Math vs. Malware

Is There A Better Way? Insights into how math can defeat malware.

“Mathematics is a more powerful instrument of knowledge than any other that has been bequeathed to us by human agency.”–Rene Descartes, French philosopher, mathematician and scientist

The problem–although few want to admit it–is that enterprise security personnel are defending a castle riddled with holes, filled with secret passageways, and protected by ineffective barriers. These weak points are a consequence of poor quality security software, older hardware, and in some cases, backdoors planted by malicious insiders. The end result is a begrudging acceptance that the attackers are winning the war.

Attacks are motivated by a variety of reasons, originate from various locales and continue to evolve in complexity as technology progresses. As part of this evolution, modern threats commonly employ evasion techniques designed to bypass existing security measures. Simply detecting these advanced threats after the fact is hard enough, let alone protecting an entire organization against them beforehand.

What if there is a better way?
What if the castle can be defended?
What if the threat could be stopped long before the damage was done?

The Human Factor

In order to keep up with modern attackers, security technologies need to evolve alongside them–without relying on human intervention. That’s where math and machine learning have the advantage. If we can objectively classify “good” files from “bad” based on mathematical risk factors, then we can teach a machine to make the appropriate decisions on these files in real time.

In the pages that follow, we posit that a math and machine-learning approach to computer security will fundamentally change the way we understand, categorize and control execution of every file. We’ll also discuss how Dell products leverage this approach and demonstrate just how different they are from every other security offering on the market.

For years industries such as healthcare, insurance, and high-frequency trading have applied the principals of machine learning to analyze enormous quantities of business data and drive autonomous decision making. At the core of each implementation is a massively scalable data processing ‘brain’ capable of applying highly-tuned mathematical models to enormous amounts of data in near real-time.

Applying Machine Learning to File Classification

Machine Learning Defined

"Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data ... The core of machine learning deals with representation and generalization. Representation of data instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory." - Wikipedia

Over the past few decades, billions of files have been created–both malicious and non-malicious. In the file creation evolution, patterns have emerged that dictate how specific types of files are constructed. There is variability in these patterns as well as anomalies, but as a whole, the computer science process is reasonably consistent.

The patterns become even more consistent as one looks at different development shops such as Microsoft®, Adobe® and other large software vendors. These patterns increase in consistency as one looks at development processes used by specific developers and attackers alike. The challenge lies in identifying patterns, understanding how they manifest across millions of attributes and files, and recognizing what consistent patterns tell us about the nature of these files.

Because of the magnitude of the data involved, the tendency towards bias and the number of computations required, humans are incapable of leveraging this data to make a determination as to whether the file is malicious or not. Unfortunately, most security companies still rely on humans to make these determinations. They hire a large number of people to look through millions of files to determine which are "good" and which are "bad".

Humans have neither the brainpower nor the physical endurance to keep up with the overwhelming volume and sophistication of modern threats. Advances have been made in behavioral and vulnerability analysis, as well as identifying indicators of compromise, but these "advances" all suffer from the same fatal flaw. They are all based on the human perspective and analysis of a problem–and humans err towards over-simplification.

Machines, however, do not suffer from this same bias.

How It Works

Machine learning and data mining go hand-in-hand. Machine learning focuses on prediction based on properties learned from earlier data. This is how Dell differentiates malicious files from safe or legitimate ones. Data mining focuses on the discovery of previously unknown properties of data, so those properties can be used in future machine learning decisions.

Machine learning leverages a four phase process: collection, extraction, learning and classification.

Collection

Much like a DNA analysis or an actuarial review, file analysis starts with the collection of a massive amount of data–in this case files of specific types (executables, .pdf, .doc. Java, flash, etc.). Hundreds of millions of files are collected via 'feeds' from industry sources, proprietary organizational repositories and live inputs from active computers with Dell agents on them[1]

The goal of collection is to ensure one:

- Has a statistically significant sample size.
- Has sample files that cover the broadest possible range of file types and file authors (or author groups such as Microsoft, Adobe, etc.)
- HAS NOT biased the collection by over-collecting specific file types

Once these files are collected, they are reviewed and placed into three buckets; known and verified valid, known and verified malicious and unknown. It's imperative to ensure that these buckets are accurate–including malicious files in the valid bucket or valid files in the malicious bucket would create incorrect bias.

Extraction

The next phase in the machine learning process is the extraction of attributes. This process is substantively different from the process of behavior identification or malware analysis currently conducted by threat researchers.

Rather than looking for things which people believe are suggestive of something that is malicious, Dell leverages the compute capacity of machines and data-mining techniques to identify the broadest possible set of characteristics of a file. These characteristics can be as basic as the PE file size or the compiler used and as complex as a review of the first logic leap in the binary. We extract the uniquely atomic characteristics of the file depending on its type (.exe, .dll, .com, .pdf, .java, .doc, .xls, .ppt, etc.).

By identifying the broadest possible set of attributes, Dell removes the bias introduced by the manual classification of files. Use of hundreds of thousands of attributes also substantially increases the cost for an attacker to create a piece of malware that is not detected by Dell.

The result of this attribute identification and extraction process is the creation of a file genome very similar to that used by biologists to create a human genome. This genome is then used as the basis for which mathematical models can be created to determine expected characteristics of files, much like human DNA analysis is leveraged to determine characteristics and behaviors of cells.

