PROCEEDINGS OF SCIENCE

Cloud-Based Anti-Malware Solution

Ismail Adel AL-Taharwa[1]
National Taiwan University of Science and Technology
Taipei, Taiwan
E-mail: tahrawee@yahoo.com

Albert B. Jeng
Jinwen University of Science and Technology, National Taiwan University of Science and Technology
Taipei, Taiwan
E-mail: albertjeng@hotmail.com

Hahn-Ming Lee
National Taiwan University of Science and Technology
Taipei, Taiwan
E-mail: hmlee@mail.ntust.edu.tw
Institute of Information Science, Academia Sinica
Taipei, Taiwan
Email: hmlee@iis.sinica.edu.tw

Shyi-Ming Chen
National Taiwan University of Science and Technology
Taipei, Taiwan
E-mail: smchen@mail.ntust.edu.tw

Abstract—In this work we focus on cloud-based malware detection. We investigate the existing academic and industry cloud-based malware detection solutions, and identify the drawbacks of those solutions. We also recommend a remedy for the drawbacks of those solutions. At the end we provide a summary of our proposed cloud-based anti-malware solution.

The International Symposium on Grids and Clouds and the Open Grid Forum
Academia Sinica, Taipei, Taiwan
March 19 - 25, 2011

[1] Speaker

1. Introduction

Malware means malicious software, which is written to figure-out and exploit the vulnerabilities of systems. It is worth mentioning that many people use virus and malware interchangeably. Malware construction has shifted from the work of novices to a commercial and financial lucrative enterprises [10]. The development in the malware industry goes beyond the imagination of security experts, which is proven by the exponential increase in the number of malware attacks detected every year, especially in the last few years. In 2008, Symantec created over 1.6 million new signatures in addition to the existing six hundred thousand signatures in 2007 [12]. Researchers have cited many causes for this evolution, here we group these causes in two categories. The first one is related to the anti-virus software vulnerabilities, while the second one is related to the evolution in malware indusrty. Next we will discuss these issues in detail. We will refer to anti-virus software by AV.

1.1 Anti-virus Software Vulnerabilities

Host-based AV is the most widespread malware protection solution, which is based on installing a detection agent in the end-user's machines, which keeps updating the signatures of the new malware attacks up-to-date to ensure a complete and through protection. This protection paradigm consists of two main entities, the first one is a group of end-user agents, and the second one is the corporation front on the Internet. For simplicity we will refer to the latter as the detection engine. The detection engine supports the user agents through the information about the new attacks on the Internet and the way to detect these attacks. Detection can be either reactive by comparing the signatures or proactive by verifying behavioral information. End-user agents scan their systems to detect potential existence of malicious software using their detection knowledge accumulated by the detection engine updates.

AV has long been a satisfactory protection solution until a decade ago. Since the beginning of the last decade, the host-based AV began to cause problems and resource burdens to the end-user host. Some of these burdens and problems are (1) End-user's heavy resources (e.g., processing, memory, storage and bandwidth) consumption, Yan W. [13] mentioned three main reasons behind the "slowing-down" of machines that installed AV products other than long signature files; (2) AV imposes tangible overhead on the network performance, especially for the private network environments, such as campuses, organizations, ministries and many others; (3) Complexity of this kind of software is increasing, which brings many leaks for attackers [10]; (4) AV decreases productivity, especially when performing system full scan or using emulators to trace the behavior of suspicious files; (5) AV suffers from subtle defects, missing many kinds of new attacks (e.g., zero-day attack), and making many false positive detections [10], [1].

1.2 Malware Industry Evolutions

The malware industry has evolved from curious hackers to profit motivated attackers, which resulted in the surge of malware creation that challenged host-based AV solutions. Some of the probable causes behind this evolution are (1) Availability of online malware generators,

(2) Availability of toolkits that facilitate generation of new variants given a malware instance with the minimum programming experience [5], [11], (3) Attackers interest in exploiting the abundant strategies and tools created to protect programs and software developer's copyrights (e.g., Packing, and obfuscation) to evade the detection of their assaults. [6], (4) Funded organizations and forums to train novices and advanced attackers how to create attacks, exploit vulnerabilities in targeted victim's machine, and avoid detection mechanisms. Those and many other causes resulted in exponential increase in the number of newly created malware. Panda Security has detected more samples in year 2008 than the previous 17 years all together [13].

