



**QUEEN'S
UNIVERSITY
BELFAST**

Dynamic Runtime Opcode Dataset Generation for Improving Malware Classification

Carlin, D., O'Kane, P., & Sezer, S. (2016). *Dynamic Runtime Opcode Dataset Generation for Improving Malware Classification*. Poster session presented at 19th International Symposium on Research in Attacks, Intrusions and Defenses, Evry, France.

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights
© 2016 The Authors.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: <http://go.qub.ac.uk/oa-feedback>

Domhnall Carlin, Dr. P. O’Kane & Prof. S. Sezer

Centre for Secure Information Technologies, Queen’s University, Belfast, N. Ireland.

Research Objective

Develop a strategy for the detection of malware, which is immune to modern obfuscation methods, and which is applicable at the hypervisor level.

Motivation

Signature-detection is the most widely used approach in commercial malware detection. New malware must be captured & analysed for a signature, which is deployed to users.

Obfuscation techniques compound the issue- 1000s of variations generated per day.

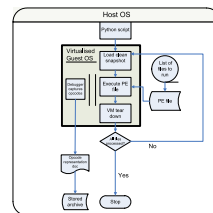
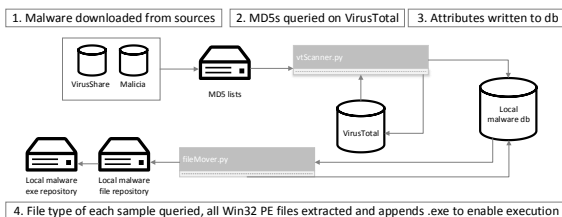
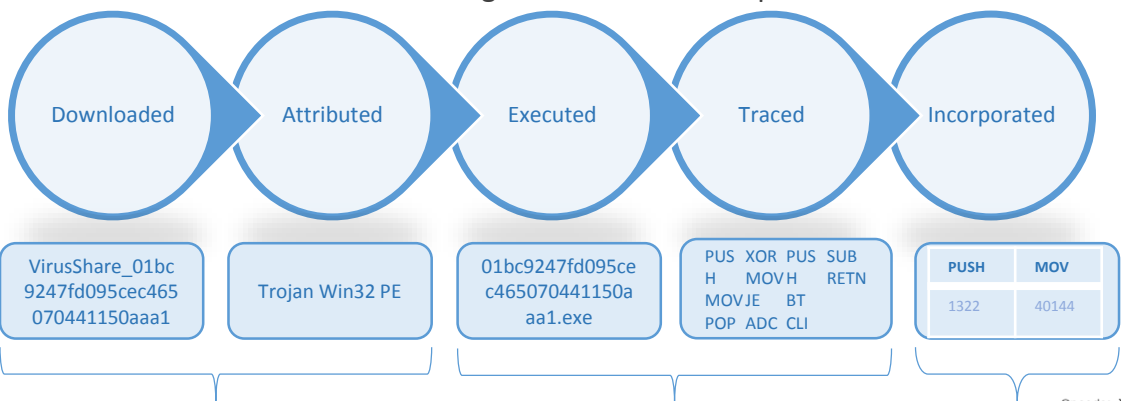
By definition, we are constantly behind the curve in the arms race against malware.

Relevant Research

- Analysis of opcodes (assembly instructions) can detect malware, bypassing issues inherent in signature-based methods (i.e. the instance can be novel).
- Packed or encrypted malware cannot generally be investigated statically. Dynamic analysis allows the obfuscated malware to reveal itself.
- Datasets in the literature have been small, poorly sampled and prone to class imbalances.

Creation of a large dynamically-generated run-trace dataset

Processing each malicious sample



Opcodes →

	JMP	CALL	PUSH	POP	POB	POB	SUB	RETN	CP	CMF	CRUD	C
10622	11559	52471	21600	0	0	0	4355	7903	0	43260	0	0
9841	19011	64099	25092	0	0	0	5826	10687	0	35714	0	0
8794	15930	13162	30787	0	0	0	4802	8464	0	30213	0	0
144	125	1130	423	0	0	0	104	183	0	809	0	0
4655	11110	8719	18487	0	0	0	3748	4828	0	7947	0	0
11845	24847	85291	33686	0	0	0	7431	12648	0	60746	0	0
61747	20564	37884	149526	0	0	0	31043	43887	0	212269	0	0
15256	15254	74211	31141	0	0	0	7882	10454	0	67411	0	0
6098	10800	36690	14200	0	0	0	3463	6205	0	22912	0	0
9209	9118	33611	13096	0	0	0	1342	2612	0	24839	0	0
6281	11071	37849	14521	0	0	0	3554	6422	0	18430	0	0
1495	12907	34607	11121	0	0	0	4617	7142	0	38852	0	0
4139	7305	27910	10888	0	0	0	2791	3674	0	20048	0	0
37	157	254	21	0	0	0	20	11	0	57	0	0
4605	9209	31281	12681	0	0	0	2398	4543	0	18406	0	0
4461	7487	2612	10487	0	0	0	1648	4000	0	19611	0	0
2846	3757	11884	4753	0	0	0	1064	2120	0	9610	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
7616	13462	50189	21513	0	0	0	4108	6487	0	49788	0	0
10230	20329	104413	43448	0	0	0	8512	13429	0	79415	0	0
1398	10571	51046	21369	0	0	0	4120	7014	0	37043	0	0
684	1292	48375	17572	0	0	0	4062	7179	0	27292	0	0
668	17556	6286	23953	0	0	0	5706	10413	0	33765	0	0

Future work facilitated by new dataset:

Dataset stands at 48,252 executed, attributed and traced malware samples.

Unsupervised Deep Learning
Neural Net Classification
Relabelling based on opcode feature clusters

Optimal dataset composition
Algorithm comparisons
Novel feature selection & reduction

One-class learning
Rule-based classifiers
Dynamically uninvestigated malware types