Learning and Training

Once the attributes are collected, the output is normalized and converted to numerical values that can be used in statistical models. This is where vectorization and machine learning are applied to eliminate the human impurities and speed analytical processing. Leveraging the millions of attributes of files identified in extraction, our mathematicians then develop statistical models that accurately predict whether a file is valid or malicious.

Dozens of models are created with key measurements to ensure the predictive accuracy of the final models used. Ineffective models are scrapped and effective models are run through multiple levels of testing. The first level starts with a few million known files and later stages involve the entire file corpus (tens of millions of files). The final models are then extracted from the test corpus and loaded into the production environment for use in file classification.

It's important to remember that for each and every file, thousands of attributes are analyzed to differentiate between legitimate files and malware. This is how our engine identifies malware–whether packed or not, known or unknown–and achieves an unprecedented level of accuracy. It divides a single file into an astronomical number of characteristics and analyzes each one against hundreds of millions of other files to reach a decision about the normalcy of each characteristic.

Classification

Once the statistical models are built, our engine can be used to classify files which are unknown (e.g., files that have never been seen before or analyzed by another whitelist or blacklist). This analysis takes only milliseconds and is extremely precise because of the breadth of the file characteristics analyzed.

Because the analysis is done using statisticalmodels, the classification is not completed in a black box. Dell provides the user with a "confidence score" as part of the classification process. This score provides the user with incremental insight that they can use to weigh decisions around what action to take on the specific file–block, quarantine, monitor or analyze further.

There is an important distinction between the machine-learning approach and a traditional threat research approach. With the mathematical approach, we build models that specifically determine if a file is valid or malicious. It will also return a response of "suspicious" if our confidence about its malicious intent is less than 20% and there are no other indications of malicious intent. In so doing, the enterprise gains a holistic perspective on the files running in their environment. It also eliminates the current industry bias in which threat researchers only determine if a file is malicious and whitelist vendors only determine if a file is good.

Other than the obvious benefits of detecting a larger amount of threats, there are more subtle benefits to this approach; every file that is analyzed is evaluated using the classification algorithms. While this may seem straight forward, traditional anti virus vendors only evaluate a specific file against a finite list of signatures designed to detect malware based on human analysis. Even if they use some automated techniques, they are limited to creating signatures based on specific parts of files that were previously identified as known malware by a human. Not only is there little-to-no proactivity possible with these techniques, they simply classify objects that do not match any particular signature as good. Our engine, on the other hand, analyzes every sample and provides a definitive classification of all files whether they are bad, suspicious or good. This provides the security team with a clear understanding of exactly what is running in their environment.

Future-Proof Security

By applying mathematical models to the endpoint, our advanced threat prevention engine easily surpasses all traditional methods of malware detection and prevention. Our mission is to stop the execution of bad files before they can cause any damage. With this approach, the endpoint remains secure and unviolated even if the file is resident on disk.

Advanced Threat Prevention by Dell

Dell Endpoint Security Suite Enterprise is our flagship enterprise product that harnesses the power of our advanced threat prevention engine to prevent the execution of advanced threats in real-time on each endpoint in the organization.

Key Features:

- Protection and detection of previously undetectable advanced threats
- Not cloud dependent for sensitive environments
- No daily .DAT updates which eliminates the need for an "always-on" connection
- Extremely low performance impact with runtime execution to dramatically reduce overhead
- Easy to manage with a simple web console

Dell provides real-time detection and prevention of malware. It operates by analyzing potential file executions for malware in both the Operating System (OS) and memory layers and prevents the delivery of malicious payloads. Memory protection is designed to be extremely low-touch as to not incur a heavy performance overhead. Instead, memory protection strengthens basic OS protection features like DEP, ASLR and EMET by providing an additional layer to detect and deny certain behaviors which are very commonly used by exploits.

These two core functions are supported by a variety of ancillary features necessary for enterprise functionality including:

- Whitelist and blacklist support for administrative granularity
- Detect-only mode (audit mode)
- Self-protection (prevention against user tampering)
- Complete control, configurability and compliance reporting from the management console

Dell Cybersecurity Professional Services

We are here to help you with end to end solutions to identify and address security risks in your environment. Our team of security professionals will assist you in finding security risks, implementing solutions, and staying secure. Our suite of services are designed from proven expertise, collaborative efforts, and urgency.

Key services offered include:

- Assessments
- Implementation Services
- Managed Services

Securing customer environments with end to end solutions

Awareness "Capture the Value"	Fortification "Get it Deployed"	Vigilance "Maintain the Value"
Consulting Services • Security Assessments • Project Based Services • Staff Augementation	Implementation Services • DDP Suite • Threat Protection Solutions • SIEM Solutions	Managed Services • Dell Data Protection • Event Monitoring • Incident Handling

Summary

Dell truly believes that mathematical modeling and machine learning are the keys to a secure future. Each product and service we offer is tightly integrated into our advanced threat prevention engine, providing unparalleled accuracy and insight into the modern threat landscape. Best of all, by continuously learning and training based on new data, the Dell engine is truly "future-proof" and will not lose efficiency over time–even as the attackers morph their strategies.

For more information on Dell Data Security solutions that help you protect data and prevent threats, visit Dell.com/DataSecurity.

[1] Files are only uploaded from active computers if the customer chooses to enable this option

Powered by CYLANCE