CA Labs Research Brief: Internet Security Solutions

Malware detection and classification

Challenge

Every day thousands of new malware variants emerge that rely on common types of previously classified malicious code, but anti-virus engineers lacked automated detection and classification technologies to help keep common malware 'families' under control.

Focus

This CA Labs research project created generic signatures for the malicious code used in malware families.

Result

Generic signatures can enable faster detection, classification and neutralization of new malware variants. Generic signatures can also increase the performance of detection software, which consumes fewer operating resources by running fewer signatures.

Malicious code exists in nearly all computer platforms and languages, yet most new malware instances are variants of previously seen malware. Anti-virus engineers were swamped by an ever-increasing research workload because traditionally every new variant had to be manually analyzed and classified, and a signature created that identified the unique byte sequence for the malware variant.

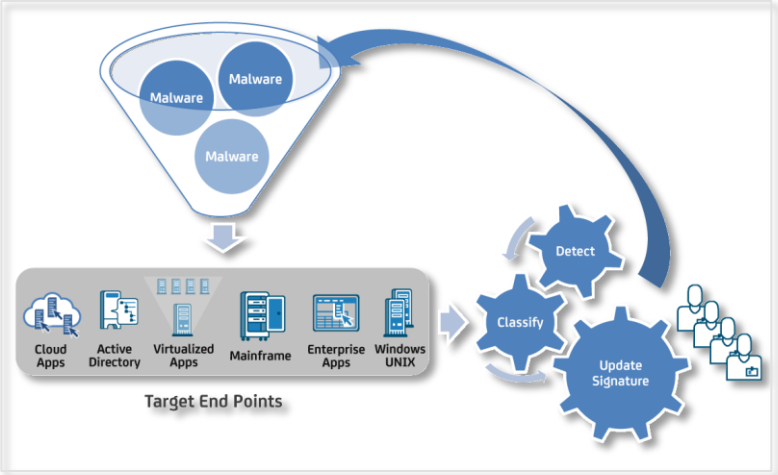

Figure 1 Detection, classification and signature updates to neutralize malware is a manually intensive process which is streamlined by generic malware DNA signatures that classify malware families automatically

The CA Labs Malware Detection and Classification research project approached this problem by identifying the malware DNA — the patterns and combinations of patterns which relate to entire families of malware. The research employed statistical analysis and machine learning algorithms on a catalog of known malicious code samples and developed methods for effectively classifying whether unknown executable files are benign or malicious and, if malicious, which malware family they belong to.

The research methods flexibly integrate with the workflows of malware detection and classification systems to help ensure easy signature maintenance and they have the potential to reduce the system's scanner codebase requirements for improved operating performance. Engineers can be more efficient and effective in fighting malware with an automated way to detect and classify malware families based on common DNA signatures. Generic malware DNA signatures can also save resources when creating new signatures and new family variants can be detected even if running without the latest signature updates.

Additionally, the successful project findings and outcomes are helping engineers focus their efforts on identifying new unclassified malware variants. Historical data captured during detection and classification processes enables deeper statistical and heuristic analysis of malware trends. The trend data is being used to advance the intelligence of antivirus engines to identify the possible presence of un-cataloged malware.

Malware detection and classification at work

The research leverages disassembly tools and a relational database of generic DNA signatures of malware families to automate detection processes. Common signature data is extracted from the database and combined into a vector for fast, scalable classification.

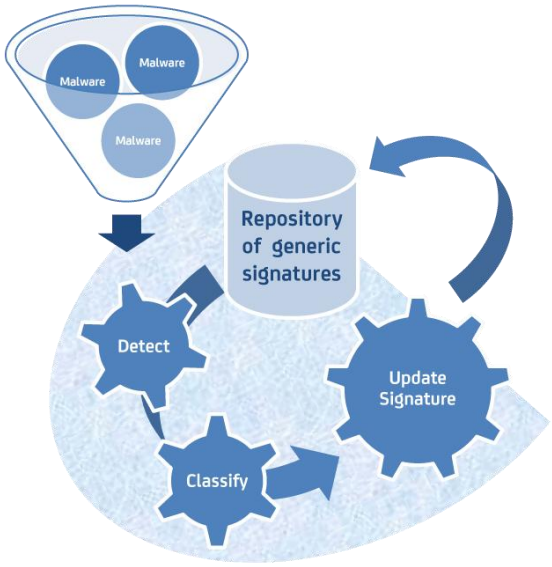

Figure 2 Generic signatures provide the foundation for analysis and classification of a malware instance. Signature updates are self-propagating.

Classification adds the malware instance to its proper place in the malware phylogenetic[1] tree and the generic signature is updated in the database in a self-perpetuating manner. The generic signatures used in detection and classification are always the most current, while periodic pushes to end user threat management software is based on what is current in the database at the time of the push.

[1] The evolutionary relationship among groups of organisms, or in this case, different malware instances.

More information on the CA Labs Malware Detection and Classification research project

CA Labs collaborated with researchers from Deakin University in Australia. The following papers have been published about this research project:

- R. Tian, L.M. Batten, R. Islam, S.C. Versteeg. "Differentiating Malware from Cleanware Using Behavioural Analysis." Fifth IEEE International Conference on Malicious and Unwanted Software (Malware'10). Nancy, France. October 2010.
- R. Islam, R. Tian, L.M. Batten, S.C. Versteeg. "Classification of Malware Based on String and Function Feature Selection". Second Cybercrime and Trustworthy Computing Workshop. Ballarat, Victoria, Australia. July 2010.
- R. Tian, L.M. Batten, R. Islam, S.C. Versteeg. "An Automated Classification System Based on the Strings of Trojan and Virus Families". In *Proceedings of the Fourth IEEE International Conference on Malicious and Unwanted Software* (Malware'09), Montreal, QC, Canada. October 2009.
- R. Tian, L.M. Batten, S.C. Versteeg. "Function Length as a Tool for Malware Classification." Third IEEE International Conference on Malicious and Unwanted Software (Malware'08). Alexandria, VA, United States. October 2008.

For additional information about this or other CA Labs projects, please contact Steve Versteeg at Steve.Versteeg@ca.com.

About CA Labs and innovation

CA Labs is the research arm for CA Technologies and a hub for the company's initiatives for innovation. CA Labs collaborates with the world's foremost researchers in academia, industry and government to perform advanced research to address cloud, software-as-a-service, security, virtualization, automation, mainframe, service assurance, and service and portfolio management challenges. For more information, visit ca.com/calabs.

ca technologies