Literature Survey

I. Introduction

Looking at the literature of malware detection, there are generally three approaches. The first is done by taking large surveys where the authors will take broad known obfuscation techniques and compare them to detection methods currently deployed [13]. The other methods take a more novel approach. The first method is they use new techniques such as Profile Hidden Markov Models (PHMM) and pair them with currently known methods of analysis such as dynamic analysis, API calls, and constructing birthmarks to try and improve detection quality and create a more robust malware detection framework. The third approach taken by the literature is using malware detection as a baseline and applying novel techniques to obfuscate itself and escape detection. While the former approach gives a good overview of the state of current techniques, the latter two approaches to literature are in a sort of arms race with one side trying to create methods to not be detected and the other side creating methods to detect them. As it stands currently there are some papers for obfuscation of a malware's signature, and there is substantial research for detection. Literature on detection ranges from dynamic and static analysis to using PHMMs as a novel method of detection. There is currently no research done on how to obfuscate PHMM based malware detection.

II. Body

The first kind of research is more broad and more surface level. It will not go as deep into a a technique, but rather it will test and measure a wide variety of techniques. Research is extensively done for exactly how polymorphic and metamorphic malware uses techniques to modify itself in intelligent ways to evade detection at a high success rate [9]. For example, they will take certain kinds of evasion techniques commonly done by polymorphic and metamorphic malware such as adding dead code, transposing code, or reordering subroutines [13]. There is more research showing how there are already various techniques for malware detection, both static and dynamic to render common signature based detection obsolete. This research mostly gives analysis on the arms race between malware detection and obfuscation, and gives good context and data to compare different methods of detection or obfuscation.

There is also novel research being done to evade detection. Novel techniques such as conditional code obfuscation were developed and was found that can evade state of the art malware detection [8]. With new obfuscation schemes, new methods of detection are needed if malware can can beat detection and can hide malicious behavior and activity.

With novel obfuscation techniques being developed, there are also novel detection methods being researched, particularly by using bioinformatics tools to analyze and catch how malware evolves. There is research done for specific kinds of evasions of malware where the malware's behavior will change if it thinks it is being analyzed [4]. Additionally there is a lot of research being done with using PHMM for malware detection. PHMM is usually used for categorizing protein families, but it has been found to be very effective in catching and categorizing malware families. PHMM has been used to detect malware in many different ways. PHMM has been incorporated in malware detection by using dynamic birthmarks [12] and has been used to classify known malware [5]. While some research found high sensitivity scores, they had a high false positive rate as well. Newer research has been done and reached one hundred percent detection rate with a very low false positive rate [1]. PHMM has also been used in a variety of ways to tackle detection using static and dynamic forms of analysis such as analyzing system call sequences [5], doing behavior based analysis [6], doing static analysis based off of op code sequences [1], and using dynamic analysis techniques [11]. PHMM in particular has also been shown to be robust on a variety of operating systems and successfully detects malware using API extractions on various platforms such as android [7].

In parallel, there is also a lot of research being done for extract API calls and using those calls for malware analysis. Research has been done for static detection of malware based off of the binary's executable [2]. Additionally there has been plenty of work demonstrating efficacy in extracting API calls and being able to classify them based off of API sequence [10] or based off of API call frequency [3]. These methods are often paired with machine learning and yield promising results. Combining this approach with PHMM has shown to be very efficacious for detecting malware.

III. Conclusion

There are three general paths of research. The first giving a general overview and comparison of current categorization techniques and methods of obfuscation. It is generally formatted as a survey. The latter two forms of published research focus on developing novel, state of the art detection or obfuscation. There is a lot of research done demonstrating that PHMM is very apt for malware detection in a variety of techniques. There is not, however, a published paper showing any kinds of methods that would lower the efficacy or render it obsolete in any capacity. While there are papers in particular examining PHMM detection using extracted API calls that seem very effective, there is no effort to examine if obfuscation of these API calls would lead to PHMM being obsolete. This paper intends to demonstrate obfuscation techniques showing ways to evade detection from PHMM based detection algorithms.

REFERENCES

[1] Alireza Abbas Alipour and Ebrahim Ansari. "An advanced profile hidden Markov model for malware detection". In: *Intelligent Data Analysis* 24.4 (2020), pp. 759–778.

[2] Wen Fu et al. "Static detection of api-calling behavior from malicious binary executables". In: *2008 International Conference on Computer and Electrical Engineering*. IEEE. 2008, pp. 388–392.

[3] Vidhi Garg and Rajesh Kumar Yadav. "Malware Detection based on API Calls Frequency". In: *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE. 2019, pp. 400–404.

[4] Dhilung Kirat and Giovanni Vigna. "Malgene: Automatic extraction of malware analysis evasion signature". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015, pp. 769–780.

[5] Ramandika Pranamulia, Yudistira Asnar, and Riza Satria Perdana. "Profile hidden Markov model for malware classification—usage of system call sequence for malware classification". In: *2017 International Conference on Data and Software Engineering (ICoDSE)*. IEEE. 2017, pp. 1–5.

[6] Saradha Ravi, N Balakrishnan, and Bharath Venkatesh. "Behavior-based malware analysis using profile hidden markov models". In: *2013 International Conference on Security and Cryptography (SECRYPT)*. IEEE. 2013, pp. 1–12.

[7] Satheesh Kumar Sasidharan and Ciza Thomas. "ProDroid—An Android malware detection framework based on profile hidden Markov model". In: *Pervasive and Mobile Computing* 72 (2021), p. 101336.

[8] Monirul I Sharif et al. "Impeding Malware Analysis Using Conditional Code Obfuscation." In: *NDSS*. Citeseer. 2008.

[9] Jagsir Singh and Jaswinder Singh. "Challenge of malware analysis: malware obfuscation techniques". In: *International Journal of Information Security Science* 7.3 (2018), pp. 100–110.

[10] Dolly Uppal et al. "Malware detection and classification based on extraction of API sequences". In: *2014 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE. 2014, pp. 2337–2342.

[11] Swapna Vemparala. "Malware detection using dynamic analysis". In: (2015).

[12] Swapna Vemparala et al. "Malware detection using dynamic birthmarks". In: *Proceedings of the 2016 ACM on international workshop on security and privacy analytics*. 2016, pp. 41–46.

[13] Ilsun You and Kangbin Yim. "Malware obfuscation techniques: A brief survey". In: *2010 International conference on broadband, wireless computing, communication and applications*. IEEE. 2010, pp. 297–300.