Dynamic Opcode Analysis of Ransomware

Dr. Domhnall Carlin
Research Fellow
d.carlin@qub.ac.uk @dqmhnallcarlin
$ whoami

- Domhnall Carlin
- Research Fellow, Centre for Secure Information Technologies, QUB
- PhD Computer Science: “Dynamic Analyses of Malware”.
  - Successfully defended April 2018
  - Research presented today is an extension of thesis
Ransomware

• Perfect example of the shift in malware generally
• From disaffected hobbyists to criminal gangs
• ‘Perfect storm’ of anonymous payment, encryption, anonymous internet and connected devices
• McAfee [1] state that 1.5 million new ransomware samples were found in Q3 of 2017, a rise of 36% on the previous quarter.

Background

Signature Detection

• Most widely utilised approach within commercial malware-detection software (Santos et al, 2009).
• New malware instances must be captured, analysed for a signature, stored and deployed.
• Obfuscation techniques compound this issue.
Background

• **Opcodes** *(OPeration CODE)* are the segments of assembly language that specify the operation to be performed on an operand.

• Dynamic analysis allows obfuscated malware to reveal itself at run time.

<table>
<thead>
<tr>
<th>Module</th>
<th>Address</th>
<th>Command</th>
<th>Modified registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>notepad</td>
<td>00095402</td>
<td>CALL notepad.00956000</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00095406</td>
<td>MOV EDI,EBX</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00095560</td>
<td>MOV ES:F</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00095560</td>
<td>MOV EDI,EBX</td>
<td>EDF=96207754</td>
</tr>
<tr>
<td>notepad</td>
<td>00095600</td>
<td>MOV ESI</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00095600</td>
<td>MOV ESI</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00095600</td>
<td>MOV ESI</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00099204</td>
<td>MOV ESI</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00099204</td>
<td>MOV ESI</td>
<td></td>
</tr>
<tr>
<td>notepad</td>
<td>00099204</td>
<td>MOV ESI</td>
<td></td>
</tr>
</tbody>
</table>

However...

• The datasets used can be small, badly sampled and need expanded to match the >20k samples used in other static research.

• Virtualization may be detected by modern malware.
Dynamic Opcode Analysis of Malware

Main Contributions

• Applying machine learning techniques to the largest dataset of its kind, both in terms of breadth (610-100k features) and depth (48k samples)

• >99% detection accuracy across 48,000 samples, using only the first 32k opcodes

• \(N=1\) is best Feature reduction techniques investigated, allowing feature set to be reduced from 610 to 50 opcodes

• This demonstrates that a dynamic opcode analysis approach can compare with static analysis in terms of speed

Motivation

Develop a strategy for the detection of malware, which is immune to modern obfuscation methods, and is applicable at the hypervisor level. I.e Detect more, faster and with less information.

1) Can opcode counts extracted from runtraces offer accurate ransomware detection over a large sample size?

2) Does a 32k opcode run-length offer superior accuracy over full-length traces?

3) Can the method(s) which successfully detect(s) ransomware behaviour, differentiate these from benign encryption (e.g. file zipping) behaviours?
Dynamic Analysis of Ransomware

• Source data VirusShare.com
• 21,378 PE .exe crypto-ransomware samples
• 3,591 benign files from Windows machines, with the SMOTE minority oversampling technique applied
• 1,000 zipping traces from 7Zip with SMOTE applied
• Used RandomForest classifier, implemented in WEKA 3.9

Sample processing

Address | Thread | Command
---|---|---
00450E11 | Main | MOV ESI,Trojan_W.00433000
00450E16 | Main | LEA EDI,DWORD PTR DS:[ESI+FFFCE000]
00450E1C | Main | MOV DWORD PTR DS:[EDI+404CC],54200703
00450E26 | Main | PUSH EDI
00450E27 | Main | OR EBP,FFFFFFFF
00450E2A | Main | JMP SHORT Trojan_W.00450E3A
00450E3A | Main | MOV EBX,DWORD PTR DS:[ESI]

Sample execution
## Results: Benignware Vs Ransomware

**MACHINE LEARNING METRICS FOR 32K RUN-LENGTH BENIGN AND, RANSOMWARE TRACES**

<table>
<thead>
<tr>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
<th>MCC</th>
<th>ROC</th>
<th>PRC</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.932</td>
<td>0.001</td>
<td>0.982</td>
<td>0.932</td>
<td>0.957</td>
<td>0.955</td>
<td>0.996</td>
<td>0.982</td>
<td>Benign</td>
</tr>
<tr>
<td>0.999</td>
<td>0.068</td>
<td>0.996</td>
<td>0.999</td>
<td>0.998</td>
<td>0.955</td>
<td>0.996</td>
<td>1</td>
<td>Ransomware</td>
</tr>
<tr>
<td>0.996</td>
<td>0.064</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.955</td>
<td>0.996</td>
<td>0.999</td>
<td>Average</td>
</tr>
</tbody>
</table>
## Results: Benignware vs Ransomware vs Zipping

**Machine learning metrics for 32k run-length benign, ransomware, and benign encryption traces**

<table>
<thead>
<tr>
<th>TP</th>
<th>FP</th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
<th>MCC</th>
<th>ROC</th>
<th>PRC</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.932</td>
<td>0.001</td>
<td>0.98</td>
<td>0.932</td>
<td>0.956</td>
<td>0.954</td>
<td>0.996</td>
<td>0.981</td>
<td>Benign</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Zipper</td>
</tr>
<tr>
<td>0.999</td>
<td>0.035</td>
<td>0.996</td>
<td>0.999</td>
<td>0.998</td>
<td>0.975</td>
<td>0.998</td>
<td>1</td>
<td>Ransomware</td>
</tr>
<tr>
<td>0.996</td>
<td>0.032</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.975</td>
<td>0.998</td>
<td>0.999</td>
<td>Average</td>
</tr>
</tbody>
</table>
TL;DR
Dynamic opcode analysis can:
• detect malicious PE software
• even when obfuscated
• with low computational expense
• in short run-times
• with high accuracy

Run-trace dataset of ransomware available.