Figure 1 summarizes the relationship between these causes and their effects on the end-users' systems. For example, whenever a host-based AV provider brought out complex solutions to encounter the new threats, then an attacker would create new ways to evade these AV detection techniques. However, the more complexity the AV has, the more vulnerable it becomes; this is illustrated by points 5, and 8 in Figure 1.

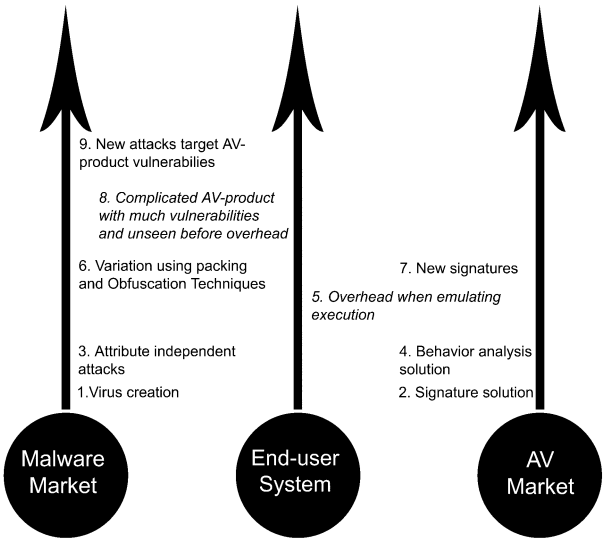

Figure 1: Malware market versus AV production and their impact on the end-user

In this work we study the impact of cloud computing on the malware detection. Specifically, we compare cloud-based anti-malware solutions from both industrial and academic perspectives. The rest of this paper is organized as follows. Section 2 surveys conventional anti-malware solutions, and the relationship between cloud computing and malware. In Section 3, we discuss industry solutions. The academic solutions are discussed in Section 4. In Section 5, we provide our suggestions and recommendation to remedy the drawbacks in both academic and industry solutions. We conclude this paper in Section 6.

2. Background

In this section we survey the background of anti-malware solutions and the impact of cloud computing on them.

2.1Malware Attacks and their Conventional Countermeasures

Two kinds of anti-malware solutions have been used in the literature. The first and the most pervasive one is the signature based detection solution. Signature can be the file name, the file size, the file digest or a combination of them. Sometimes it may be rule-based detection policy [6], [2]. The second solution is behaviour based systems. The motivation behind this solution is the possibility to generate variants of the given exploit by modifying it. Such kinds of modifications include but not limited to obfuscation, changing file name, packing PE format files and applying some code optimization attitudes [6], [4]. During the second half of 2006, Microsoft found more than 45,000 new variants of backdoors, trojans, and bots [8]. Moreover, many toolkits can provide malware writers with regularly updated exploit codes and enable them to generate variants of the given malware. Metasploit is an example of such tools [3].

2.2Cloud Computing and Malware Detection

Researchers who study cloud computing and malware protection can be divided into two groups. The first one aims to protect cloud infrastructure and users from new attack strategies and models that exploit drawbacks and risk probabilities in the cloud. The second one aims to extend the services of traditional anti-malware solutions by integrating them as cloud services. In this work we focus on the second group who are motivated by the opportunities that cloud can provide to the malware detector, such as solving the problems of constrained resources machines, minimizing the wasted network resources, providing in-time protection, minimizing the vulnerability window, and facilitating forensics & retrospective detection [10], [9], [6].

Academic researchers focused on the computational capabilities of cloud, which can minimize overhead on the end-users machines, While, industry people exploit cloud computing capabilities in addition to their already existing solutions to facilitate intelligent knowledge extraction from their numerous users' data. In this work we survey a group of academic views, and two representative cloud-based industry AV solutions. Academic views include (1) CloudAV [10]; (2) Behavior-Based Malware Analysis in the Cloud [7]; and (3) Retrospective Detection of Malware [6]; while the industry solutions are taken from two distinct AV-provider companies including: (1) Trend Micro; and (2) Panda Security.

