Accepted Manuscript

Title: Auto-detection of sophisticated malware using lazy-binding control flow graph and deep learning

Author: Nguyen Minh Hai, Le Nguyen Dung, Nguyen Xuan Mao, Quan Thanh Tho

PII: S0167-4048(18)30088-9
DOI: https://doi.org/10.1016/j.cose.2018.02.006
Reference: COSE 1291

To appear in: *Computers & Security*

Received date: 2-6-2017
Revised date: 12-12-2017
Accepted date: 11-2-2018

Please cite this article as: Nguyen Minh Hai, Le Nguyen Dung, Nguyen Xuan Mao, Quan Thanh Tho, Auto-detection of sophisticated malware using lazy-binding control flow graph and deep learning, *Computers & Security* (2018), https://doi.org/10.1016/j.cose.2018.02.006.

Auto-Detection of Sophisticated Malware using Lazy-binding Control Flow Graph and Deep Learning

Nguyen Minh Hai[a], Le Nguyen Dung[a], Nguyen Xuan Mao[b], Quan Thanh Tho*[a]
[a]Ho Chi Minh City University of Technology, Vietnam; [b]YouNet Group, Vietnam

*Corresponding author: qttho@hcmut.edu.vn

Mr. Nguyen Minh Hai is a PhD student at Ho Chi Minh City University of Technology, Vietnam. His research interests include binary analysis, formal method, malware detection and intelligent systems.

Mr. Le Nguyen Dung is a master student at Ho Chi Minh City University of Technology, Vietnam. His research interests include binary analysis and malware detection.

Mr. Nguyen Xuan Mao is a research scientist at YouNet Corporation, Vietnam. His research interests include deep learning and intelligent systems.

Dr. Quan Thanh Tho is an Associate Professor in the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam. He received his B.Eng. degree in Information Technology from HCMUT in 1998 and received Ph.D degree in 2006 from Nanyang Technological University, Singapore. His current research interests include formal methods, program analysis/verification, the Semantic Web, machine learning/data mining and intelligent systems. Currently, he heads the Department of Software Engineering of the Faculty.

ABSTRACT

To date, industrial anti-virus tools are mostly using signature-based methods to detect malware occurrences. However, sophisticated malware, such as *metamorphic* or *polymorphic virus*, can effectively evade those tools by using some advanced obfuscation techniques, including *mutation* and the *dynamically executed contents (DEC) methods*, which dynamically produce new executable code in the run-time. Common DEC methods used by malware programs are *packing* or *calling external code*. In the research community, the approach of program analysis to detect suspicious behaviors has been emerging recently to handle this problem. *Control flow graph* (CFG) is a suitable representation to capture common behaviors from various mutated samples of virus. However, the current typical CFG forms generated by state-of-the-art binary analysis tools, such as IDA Pro, do not precisely reflect the behaviors of DEC methods. Moreover, this approach suffers from an extremely heavy cost to conduct and analyze the CFGs from binaries. This drawback causes the method of formal behavior analysis to be virtually not applicable with real-world applications.

In this paper, we propose an enhanced form of CFG, known as *lazy-binding CFG* to reflect the DEC behaviors. Then, with the recent advancement of the *deep learning* techniques, we present a method of producing image-based representation from the generated CFG. As deep learning is very popular to perform image classification on very large dataset, our proposed technique can be applied for malware detection on real-world computer programs and thus enjoying very high accuracy. We also