Collective intelligence approaches to malware recognition

Iñaki Urzay, chief technology officer, Panda Security

Iñaki Urzayr

In 2007, PandaLabs received an average of 5000 new strains of mail every day. In total, the amount of malware that appeared last year increased tenfold with respect to 2006 which, in turn, saw the same amount of new malware as in the previous 15 years combined. Put simply, the amount of malware in circulation is increasing dramatically.

The reason for this chasm between reality and perception is that cyber-criminals are no longer trying to cause headline-grabbing epidemics; they are trying to infect systems silently. As part of this strategy to profit from their nefarious activities (theft of bank details, selling email addresses for spamming, adware contracted by third-parties, and so on) they strive to keep users and security companies in the dark regarding new malware samples.

This has resulted in a change in the type of malware used. The number of new bots detected increased by 1800% between 2003 and 2005. Something similar has happened with other types of malware designed for financial profit, such as banker trojans or key loggers. More evidence? Malware that uses stealth techniques increased by over 32 000 percent between 2003 and 2006.

Intelligence in the cloud

A current trend in malware creation is that the binary infecting the user's PC is 'dumb', and the intelligence is 'in the cloud'. The code that resides on the PC has some simple functions that it passes on to a remotely compromised server. The server then returns instructions on what to do. Borrowing the (perhaps overused) '2.0' term from current web trends, we will refer to 'Malware 2.0' as malware that separates its intelligence from its code base.

PandaLabs has reported the '2.0' approach in banking trojans that target individuals to remotely monitor their habits. Based on the online banking landing page and authentication scheme, they inject some type of code to modify the banking session. Known banker trojans such as Limbo/NetHell and Sinowal/Torpig use these techniques quite extensively.

"The number of new bots detected increased by 1800% between 2003 and 2005"

Other '2.0' techniques recently used by malware are server-side compilation, where the web server recompiles a new binary every few hours. Finally, bot nets are using fast-flux DNS networks for improved resistance against take-down efforts. These last techniques are more visible in the recent Storm/Nuwar attacks.

This whole scenario has been dubbed 'the new malware dynamic', and has led to a panorama in which even users with up-to-date security solutions installed can find themselves infected with malware. This is why we talk about a silent epidemic.

As malware has evolved, so have the techniques used to evade detection and hide from prying eyes.

There are several technologies available to combat these hiding techniques. Techniques such as deep code inspection, generic unpacking, native file access, and rootkit heuristics are able to inspect any item as deeply as necessary, even if the item is making use of stealth techniques to remain hidden in the system, and pass on the results to the scanning and monitoring technologies. These intelligent technologies can detect unknown threats by analysing their behaviour, providing an extra level of security.

Within the new malware dynamic, in which cyber-criminals try to hide their creations, a security solution that does not incorporate proactive technologies will not be able to effectively protect a computer.

These technologies act as the last defence line against new malware samples being run on the computer that have bypassed signature protection, heuristic scans and behaviour blocking. While running, they intercept the operations and calls made to the API by each program and correlate the information before allowing the process to be fully run. This real-time correlation allows or denies a process to run depending on its behaviour. If it considers a process suspicious, it blocks and eliminates it before it acts, and prevents it from being run again.

Collective intelligence

Today, there are over ten times more malware variants in distribution than there were two years ago. The obvious conclusion is that a security solution must detect ten times more malware to provide adequate protection to users. While a full-fledged hosted intrusion detection system raises the bar substantially by detecting and blocking most of these, it is still possible for unknown malware to slip through its defences.

We need to consider the fact that, while 80% or 90% of proactive effectiveness is relatively speaking an excellent score, in absolute terms it may lead to hundreds or thousands of malware samples being missed over time, since even a small fraction of a large enough number will still be a 'big' number.

Panda Labs has developed an approach called collective intelligence, which uses community knowledge to protect others, while automating and enhancing malware collection, classification and elimination. It also gains knowledge of techniques to improve existing technologies, while deploying new generation of security services from the cloud.

The shortcomings of PC-centric protection

Traditional security solutions are architected with a PC-centric philosophy. This means that a PC is treated as a single unit in time and any malware detected within that PC is considered separately from the rest of the malware samples detected in millions of other PCs.

