Dynamic Opcode Analysis of Ransomware

Dr. Domhnall Carlin

Research Fellow

d.carlin@qub.ac.uk @domhnallcarlin

$ 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.

[1] “McAfee Labs Threats Report December 2017,” McAfee, Tech. Rep., 2017. [Online]. Available:https://www.mcafee.com/us/resources/reports/rp-quarterly-threats-dec-2017.pdf

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

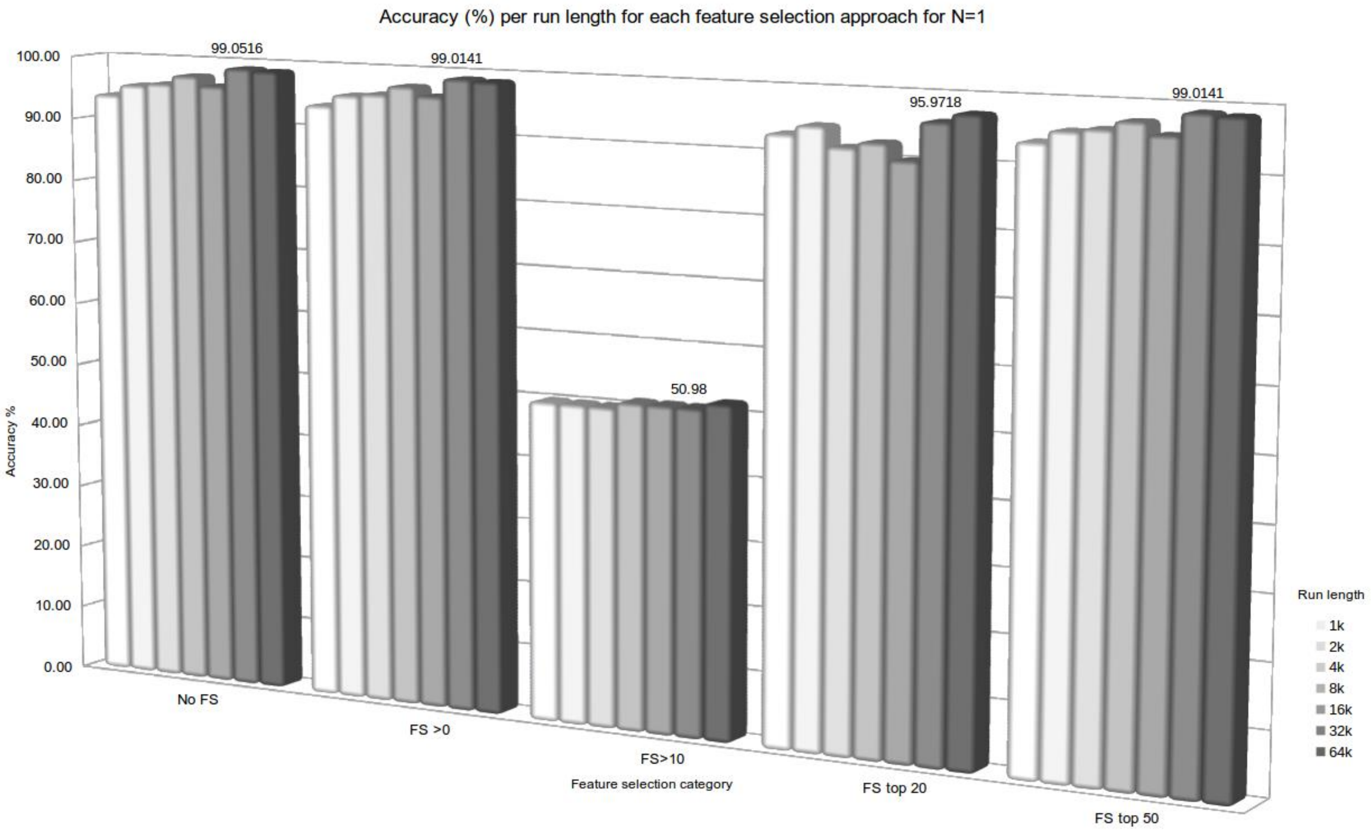
- **Opcodes** (*OPeration CODE*) are the segments of assembly language that specify the operation to be performed on an operand.
- Dynamic analysis allows obfuscated malware to reveal itself at run time

