**Malware Categorization using Machine Learning**

A dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science

In

**The Queen’s University of Belfast**

By

**Thomas Pickup**

# Declaration

Declaration

# Acknowledgements

Acknowledgements……

# Abstract

An important factor in risk assessment is categorisation of malware and its behaviour. It should be noted, a high number of new malware types does not necessarily imply high risk, as malware such as adware does not constitute a high risk. However, a low number of new signature variants does not indicate a low risk, as the new malware signature may relate to a rootkit. Malware programs are often categorised based on Propagation, infection mechanism, Self-Defense (concealment/evasion) or Payload (Criminal Software functionality).

When malware is correctly categorised, it enables an assessment of the risk associated with particular types of malware attacks, thereby enabling Security Operation Centers (SOC) to focus on the highest current threat. Many SOCs have adapted malware categorisation according to type, family and strain is a difficult task and may be impossible to achieve fully. The result is that 66 different AV scanners (VirusTotal) often produce different results, adding to the confusion and impact the ability to assess malware attacks. Therefore this investigates new methods of malware classification that will improve the ability to determine risk assessment of malware. A dynamic runtime dataset (PE file execution) will be mined using unsupervised/clustering algorithms to identify new methods of malware categorisation based on API call structure, which hopefully provides insight to malware risk assessment.

The project will involve:

* Study current publications about dynamic malware analysis techniques
* Establish a run-time environment that can be used to create a programme execution trace dataset (such as cuckoo)
* Write a parser to extract features from the dataset. A literature review is required to determine those features that may yield the best machine learning features.
* Use machine learning clustering algorithms to categorise malware into a cluster that correlates its: risk, family, structure, etc.
* The data mining should be repeated for multiple malware family/categories to determine the optimal category definition.
* Develop and implement an algorithm for measuring agreement/different between existing labels and the new label sets (novel labelling).

# Contents

[Declaration 2](#_Toc529181536)

[Acknowledgements 3](#_Toc529181537)

[Abstract 4](#_Toc529181538)

[Contents 5](#_Toc529181539)

[Introduction and Problem Area 6](#_Toc529181540)

[Solution Description and System Requirements 6](#_Toc529181541)

[Design 7](#_Toc529181542)

[Implementation 7](#_Toc529181543)

[Testing 8](#_Toc529181544)

[System Evaluation and Experimental Results 8](#_Toc529181545)

[Appendices **Error! Bookmark not defined.**](#_Toc529181546)

[References 9](#_Toc529181547)

[Appendices 10](#_Toc529181548)

# Introduction and Problem Area

Malware: any malicious program or code that is harmful to computers.

With an expected 20 Billion Internet connected devices to come online by 2020 (Forbes 2018), protection against and identification of Malware is becoming more important by the day. The average cost of a malware attack on a company is $2.4million and the cost in time of a malware attack is 50 days (Varonis 2018).

It’s clear that Malware is becoming even more varied, with the NHS Ransomware attack spanning off multiple clones of the WannaCry virus, G Data Software estimates that in 2017 alone there was 7.41 million new malware specimens (G Data Software 2017).

The cost and rate at which malware is growing, makes this field one of the most important in the Computer Science industry. Current methods of analysing malware are not proving useful for this ever changing field. Static Analysis is prevalent, but the major downside is that it doesn’t protect against zero day attacks or new strains. Even polymorphic code can fool malware detectors that rely upon Static Analysis.

Machine Learning algorithms using dynamic analysis provide a viable alternative to this limitation, by basing their result on the behaviour of the specimen, the model theoretically can predict not only whether it is benign or malicious but could also be used to classify what family of malware the specimen belongs to.

The goal of this report it to provide an in depth look into how we could use machine learning in the future to classify malware. The report will look into different methods of malware analysis techniques, it will then document the process of setting up a Cuckoo Sandbox environment that will allow us to analyse the behaviour of the specimen. This dataset will be used with a machine learning algorithm to predict what type of malware (or benign) a specimen is. This process will be repeated to determine the optimal category definition.