Traditional security companies do not have visibility into what PC a particular piece of malware was first seen on. Neither is there visibility of the continuity of that malware's evolution over time in different PCs.

Another significant barrier to reliable PC-centric malware detection ratios is the fact that the process of creating a signature against a single sample takes too long. Each malware sample needs to be sent to the lab by an affected user or fellow researcher, reversed engineered by a lab technician who in turn needs to create a detection signature and disinfection routine for it. These in turn need to be quality-assured, uploaded to production servers, replicated worldwide and finally downloaded and applied by customers.

This entire process is, in most cases, mostly manual and can take up anywhere from minutes, to hours or days or even weeks, depending on the workload of the lab engineers and other factors such as sample priority, prevalence, damage potential, and media coverage.

> ***"A significant barrier to reliable PC-centric malware detection ratios is the fact that the process of creating a signature against a single sample takes too long"***

The process can even be delayed much longer when functional upgrades to the anti-malware or behavioural engines are involved. It is typical of an anti-malware vendor to upgrade its solutions once or twice a year, as each upgrade has a costly testing and deployment process for corporate customers. Ultimately this results in traditional approaches being too slow to combat today's rapidly moving malware.

In a collective intelligence approach, as soon as a malicious process is detected in a user's PC (whether by system heuristics, emulations, sandboxing, or behavioural analysis), the rest of the users worldwide will automatically benefit in real time from that specific detection. This results in a close to real-time detection not only for initial malware outbreaks but also of targeted attacks whose objective is infecting a small number of users to stay below the radar. In a collective intelligence infrastructure, this entire process of malware collection, classification and elimination can be automated and performed online for the vast majority of samples.

Let's walk through the process from the point of view of a computer that has just been exploited and infected by a malicious code.

Automated malware collection

A collective intelligence (CI) agent gathers information of processes and memory objects, and performs queries against the CI central servers which perform a variety of checks against those.

If certain conditions are met, the suspicious file or parts thereof are automatically uploaded to the CI servers, where they are further processed.

Since processes loaded in memory are not subject to many of the cloaking techniques, the agent component does not need to contain a large amount of intelligence and uncloaking routines, and can therefore be very light.

Panda has built a vast database of automatically collected malware samples, which provides the CI web service with areal-time feed of new malware classification entries.

Automated malware classification

Server-based processing is not limited by the CPU and memory constraints of personal computers. Therefore, scanning routines at the CI servers undergo much more in-depth processing by more sensitive technologies (signature and sensitive heuristic scanning, emulation, sandboxing, virtualisation, white lists, and so on) to reach a final classification.

It is important to note that the scanning power used by CI servers is only limited by hardware and bandwidth scaling, unlike a typical scenario at a PC, desktop or server. This means that more computationally intensive malware detection and analysis techniques with higher detection rates can now be used without touching customers' CPU and memory resources. With this approach the majority of new malware samples can be analysed and classified automatically in a matter of minutes.

The CI servers are managed by PandaLabs, and therefore samples that cannot be classified automatically are ultimately looked at by an analyst at the lab.

Automated malware elimination

The elimination module of the CI is in charge of automatically creating detection and disinfection signatures for the samples previously analysed by the processsing and classification module. These signatures are in turn used by the community of CI users to proactively detect and disinfect new or even targeted attacks with very low numbers of infected hosts.

The traditional anti-malware and hosted intrusion prevention solutions have also started to benefit from the CI approach. During the initial three months, the elimination module has created protection for a few hundreds of thousands of malware samples, which have been gradually deployed to our existing products.

One of the main benefits of the collective intelligence approach is that these signatures do note need to be downloaded to each client as they operate from the cloud. This however, does not mean that the client machine will not need to maintain updated signatures.

A potential threat to such an approach is the availability of the CI servers. However, our approach for integration of the Collective Intelligence technology on current solutions is designed as an additional layer of protection. Therefore, under non-availability of the platform for whatever reason, security protection would fall back to the regular hosted intrusion prevention solution, which provides well-above-average protection.