3.Industry Views

Most antivirus corporations started to advertise and deploy their cloud solutions in 2008. They claimed to provide better detection coverage and to minimize overhead in their client machines. Since most of these companies do not provide technical details about their implementations, we are going to rely on the best available public domain information to make a subjective qualitative judgment of their value-added services. We choose two AV providers which provided cloud-based services earlier than their competitors. The first one is Trend Micro which is a well-known anti-virus provider. The second one is Panda Security which provides a free tool with limited capabilities for Internet users. Table I shows these two companies, their corresponding cloud- based AV products, and their underlying techniques.

Table I

Industry AV cloud-based products and technologies

Corporation name	Products	Underlying Technologies
Trend Micro	1. InterScan Messaging security Virtual Appliance, 2. Trend Micro Smart Protection Network, 3. Trend Micro OfficeScan 10 with File Reputation	1. Web, and file reputation service, 2. Machine learning techniques, 3. Automatic Malware Signature Discovery System (AMSDS)
Panda Security	1. Malware Radar	1. Expert systems, 2. Rule-based policies, 3. Machine Learning techniques

3.1Technical Details

3.1.1Trend Micro

Trend Micro first deployed Smart Protect Network to integrate and analyze reactions on scanned files among their users by collecting tremendous amount of data that were used to generate trusted heuristic information about suspicious files and web pages on the Internet. InterScan Messaging Security and OfficeScan 10 use this information in different ways. InterScan Messaging Security scans suspicious files and Web-links instantly without bothering their clients when they are reading their emails. The scanning process is shared between the cloud engines that remove and heal high percentage of those suspicious messages, and the on-premise VMware Ready Virtual Appliance which provides each enterprise both the fine level of policy enforcement they require and the privacy they prefer. This mechanism preserves clients' privacy. On-premise protection is achieved by incorporating three modern anti-spam solutions together, namely, (1) Directory Harvest Attack (DHA) protection; (2) Granular content filtering controls; and (3) Adaptive technology using machine learning techniques.

OfficeScan 10 with File Reputation scans all files injected into clients' machines regardless of their source with minimum processing overhead. This product uses the same client agent's and cloud server's architecture, however in a more sophisticated manner. Here client agents do not need to download the whole ineffective signature updates any more. Instead, Trend Micro file reputation solution decouples pattern files from the scanning engine and conducts pattern file lookups over the network to a Smart Scan Server. Figure 2 illustrates the architecture of OfficeScan 10 with File Reputation as shown in the white paper of Trend Micro. This architecture consists of two main parts: (1). File Reputation and (2). Automatic Malware Signature Discovery System (AMSDS) [14]. AMSDS is represented by the communication patterns between the smart query filter and smart scan server entities.

Figure 2: Trend Micro Smart Scan Server architecture

File Reputation leverages the anonymous software usage patterns of millions of Trend Micro users to automatically identify new threats and informs all Trend customers about those threats in the shortest time. This solution exploits the availability of huge amount of data and machine learning techniques to discover attack patterns. Moreover, this mechanism removes the complexity of the AV product from the users' machines which minimize their vulnerability risks. AMSDS represents the long size signatures by "hashed" patterns which are hundred times shorter [14] but sufficient to determine whether the host machine under consideration matching the attack pattern or not. In such situation, the client agent (i.e., Smart Query Filter) will access the updated information in the smart scan server to perform exact matching with the suspicious pattern. There are two main issues related to this scenario. First, the client agent does not query the cloud service for every single file. Instead, it tries to determine with a high degree of accuracy whether the file under scanning can be detected by using the actual pattern file. Second, Trend clients do not need to transfer their suspicious files over the cloud. They ask for the corresponding actual signature files, which is important to preserve their privacy.

