Spring 3-27-2019

Bust-A-Binary: Active Attribution and Analysis of Malware Campaigns

Micah Flack
Dakota State University

Nathan Kramer
Dakota State University

Zayn Snyder
Dakota State University

Ezra Chona
Dakota State University

Matthew Steckelberg
Dakota State University

See next page for additional authors

Authors
Micah Flack, Nathan Kramer, Zayn Snyder, Ezra Chona, Matthew Steckelberg, and Bramwell Brizendine

Bust-A-Binary: Active Attribution and Analysis of Malware Campaigns

MICAH FLACK, NATHAN KRAMER, ZAYN SNYDER, EZRA CHONA, MATTHEW STECKELBERG
ADVISOR: BRAMWELL BRIZENDINE

Abstract

The applied research provides for the implementation of several systems to create a new web platform framework for analyzing binaries; it is defined by Django with MongoDB, Nginx, and Gunicorn on Ubuntu for the server-side rendering of client requests, Cuckoo for runtime behavioral analysis, Suricata and YARA based signature defining, NSRL querying for whitelisted MD5 hashes, and open-source intelligence via VirusTotal and AlienVault OTX. This allows for competent management of incoming web application requests and outgoing, high-level, analytical data which will yield greater depth through the extraction of more trivial indicators. This open and collaborative setting will allow other researchers to readily review and shorten the necessary time for investigations.

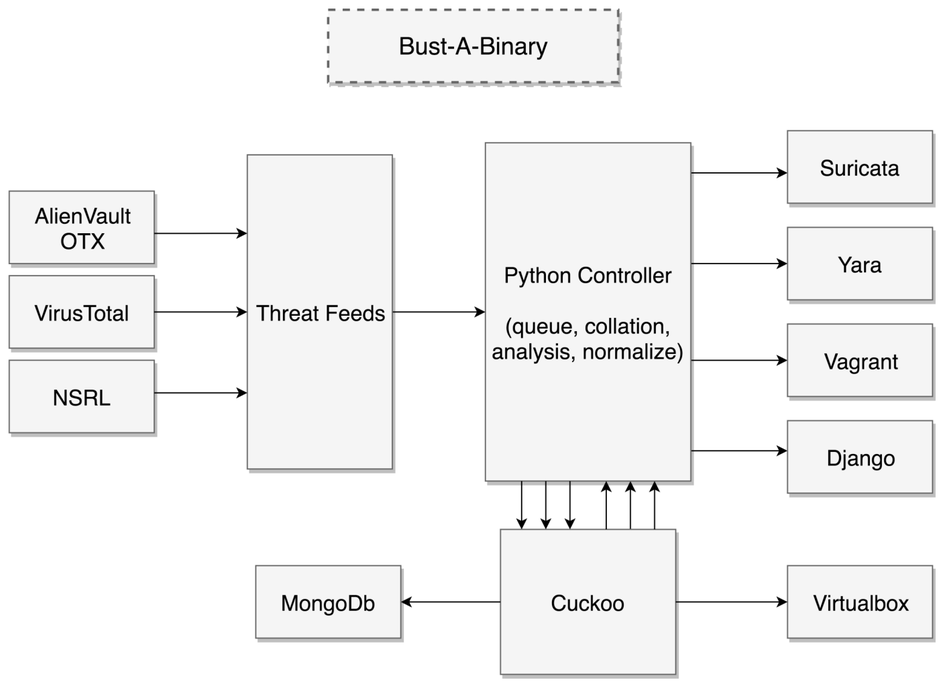
Background

Malware is typically studied in two ways: the first is static analysis, which requires manually abstracting information from several key indicators without running the binary and potentially compromising the system; the second is dynamic analysis, which requires running the potential malware and observing how it behaves or reacts to the environment and user interaction, notwithstanding the files located on the system or network which were spawned by the malware. The research here proposes a novel application of several areas of study. Bust-A-Binary encompasses machine learning, the application of an algorithm; this is capable of finding the properties of previously unseen samples between massive sets of data. Machine learning has a broad variety of approaches that it takes to a solution, rather than a single method. These approaches have different capacities and different tasks that they suit best. Utilizing different approaches, this research aims to provide a tool that can effectively combine static and dynamic analysis with machine learning to identify key similarities.

Literature Review

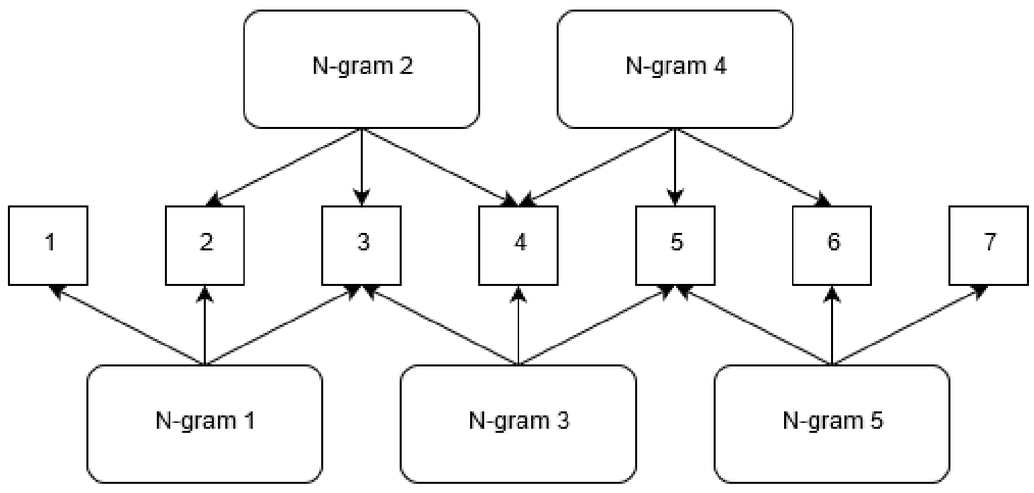
Ortiz introduces several novel applications of deployment software for content between local networks as well as guidelines for managing and creating flexible, dynamic virtual machines (2015). DiRaimondi provides guidance to analysts by reviewing concepts core to the Laika BOSS framework, integrating custom Yara rules for file-based detections, and searching and filtering scan object metadata. He describes how to develop, test and implement new Laika BOSS modules to extend and automate new functionality into the framework (2018). Further capability is gained when extended with machine learning concepts that allow these relationships and key indicators to identify families or similarities between samples (Saxe, 2018).

Methodology



Results and Conclusion

Current results have indicated that the suggested design proposal, however theoretical, is both practical and useful. By comparing several samples or entire malware families and their indications against complex decision boundaries for logistic regression, we can demonstrate clear relationships between common metadata, such as context triggered piecewise hashes. This is further evidenced by the use of Jaccard indexes and sequence based similarities. However, the intended efficacy of these techniques varies widely depending on the implementation and the data types used. Using n-gram comparisons, the order with which certain behaviors were observed, is one such method used to distill key features between samples.



Demonstrating n-grams extracted from malware execution threads

As such, the provided results represent a combination of all the research methods into one, open and seamless source, capable of accessing current analysis results and similar attributes, dictated by their relational metadata. Despite the current results, future goals aim to introduce other techniques, such as Vagrant, Packer, Ansible, Chef, or Volatitiy, for deployment management and information dissemination across the platform.