**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**

[Date]

# 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 6](#_Toc529181542)

[Implementation 6](#_Toc529181543)

[Testing 6](#_Toc529181544)

[System Evaluation and Experimental Results 6](#_Toc529181545)

[Appendices 6](#_Toc529181546)

[References 7](#_Toc529181547)

[Appendices 8](#_Toc529181548)

# Introduction and Problem Area

# Solution Description and System Requirements

# Design

# Implementation

# Testing

# System Evaluation and Experimental Results

# Appendices

# References

# Appendices