Malware Detection Using Context-Free Grammars

Daniel Ovšonka
iovsonka@fit.vutbr.cz

This work will discuss two different approaches for using context-free grammars (CFG) for intrusion detection systems. My PhD thesis deals with malware detection algorithms based on artificial intelligence. Hence, I would like to show the possibilities that CFG offers in the field of information systems security. Two widely different examples were picked for better demonstration, where CFG can be used. First example focuses on the detection of polymorphic malware [1], which represents major problem of the standard detection mechanisms. Second approach points to the possibility of using probabilistic CFG to detect botnet network traffic [2].

Various kinds of the malicious software continue to spread, despite the use of network intrusion detection systems, firewalls or antivirus programs. This is mainly due to its inability to detect obfuscated variant of malware, because of the use of signature based algorithms. In this case, when the original signature of known attack, which is stored in system database, and the signature acquired while analyzing obfuscated suspicious data is compared, comparison fails. More advanced methods, which are implemented in commercial detection systems, usually use heuristic or behavioral analysis. These methods are known because of their high number of false alarm ratio. This fact forces us to explore new ways of detecting suspicious behavior in network communication or on the host station. This work briefly describes the approach based on the formal models created by using CFG.

On the other hand, botnets are becoming a major source of distributed denial of service (DDoS) attacks, spam and other cybercrimes. Botnets have evolved over time into the powerful tools, which use strong encryption algorithms and peer-to-peer technologies. Therefore, detecting and removing botnets is becoming a challenging and important task for the security experts. Approach, described in this work, uses probabilistic CFG to represent recursive patterns in botnet network traffic. These formal models are further investigated using timing analysis.

The main goal of this work is to introduce the basic principles of malware detection methods, based on formal models represented by CFG. First of all, we will briefly introduce the mathematical definition of CFG and probabilistic CFG. Then, we will discuss how it can be used for detection of suspicious programs or suspicious network flows. Two methods were chosen as examples, in order to establish the universality of this approach.

References

[1] Gerald R. Thompson and Lori A. Flynn. Polymorphic malware detection and identification via context-free grammar homomorphism. *Bell Labs Technical Journal*, 12(3):139–147, 2007.

[2] Chen Lu and Richard R. Brooks. Timing analysis in P2P botnet traffic using probabilistic context-free grammars. In *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop*, CSIIRW '13, pages 14:1–14:4, New York, NY, USA, 2013. ACM.