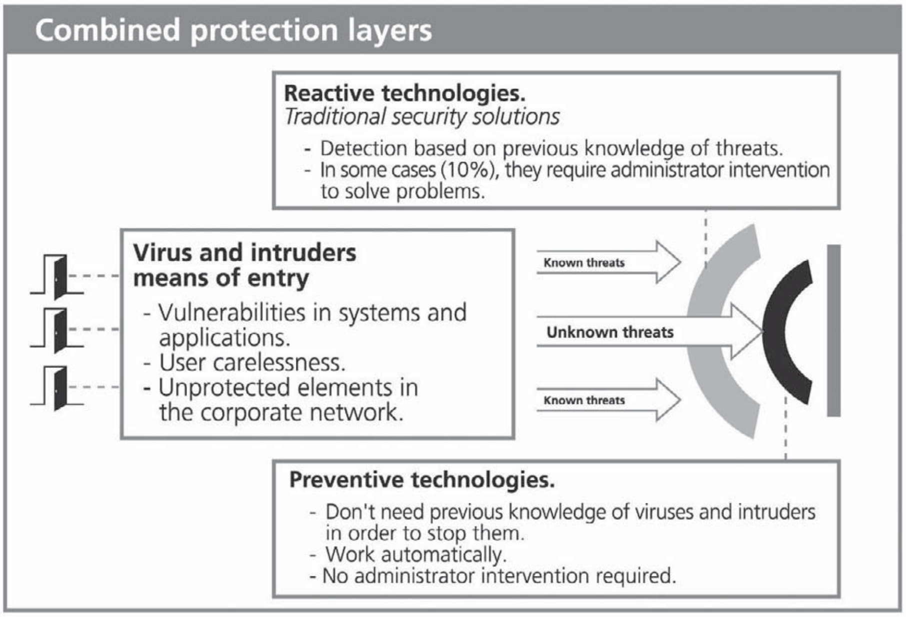
Combined protection layers
Reactive technologies.
Traditional security solutions
- Detection based on previous knowledge of threats.
- In some cases (10%), they require administrator intervention to solve problems.
Virus and intruders means of entry
- Vulnerabilities in systems and applications.
- User carelessness.
- Unprotected elements in the corporate network.
Known threats
Unknown threats
Known threats
Preventive technologies.
- Don't need previous knowledge of viruses and intruders in order to stop them.
- Work automatically.
- No administrator intervention required.

Figure 1: Threat types and protection layers.

Gaining knowledge on malware techniques

Another main benefit provided by the community feature of collective intelligence is that of giving insight to our engineers about new malware techniques and entry points. Questions such as where a specific piece of malware was first found, and how it spread, allow us to model additional intelligence into specific malware families and even creators of specific malware variants.

This approach of applying data warehousing and data mining techniques to malware detections by the community provides significant knowledge on how malware and targeted attacks are carried out. The type of knowledge that can be gathered using this approach becomes especially useful if it can be applied for tracking infection origins, which in turn might have some interesting applications and benefits for law enforcement efforts.

Deploying security services 'from the cloud'

We have developed and are currently distributing some services that are exclusively based on a collective intelligence platform. These online services are designed to carry out in-depth audits on computers and detect malware that has gone undetected, and security solutions installed.

> ***"Applying data warehousing and data mining techniques to malware detections by the community provides significant knowledge on how malware and targeted attacks are carried out"***

Traditional security solutions can be effective, but not all antivirus products are the same, as they do not detect the same number of threats or the same malware samples. Therefore, the level of protection varies with each antivirus solution. As a result, many computers are infected even though they have an updated antivirus solution installed. This model allows Panda to protect its customers without increasing resource consumption.

So, collective intelligence is Panda's response to the current malware situation. The basic premise may not be entirely new. For example, the Vipul's Razor algorithm uses similar collective intelligence techniques, but uses user input and applies to spam email. Nevertheless, as malware becomes more complex and stealth-driven, using community resources to help identify and classify it will become an increasingly important concept.

About the author

Inaki's interest in IT security arose in 1988, designing and developing the first antivirus products for AmigaDos operating systems. In 1990, he founded the IT company Bitmap Multimedia, dedicated to developing medical software. In 1996, he joined Panda Software, working in software analysis and programming. He later managed software development teams until 1998, when he took charge of the R&D division. In 2002, he became chief technology officer at Panda, managing the research and innovation area, known as Panda Research.

Iñaki Urzay holds a degree in Computer Science from the University of Deusto, Bilbao.