3.1.2Panda Security

Panda Security came out with an interesting solution which integrates knowledge extracted from community to their already existing collective intelligence solution and exploits cloud services to improve the performance of this solution. Actually, they employed their earlier products to collect massive information and data about attack patterns and behaviours by simply transferring any suspicious file undetected by their installed security products to the scanning labs. Collective Intelligence technique played the fundamental rule of collecting behavioural patterns, file traces and new malware from transferred data. They applied many artificial intelligence techniques to the data, such as: Expert system and machine learning to create PandLabs' malware knowledge database and eventually transfer these information to the end-users either in signature files or as a web service delivered by the cloud. Client agents send suspicious files to the cloud, where cloud emulates their execution to determine their real behaviour. Moreover, cloud correlates the results of such files with other suspicious files from other users, to provide a comprehensive and complete behavioural examination.

3.2Effective Comparison

Frankly speaking, all those industrial corporations lure their customers by concealing the underline technologies used in their products, which make end-users unaware for the drawbacks and shortcomings of those companies' products. Table II lists some of those serious issues and compares them between Trend Micro and Panda Security cloud-based AV solutions.

Comparison provided in Table II does not reflect the difference between AV-cloud products and their corporation traditional products. Instead, it reflects the difference between their AV-cloud products actual performance and the advertised performance as documented by their corporations. For example, Trend Micro AMSDS solution requires higher processing resources overhead than advertised. However, this overhead is still lower than traditional products overhead. SplitScreen [2] is an interesting anti-virus framework that replaces the long signature files with shorter patterns and uses bloom filters to perform slight comparison with end-users' files. When a suspicious file is detected, it queries its corresponding full signature

from the anti-virus server. This solution tolerates the host side to make false positive detection only, which can later be removed when doing the exact matching using the full signature. Minimizing the number of the false positives is a preferred target; in order to achieve this goal, SplitScreen [2] requires more processing computations in their end-users machines. No matter whether Trend Micro AMSDS solution uses the same technique or not, it will resemble the same trade-off problem between processing overhead and other resources consumption, which should be analyzed carefully.

Table II

Comparison between trend Micro and Panda security AV cloud-based solutions (✗✗✗ means excellent performance, ✗✗ means good performance, and ✗ means OK performance) with respect to the given criteria

Issue	**Trend Micro**	**Panda Security**
Storage and memory optimization in the end-user machines'	✗✗✗	✗✗
End-user network bandwidth consumption	✗✗✗	✗✗
Bandwidth consumption on the corporation front	✗✗	✗✗
Processing time	✗	✗✗✗
Data Privacy preservation	✗✗✗	✗
Flexibility to end-users security preferences	✗✗✗	✗✗
Forensic tracing support	✗✗✗	✗✗✗
Retrospective detection support	✗✗	✗✗✗

4.Academic Views

Academic people are making hard efforts to exploit cloud computing capability for better malware detection; even those works are few and mostly limited to their organizations and universities resources, they provided a fertile truth ground for effective research in the future. As we mentioned earlier in Section 2, we'll focus on three of those works. Next, we summarize these works concisely. We'll just focus on their main ideas, and architectural components.

1) CloudAV: N-Version Antivirus in the Network Cloud [10]: The basic idea of this work is based on exploiting the detection capabilities of multiple, heterogeneous detection engines that located in the cloud environment, and using a lightweight end-user agent to transfer suspicious files to the cloud to be checked by all scanning engines there, in parallel.

2) A framework for behavior-based malware analysis in the cloud [7]: It proposes a new framework for behaviour-based analysis that allows end-users to delegate security labs in the cloud for the execution and analysis of their suspicious programs, and makes those programs behave as if they were executed directly in their original end-user environments'. Client agents and security labs in the cloud are the main components of this framework. However, security labs in this solution need to communicate with original client systems when interrupting system calls.

3) Retrospective Detection of Malware Attacks by Cloud Computing [6]: Liu et al. proposed a model to clean up infected machines by monitoring their Portable Executable (PE) files creation/written operation logs, sending these logs to cloud which index the information about PE files operation at the first time according to their local machines, and adding one more

indexing according to the relationships among PE files. This work was evaluated by a prototype which implemented using MapReduce functionality provided by Hadoop project. Actually, this solution composed of two components, client agents and indexing servers. However, here we have two indexing servers (1) File indexing server; and (2) File relations indexing server.

