2017-12-01

A Shift in Malware Protection

Baptista, I

10.1093/itnow/bwx134
ITNOW
Oxford University Press (OUP)

The necessary shift in malware prevention

Irina Baptista, holds a BSc(Hons) Computer and Information Security and Dr Stavros Shiaeles, is a Lecturer in Cybersecurity at CSCAN, Plymouth University as well as Cybersecurity consultant.

Whether it was for improper usage of your device or an unavoidable situation, chances are you have been afflicted with malware at some point in time; new malicious binaries are reported daily, and the current defence mechanisms are struggling to persevere – a better solution is therefore required.

Most security systems use a signature-based approach in which malware is detected and for us, users, to remain protected the system needs to be constantly updated. Although these systems are quick to detect known signatures, if, say, the malware has only been launched recently the odds of it being detected are very low since the signature is still unknown.

Additionally, those security systems have proved insufficient because of the rapid growth and easy adaptation of malware and the methods used to avoid them. Malware writers are constantly exploring ways to improve their creations, take for example a polymorphic malware which constantly changes its underlying code to evade security or a "program" with a dormant functionality to protect its malicious intents.

And it doesn't take an expert to cause panic with a single piece malware. Various tools to create malicious programs or attacks are available online at no cost; anyone with a computer and internet connection available can create the next major attack.

From paid ransoms to shutting down systems, there is no limit to the damage malware can cause. An example of this was the recent attack by the WannaCrypt ransomware which spread across the globe with what is estimated to be over 150 countries affected. Once inside a device the ransomware, which contained a worm in its code, spread over the network and at that point we can say that security defences were defeated because of an unpatched system.

The point is, it is becoming increasingly difficult to protect our devices against security threats with the existing systems, and it is necessary to start addressing malware differently. Otherwise, we risk losing our assets or worst.

Is there a solution?

Possibly yes. Until now, computers have outperformed humans in managing and executing more tasks in a less amount of time, so why shouldn't we give them more responsibility when concerning malware detection? As technology evolves so should its applications and what started as a giant calculator has now become capable of making decisions for itself, and this is known as machine learning.

Machine learning has been designed to give devices the ability to learn from data, – some in a similar way to how our brain works – and this includes the distinction between malware and normal programs. But to make use of its abilities, it is important to decide which of the different algorithms available will achieve our objective with best results.

The first step is to identify what data is relevant and should be fed to the algorithm. Currently, most detection systems using machine learning make use of the Portable Executable (PE) File Format to detect malware. Wanting to shift from the "traditional" approach we ran across binary visualisation.

In case you have never heard of it, this technique can help identify potential anomalies in a file since analysing the binary values provides information about the file structure and aids in the identification of obfuscated data.

How does it work? Simply put, it converts the contents of a file to an image reflecting its binary values. Each pixel will represent a value, and the differences between the pixels will create patterns in the image, which can help identify malicious programs.

Having identified the data, we can now proceed to choose an algorithm which in this case should be capable of spotting patterns in images. Since it is important to detect a threat promptly, thus avoiding the user to execute a malware mistakenly, the system performance needs to be taken into consideration.

There are many algorithms which could fit that description, however, around six years ago, a research team led by Osamu Hasegawa at the Tokyo Institute of Technology developed an interesting robot capable of learning and reasoning. The robot can perform tasks such as pouring a glass of water or preparing a cup of tea, and if it is unfamiliar with a task or object, it gathers information from other robots similarly to how humans transmit experiences.

But what is so important about this robot for our detection system? Apart from other senses, the robot successfully uses images to identify objects and does it with a machine learning algorithm. The algorithm is known as Self-Organizing Incremental Neural Network (SOINN).

If you are not familiar with machine learning, you need to understand that most algorithms require previous knowledge of their network structure, for instance, the amount of data (input) and the number of classes (output) necessary, but SOINN doesn't. Another one of its perks is noise reduction; SOINN can eliminate "insignificant" values to understand the overall pattern of the data.

Having the data and the algorithm, what comes next?

A process called pre-processing or normalisation is done before sending data to the algorithm to reduce redundancy and improve both accuracy and performance. This process extracts key features to classify the data.

In our system at normalisation stage, after the file contents are converted to a picture, the image is divided into 4 parts separating top, bottom and upper/lower middle and the sum of the colours in the RGB (Red, Blue, Green) space are calculated for each part.

As there are 256 possible colour combinations in the RGB space, the combined sums will produce a vector with only 1024 dimensions, which should be significantly smaller than the vector that we initially had (width * height), thus reducing the time necessary to classify a file.

This vector is now ready to be sent to SOINN which will determine whether a file is malicious or not.

How well it performs?

At the moment, the system is showing good results and can correctly detect between 70% and 89% of the files with a higher probability of raising false negatives (FNs) than false positives (FPs). However, without trying to justify one over the other, having numerous FNs can leave the system vulnerable to unauthorised access, infection, data theft or other forms of intrusion.

Nonetheless, it is time to let go of our reliance on antivirus and evolve our methods in the same way that attackers do to face the upcoming challenges of cybercrime.

Boxout

The files were converted to an image based on the online tool binvis.io. Binvis.io (http://binvis.io/#/) is "an interactive tool for binary visualisation" developed by Aldo Cortesi, CEO at Nullcube. The tool provides different visualisation schemes each focusing on a particular feature.