Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Classification and Phylogenetics of Malware

CHALLENGE

In the current environment of advanced nation-state and highly organized criminal actors creating malicious software (malware), characterization, identification, and attribution are difficult problems that must be addressed to respond to these threats. Detecting and understanding malware are significant challenges because of many factors, including:

- The rate new variants are generated from old attacks;
- The sheer volume of data that must be analyzed to detect or characterize malware on systems or on a network; and
- Adversarial attempts at obfuscation to thwart the current state of the art of characterization, detection, and attribution.

Experienced malware analysts' "hands-on" time with individual samples of malware is a precious resource. To better use human resources, we have leveraged high-performance computing in ways that help analysts identify malware samples of interest by their similarity, both in structure and attribution.

Using a combination of bio-inspired techniques with high-performance dynamic analysis to characterize malware and its lineage.

```
>SERVER1
QLMAQMLQQANNNNNNNQLMAQIIMQALQLMQATMGNIQINAQQQQMALQL
MALAWRWRWVWVQTAGMMLLQAAQLMAMLQQAQLMLAAMLAMLATMAGQ
MLQMALATMAGQQQTAGMQQIQMALILQMQALALAWVWVLAWRWRWVWVT
MAGQQWVVWMQLAQLMAQMALQTAGQMLQMALRAWWRWVWVMLQQAQ
MLAMLQQANNNNNNNQQMLQIAQMLIQAQMLAMQLAQTMG
>SERVER2
QMLQQAQLAMQMLQQAQLAMIIQQMLQAQQMLAMQQLAQTGMAMLAMLAM
QQLAQLAWVQLAMMQQLATMAQTMAGQQTMAGQAMLQIMLQANNNNNNNI
LALALALALALALALANNNNNNNNNNNNNNNNNNNLALALALALALALALA
NNNNNNNNNNNNNLALALALALALALALNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNLLLLLLLLLLLLLLLLLAAAAAAA
AAAAAAAAAAAAAAAAAALAMLQAQTMAGQLAM
>SERVER3
MLAMLANNNNNNNQMLAMLAQMLAMLAQQMLAMLATGMAMLAAMLAQQML
AAMLATGMALALLAALLAALLAALLAALLAALLAATGMAQQMATGMAMLALLAA
QLLAAMLAMLATGMAQLLAAQQTGMQQQLLAAMLATGMAQQTGMQQLALLA
ALLAALLALAAMLATGMAMLALLAALLAAMLALLAAMLANNNNNNNVTGMANN
LLAAMLAMLAQLLAAMLAMLAAMAMLATGMA
```

Biology concepts, such as protein identification, use "cyber proteins" to generate families of malware.

CURRENT PRACTICE

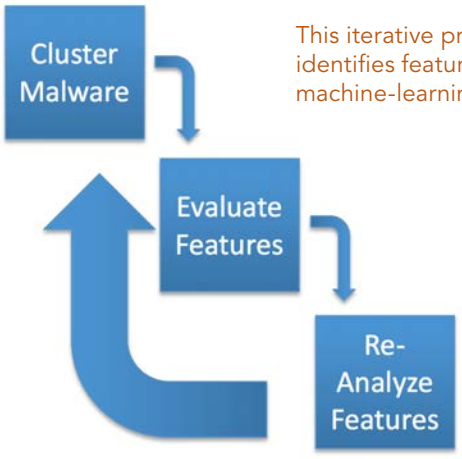
Current malware classification is accomplished mostly through manual and semi-automated processes. Generating signatures typically happens in one of two ways: reactively, or with expert knowledge. Similar attempts at attribution typically are accomplished by manual inspection of code and comparison to known signatures. This is trivial for malware to evade, and also prone to false classification and

attribution, depending on the quality of the signatures. We have developed two approaches to solve these problems that take advantage of high-performance computing, machine learning, and bio-inspired techniques. These two approaches can be combined to generate better classifications and identifications of malware. The approaches will support the concepts of generating family trees of malware, which can then inform attribution.

TECHNICAL APPROACH

The team at Pacific Northwest National Laboratory (PNNL) developed Machine Learning String Tools for Operational and Network Security (MLSTONES), a computational high-performance computing capability for applying biosequence analysis to cybersecurity applications. MLSTONES can be used to discover similarity in software, particularly for malware. The MLSTONES process creates "cyber protein" representations of malware and then uses protein alignment techniques to quickly generate families of proteins. With this method, we can create a single representation of an entire family of entities, significantly reducing the amount of data to analyze. MLSTONES has been applied to the analysis of a malicious software corpus to produce "family trees" of malware artifacts and a signature library. These family trees can be shared, to deploy the analytical technique into operational environments. Our high-performance computing implementation of this biosequence-based capability is ideal for malicious code analysis because we can infer the function of a "cyber protein" by its relationship to other similar proteins, the same process used in biology.

The team at Mississippi State University (MSU) is identifying features of malicious software that can be used to help define capabilities and relationships of malware. MSU's focus is on dynamic features, those generated at runtime. The dynamic features are generated while executing the malware in a sandbox environment. During execution, any interactions with the operating system



This iterative process identifies features for machine-learning clustering.

are recorded and stored as features. Once execution is complete, memory is captured and analyzed to extract additional features that the sandbox might have missed at runtime.

These features are stored in a database and can be used to identify the malware capabilities and relate groups of similar samples. High-performance computing resources are necessary to effectively process the volume of new malware being discovered on a daily basis.

The combination of PNNL's MLSTONES with MSU's dynamic features enables the rapid characterization, classification, and detection of malware in the wild.

IMPACT

Work on this project increases the likelihood of finding the source of an attack or widespread malware infection, as well as identifying new zero-day types of attacks. With the ability to leverage high-performance computing resources, analysts will be able to take current sources of potential new malware samples and quickly classify them by potential authorship or common source. We will also be able to identify the type of those new malware samples–even if they are not exact matches to standard signatures. New malware samples may be grouped with existing ones that have a high degree of similarity, giving analysts more information on specific threat actors and new types of malware. This benefits those who wish to attribute cyber attacks by providing more data that can be examined for operational mistakes by the attackers.

Contacts

Elena Peterson
Principal Investigator
(509) 372-4573
Elena.Peterson@pnnl.gov

Dae Glendowne
Principal Investigator
(662)-325-5964
jg905@msstate.edu

David A. Dampier
Director, Distributed Analytics and Security Institute
(662) 325-0779
dampier@dasi.msstate.edu

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

MISSISSIPPI STATE UNIVERSITY