# Solution Description and System Requirements

## Solution Description

The system that I will develop will consist of two parts. The first part is the Cuckoo environment that will perform dynamic analysis on the specimens, this will output a JSON report.

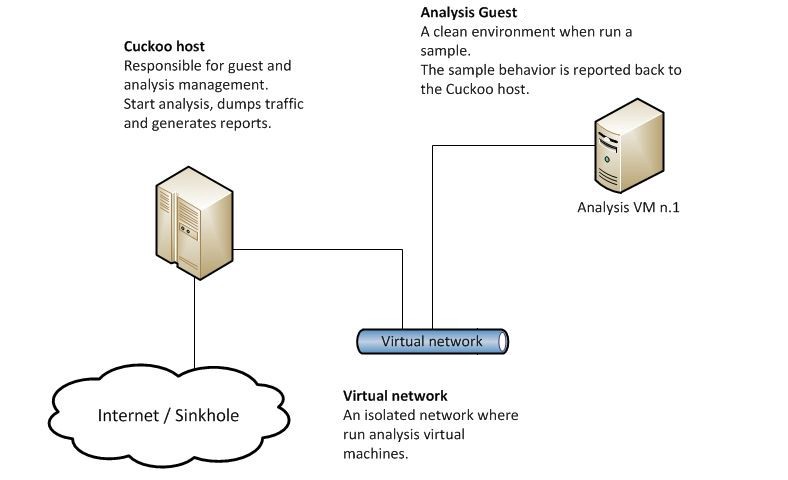


Figure 1. Example of Cuckoo Environment

The environment consists of an Ubuntu machine that has the Cuckoo software installed as well as virtual box by oracle. On this machine there will be a Windows 7 virtual machine.

The Windows 7 virtual machine will be used to run the specimen file. This will allow the Cuckoo Host to perform dynamic analysis, looking at what that file is doing when activated. After it has finished executing, the Cuckoo host will generate the JSON report.



Figure 2. An example of the JSON reports info collection.

Out of this report we are interested in the API Calls. API stands for Application Programming Interface, these are functions in dll files that an executable would run to perform various system tasks.

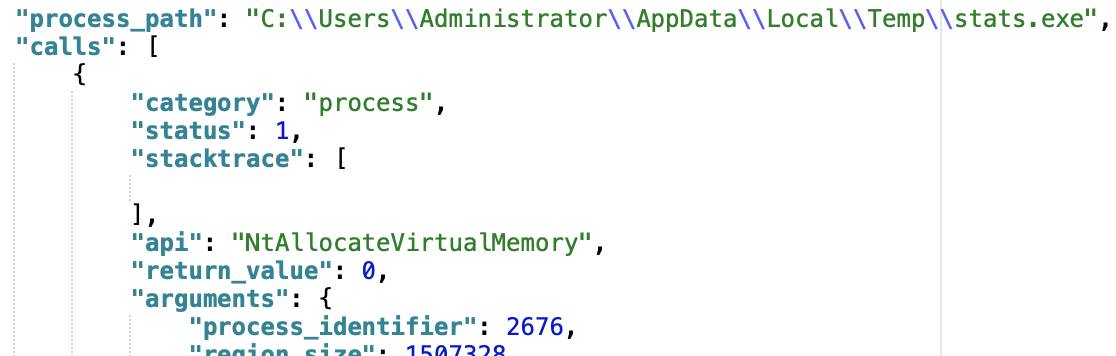


Figure 3. Example of API Call in JSON File

In the JSON report we can see a whole host of details about what the process, in this case a malicious file called “stats.exe” is trying to do on the computer. In this case it is running an API called NTAllocateVirtualMemory.

From there will we will parse the collection of JSON files into one. From there we will be able to apply a machine learning algorithm.

## System Requirements

# Design

# Implementation

# Testing

# System Evaluation and Experimental Results

# References

<https://www.forbes.com/sites/forbestechcouncil/2018/03/16/developing-the-connected-world-of-2018-and-beyond/#156cda811e51>

<https://www.varonis.com/blog/cybersecurity-statistics/>

https://www.gdatasoftware.com/blog/2017/04/29666-malware-trends-2017

# Appendices

## Work Plan