Poster: Android Malware Detection using Multi-Flows and API Patterns

Feng Shen, Justin Del Vecchio, Aziz Mohaisen, Steven Y. Ko, Lukasz Ziarek
University at Buffalo, The State University of New York
{fengshen, jmdv, mohaisen, stevko, lziarek}@buffalo.edu

ABSTRACT

This paper proposes a new technique for detecting mobile malware based on information flow analysis. Our approach focuses on the *structure* of information flows we gather in our analysis, and the *patterns* of behavior present in information flows. Our analysis not only gathers *simple* flows that have a single source and a single sink, but also *Multi-Flows* that either start from a single source and flow to multiple sinks, or start from multiple sources and flow to a single sink. This analysis captures more complex behavior that both recent malware and recent benign applications exhibit. We leverage *N-gram analysis* to understand both unique and common behavioral patterns present in Multi-Flows. Our tool leverages N-gram analysis over sequences of API calls that occur along control flow paths in Multi-Flows to precisely analyze Multi-Flows with respect to app behavior.

Using our approach, we show that there is a need to look beyond simple flows in order to effectively leverage information flow analysis for malware detection. By analyzing recently-collected malware, we show there has been an evolution in malware beyond simply collecting sensitive information and immediately exposing it. Many previous systems focus on identifying the existence of *simple* information flows—i.e. considering an information flow as just a (source, sink) pair. However, modern malware performs complex computations before, during, and after collecting sensitive information and tends to aggregate data before exposing it. A simple (source, sink) view of information flow does not adequately capture such behavior.

The uniqueness of our approach comes from the following two features. First, our information flow analysis represent an information flow not as a simple (source, sink) pair, but as *a sequence of API calls*. This gives us the ability to distinguish different flows with same sources and sinks based on the *computation* performed along the information flow. Second, our information flow analysis detects Multi-Flows, flows that either start with a single source and flow to multiple sinks, or start with multiple sources and flow to a single sink. We treat such flows as a single flow, instead of multiple distinct flows. Fig. 1 shows a comparison of a Multi-Flow and its corresponding simple flows. This allows us to examine the *structure* of the flows themselves. We leverage machine learning techniques to extract features from Multi-Flows and their API sequences (N-gram analysis) and use these features to perform SVM-based classification.

Based on this approach, we build an open source implementation of Multi-Flow analysis and API sequencing in the BlueSeal framework [2] [3] [1], along with N-gram analysis and a SVM-based classifier. We also conduct a detailed evaluation study, highlighting the differences in old and new apps. We leverage the app behavior extracted as features from both Multi-Flows and their API usage patterns and apply machine learning techniques to automatically identify malware based on the structure of its computation over sensitive data. We test our tool on a set of 1,576 benign apps downloaded from Google Play and 2,422 known malicious apps. Our results show that app behavior difference on sensitive data can be a significant factor in malware detection.

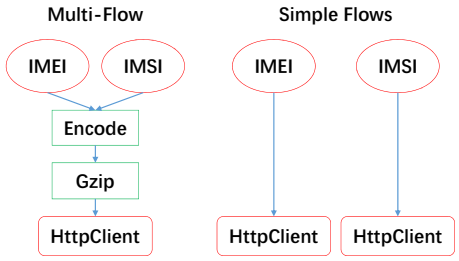

Figure 1: Multi-Flow vs Simple Flows

References

[1] S. Holavanalli, D. Manuel, V. Nanjundaswamy, B. Rosenberg, F. Shen, S. Y. Ko, and L. Ziarek. Flow permissions for android. In *Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering (ASE 2013)*, 2013.

[2] F. Shen, J. Del Vecchio, A. Mohaisen, S. Y. Ko, and L. Ziarek. Android malware detection using complex-flows. In *Proceedings of The 37th IEEE International Conference on Distributed Computing Systems*, ICDCS '17.

[3] F. Shen, N. Vishnubhotla, C. Todarka, M. Arora, B. Dhandapani, E. J. Lehner, S. Y. Ko, and L. Ziarek. Information flows as a permission mechanism. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*, ASE '14.

MobiSys'17 June 19-23, 2017, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-4928-4/17/06.
DOI: http://dx.doi.org/10.1145/3081333.3089315