Hierarchical Density-based Clustering of Malware Behaviour

Navein Chanderan, Johari Abdullah
Faculty of Computer Science & Information Technology,
Universiti Malaysia Sarawak, 94300, Kota Samarahan, Sarawak, Malaysia.
15020358@siswa.unimas.my

***Abstract*—The numbers and diversity of malware variants grows exponentially over the years, and there is a need to improve the efficiency of analysing large number of malware samples efficiently. To address this problem, we propose a framework for the automatic analysis of a given malware's dynamic properties using clustering technique. The framework also provides outlier discovery, abnormal behaviour analysis and discrimination of malware variants. We also created a module for normalisation of malware labelling based on the labels we get from VirusTotal, which provides consistency of malware labels for accurate analysis of malware family and types. An evaluation model for the proposed framework is also discussed. Ultimately, the proposed framework will ensure rapid analysis of malware samples and lead to better protection for various parties against malicious malware.**

***Index Terms*—Anomaly Detection; Automated Dynamic Malware Analysis; Clustering; Malware Behaviour.**

I. INTRODUCTION

Malicious software, which is also popularly known as malware, is one of the major cybersecurity threats today. In fact, many cybersecurity incidents are usually caused by malware [1]. It comes in various forms, such as viruses, Trojans, worms, botnets, and rootkits, to name a few. Recent report from AV-Test reveals that it registers over 390,000 samples daily [2]. Due to the exponentially growing numbers of malware over the years, a problem that is faced by analysts is large scale malware analysis. The high number of malware samples posed difficulty for analysis as analysts need to extract meaningful information from the samples. To add to this problem, the complexity of modern malware employing evasion techniques such as polymorphism, code obfuscation and metamorphism makes analysis harder. These techniques are effective against static analysis of malware binaries [3]. In contrast, dynamic analysis of malware binaries does not have this limitation for the most part, as these evasion techniques are hard to conceal during run-time. Due to this, there are many researches which focused on dynamic analysis [5-8].

While dynamic analysis is a good approach for analysing malware samples, it does not scale as it is a time-consuming process. It also does not alleviate the problem of exponential malware sample analysis. Therefore, the ability to efficiently and automatically analyse malware behaviour is needed. This is not a new concept as it has been studied and applied before, either by clustering or by classification, usually by applying different algorithms of clustering or classification, and by applying different behaviour representations. The goal of clustering is to discover patterns of similar behaviour and to discover novel malware classes and variants [9-11]. Meanwhile, the goal of classification enables unknown malware variants to be added to existing classes of behaviours [12, 13].

In this paper, we proposed clustering of malware behaviour using hierarchical and density based algorithm (HDBSCAN) to cluster malware samples, and discover unknown variants of malware in an efficient manner.

II. RELATED WORK

Machine learning is employed to automate analysis as it can analyse large number of samples efficiently for the discovery of novel malware, reduce analysis efforts and provide insights into patterns and trends. There are basically two main approaches to machine learning for malware analysis, classification, and clustering. It can be done based on static analysis or dynamic analysis. Based on previous research [10], it has been shown that machine learning based on dynamic analysis gives better results than machine learning analysis based on static analysis, due to the limitations of static analysis.

Since our work focuses on the clustering of malware behaviour for unknown malware, to aid the discovery of unique samples, reduce manual analysis time, and to discover patterns of malware behaviour, we focus on these line of research works. Various methods have been proposed for this purpose with varying level of success. In [10], the authors modelled malware behaviour as a non-transient state changes. Although their technique achieves good results by abstracting higher level calls, the system fails to recognize the relationship between state changes, and thus does not paint a complete picture of malware behaviour, as compared to fine-grained analysis. In [28], the authors use dynamic analysis and machine learning to estimate malware functions, which is useful in identifying the characteristics and behaviour family of malware. The methods and results of this research looks promising.

Bayer et al. [11] proposed a fine-grained malware behaviour analysis. Their framework utilised local sensitivity hashing (LSH) on features extracted, to reduce the number of comparisons during clustering. However, the variable length feature representation makes their approach less scalable. Rieck et al. [23] on the other hand, uses prototype-based clustering to approximate malware behaviour, which reduces the run-time complexity. However, the n-gram approach that the framework uses is susceptible to behaviour obfuscation. In [29], the authors used hybrid deep learning approach to model malware call sequences for classification by combining recurrent neural networks with convolutional neural networks. Using these techniques, the algorithm gets a