Module	Address	Command	Modified registers
notepad	00B964B2	CALL notepad.00B96368	
notepad	00B96368	MOV EDI,EDI	
notepad	00B9636A	PUSH EBP	
notepad	00B9636B	MOV EBP,ESP	EBP=002BFF54
notepad	00B9636D	SUB ESP,0x14	
notepad	00B96370	AND DWORD PTR SS:[EBP-0xC],0x0	
notepad	00B96374	AND DWORD PTR SS:[EBP-0x8],0x0	
notepad	00B96378	MOV EAX,DWORD PTR DS:[0xBA7390]	EAX=081C7BBD
notepad	00B9637D	PUSH ESI	
notepad	00B9637E	PUSH EDI	
notepad	00B9637F	MOV EDI,0xBB40E64E	EDI=BB40E64E
notepad	00B96384	MOV ESI,0xFFFF0000	ESI=FFFF0000
notepad	00B96389	CMP EAX,EDI	
notepad	00B9638B	JE notepad.00BA21DD	
notepad	00B96391	TEST ESI,EAX	

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

Carlin, D., O'Kane, P. & Sezer, S. 25 Jan 2017 Dynamic Analysis of Malware using Run Time Opcodes. Chapter in: Data Analytics and Decision Support for Cybersecurity - Trends, Methodologies and Applications. Springer.

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.**

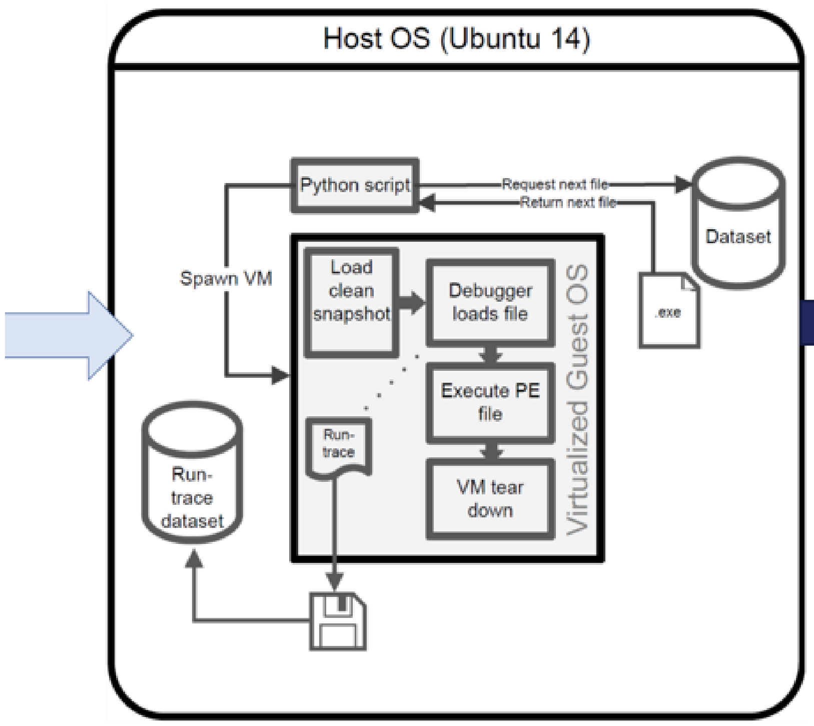
1) Can opcode counts extracted from runtraces offer accurate ransomware detection over a large sample size?
2) Does a 32k opcode run-length offer superior accuracy over full-length traces?
3) Can the method(s) which successfully detect(s) ransomware behaviour, differentiate these from benign encryption (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

- N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial, special issue on learning from imbalanced data sets," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, p. 1, Jun 2004.

Sample processing



Sample execution

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]

XOR	CALL	PUSH	POP	POR	SUB	RETN	CP	CMP	CPUID	CMPB	ADC	ADCL
10622	11559	52471	21600	0	4535	7303	0	43260	0	0	3	
9841	19011	64099	25092	0	5826	10697	0	35774	0	0	0	
8256	15538	51302	20707	0	4860	8444	0	30113	0	0	0	
164	325	1130	423	0	104	183	0	609	0	0	0	
4635	11710	37739	18407	0	3746	8528	0	7587	0	0	638	
13845	24347	80291	33686	0	7431	12648	0	60746	0	0	2	

Results: Benignware Vs Ransomware

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

TP	FP	Precision	Recall	F	MCC	ROC	PRC	Class
0.932	0.001	0.982	0.932	0.957	0.955	0.996	0.982	Benign
0.999	0.068	0.996	0.999	0.998	0.955	0.996	1	Ransomware
0.996	0.064	0.996	0.996	0.996	0.955	0.996	0.999	Average

Results: Benignware vs Ransomware vs Zipping

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

TP	FP	Precision	Recall	F	MCC	ROC	PRC	Class
0.932	0.001	0.98	0.932	0.956	0.954	0.996	0.981	Benign
1	0	1	1	1	1	1	1	Zipper
0.999	0.035	0.996	0.999	0.998	0.975	0.998	1	Ransomware
0.996	0.032	0.996	0.996	0.996	0.975	0.998	0.999	Average

CENTRE FOR SECURE INFORMATION TECHNOLOGIES

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.

Domhnall Carlin : d.carlin@qub.ac.uk