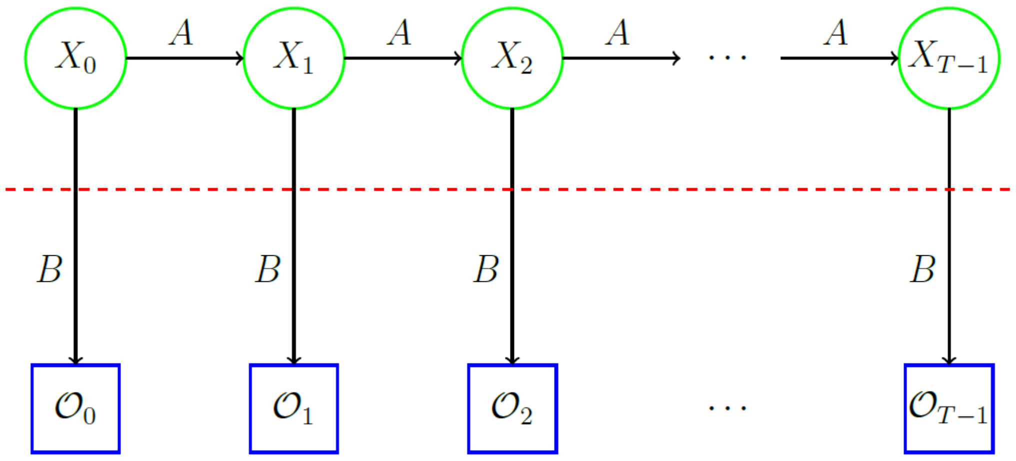

Figure 1: Generic Hidden Markov Model

$$
\begin{aligned}
T &= \text{ length of the observation sequence} \\
N &= \text{ number of states in the model} \\
M &= \text{ number of observation symbols} \\
Q &= \{q_0, q_1, \ldots, q_{N-1}\} = \text{ distinct states of the Markov process} \\
V &= \{0, 1, \ldots, M-1\} = \text{ set of possible observations} \\
A &= \text{ state transition probabilities} \\
B &= \text{ observation probability matrix} \\
\pi &= \text{ initial state distribution} \\
\mathcal{O} &= (\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1}) = \text{ observation sequence.}
\end{aligned}
$$

I don't want to write much about this part, this article simply applies HMM to malware detection. The **Problems** to solve are also classic. It is very likely that LSTM will achieve better results.

Related Work

Static Analysis

Based on opcode sequences, control flow graphs, function call graph, etc. Using some techniques or ML such as SVM+PCA.

Dynamic Analysis

Based on API calls, system calls, instruction traces, registry changes, memory writes, and so on.

Hybrid Approaches

Seems can be obtain more features.

Experiments

- Tools: IDA Pro, Buster Sandbox Analyzer, Ether.
- Datasets: Harebot, Security Shield, Smart HDD, Winwebsec, Zbot, ZeroAccess
- Metrics: AUC

Training/Scoring with specified analysis.

