Security

MS Researchers Tackle Automated Malware Classification

By Ryan Naraine

May 11, 2006

Researchers from Microsoft's anti-malware engineering team are working on an automated way to sort through the thousands of malware families and variants attacking Windows computers.

The company unveiled its plans at the EICAR (European Institute for Computer Anti-Virus Research) conference in Hamburg, Germany, proposing the use of distance measure and machine learning technologies to come up with automatic classification of viruses, Trojans, spyware, rootkits and other malicious software programs.

A research paper presented by Microsoft's lead anti-virus researcher, Tony Lee, described the existing process of manual human malware analysis as "inefficient and inadequate" and suggested an ambitious method that combines runtime behavior analysis, static binary analysis and adaptable algorithms to automate classification.

"In recent years, the number of malware families/variants has exploded dramatically…Virus [and] spyware writers continue to create a large number of new families and variants at an increasingly fast rate," Lee said, arguing that automatic malware classification has become an important research area.

He said Microsoft's attempts to automate static file analysis present "considerable challenges" because of the way malware families evolve.

Lee, a graduate at the University of California at Berkeley, said the **dramatic rise in malware prevalence** in recent years has forced the anti-virus industry to change the way the threats are detected, analyzed, classified, described and eventually removed.

"[We believe] that an effective classification method can serve better detection, cleaning and analysis solutions," Lee added.

In a **white paper** co-written with Microsoft program manager Jigar Mody, Lee said the automated process would get around the traditional way in which new malware samples are sorted.

Microsoft says that recovery from malware is becoming impossible. Click here to read more.

"[Today], human analysts classify these samples by memorization, looking up description library or searching sample collection. Human analysis is time consuming, subjective and results in considerable information loss," he said.

Microsoft's proposal will take a "holistic approach" to tackle the classification problem, Lee said, pointing out that the machine learning aspects will deal with everything, from knowledge consumption, representation and storage, to classifier model generation and selection.

It aims to consume knowledge about the malware sample efficiently and automatically and represent that knowledge in a form that results in minimal information loss.

The process calls for the knowledge to be structured, stored, analyzed and referenced efficiently. Once the knowledge of the sample is stored, it can be automatically applied to identify familiar pattern and similarity relations in a given target.

"The process is adaptable and has innate learning abilities," Lee and Mody wrote.

eWEEK.com Special Report: Worm Attacks

Microsoft isn't the only company working aggressively in the automated malware classification field. Halvar Flake, CEO and head of research at Sabre Security, has used the company's BinDiff tool to pinpoint visual evidence of related malware families.

According to data from **Flake's research,** which combined reverse engineering techniques with a clustering algorithm, similar code has been found in the most prevalent malware families.

Flake used 200 malware samples and found that they all related to two large virus families, three small families and two pairs of siblings.

The researchers believe that better classification of malware will help cut through the confusion of naming virus families where anti-virus vendors all append different names to newer threats.

Check out eWEEK.com's **Security Center** for the latest security news, reviews and analysis. And for insights on security coverage around the Web, take a look at eWEEK.com Security Center Editor **Larry Seltzer's Weblog.**