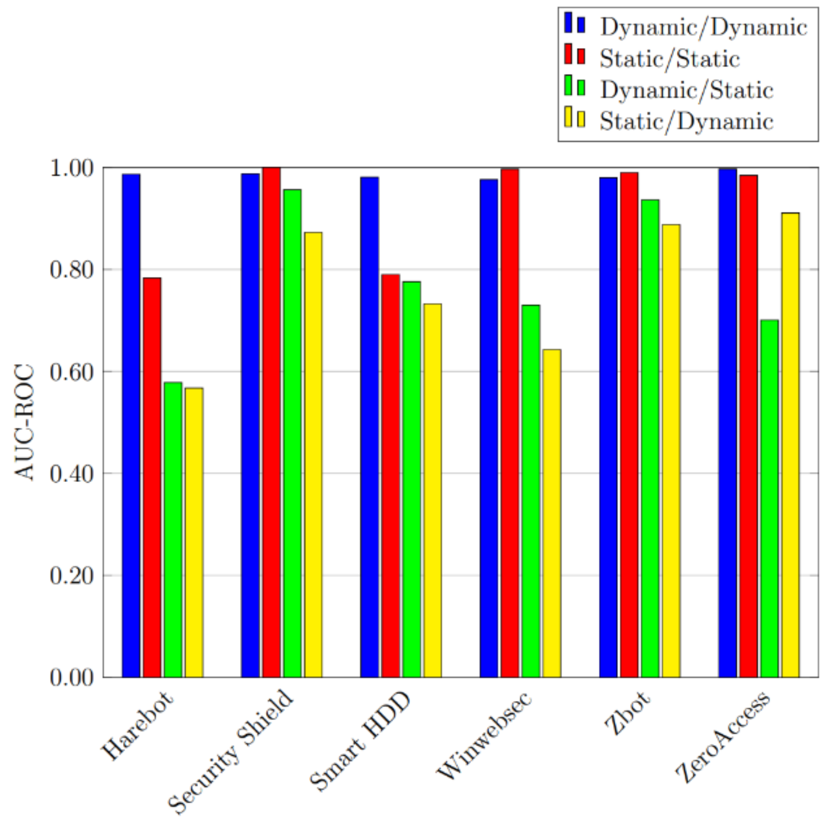

Figure 4: ROC Results for API Call Sequence

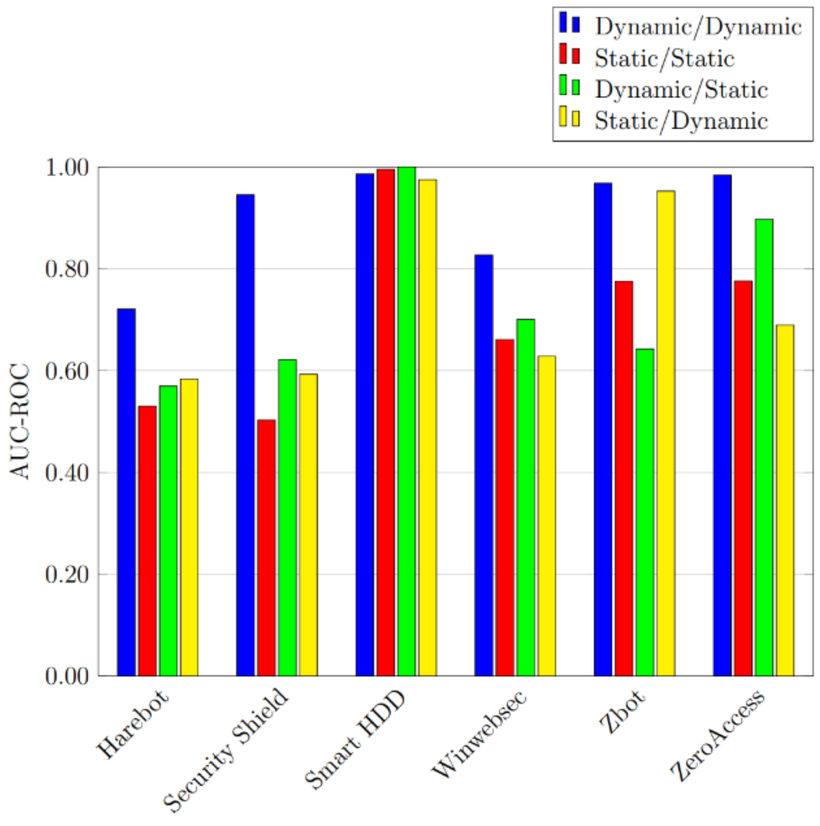

Figure 7: ROC Results for Opcode Sequences

A straightforward hybrid approach is unlikely to be superior to fully dynamic detection. Use hybrid with caution.

Other

- Imbalance Problem: In Recommender System, negative sampling is a good way to solve it. But the scale of these datasets are too small.
- It's very likely that using LSTM will achieve better performance.
- As authors mentioned, more scoring techniques can be used.