These three solutions use the similar basic architecture (i.e., lightweight agents and powerful cloud servers). However, each one of them has distinct approach and contribution, which imposes different requirements on the end-users' machines. Table III shows the pros and cons of these solutions.

Table III

Pros and Cons of cloud-based anti-malware academic solutions

Technique	**Pros**	**Cons**
CloudAV	1. Multiple heterogeneous detection engine 2. Tuneable parallel processing 3. Enhanced forensic capabilities 4. support retrospective detection	1. Works for private networks 2. Exposure of enterprise data 3. lack adaptive integration among the different detection engines
Behavior-based analysis	1. Resembles the exact execution behavior of the desired system 2.leverage tracing of all possible execution paths 3. protect user machines' from potential attacks	1. Compromise users sensitive data 2. No guarantee that end-users will execute such suspicious data 3. Memory and CPU overhead are proportional to the number of system calls
Retrospective detection	1. Reliable detection performance 2. Overcome some kinds of evasion techniques (e.g., obfuscation, and packing)	1. Need long time to complete file relation indexing 2. Compromises users' privacy

4.1Discussion

All previous academic solutions have the same defect. Namely, all of them rely on the traditional detection engines, and work as consolidating solutions to the conventional host-based anti-virus solutions. Next we discuss some drawbacks of these solutions from the end-user perspective.

CloudAV: It is constrained to run as an extra layer of protection in addition to the already existing host-based anti-virus solution. That means this solution will not improve resources utilization of the end-users' machines. Instead it will pose extra network bandwidth overhead. Moreover, this solution is implemented for organizational network environment. Consequently the majority of computer users (i.e., Home users) will not benefit from it. Nevertheless, this solution improves the detection capabilities of end-users' machines. It requires them to send suspicious files to be scanned in the cloud which violates those end-users' privacy. Also, those commercial detection engines in the cloud will have to pay periodic licensing charge.

Behaviour-based and retrospective detection: Both provide frameworks for limited regions of malware detection. The former focus on the behaviour-based detection but it cannot work apart from signature-based solutions [2], while the later focus on determining which hosts or users open similar files once a threat is identified. Moreover, Liu S.-T [6] stated that "although cloud computing is suitable for processing a small number of huge files, it has shortcoming in dealing with a large number of small files". The proposed behaviour-based framework [7] exchanges system calls between security labs in the cloud and users machines, which is supposed to generate extensive overhead on the network, and degrade the execution

behaviour of the considered program. Also, the execution of these files will depend on the availability of interested users. One thing missing in this work is how to ensure that the suspicious file sent to distinct behaving systems and environments, and furthermore how to force those systems' users to run such files.

5.Recommendation

Sophistication in malware industry increased day by day. Those people who created such attacks are not doing this for fun anymore. Some of them are motivated by financial gains; while others want to compromise their competitors. Most anti-virus providers are aware of these situations and they are seriously overworking to resolve drawbacks in their existing systems, which is evidenced by their move toward cloud services. Section 3 shows a faster and a better cloud adaption by private industries compared to the academic researches as discussed earlier in Section 4. However, there are some good practices adopted by academic people, such as: (1) Cumulative detection engines can improve detection capabilities significantly; (2) Cloud structure can foster behaviour analysis by leveraging different users' actions. Also, industrial solutions provided practical guidelines that enrich this area of research. Some of these guidelines are (1) The idea of replacing long signatures by smaller pattern representation; (2) Information integration resulted by executing suspicious files in different systems; and (3) Storing logging information by end-users, which can facilitate both retrospective detection and forensic capabilities. We recommend that any cloud-based solution should meet the following criteria to qualify itself as an applicable solution:

1. Protect end-users privacy.
2. Minimize overhead on the end-user systems.
3. Be applicable for both enterprises and stand-alone end-users systems
4. Support different levels of protection according to the end-users preferences.
5. Provide more than a single detection engine in the cloud.
6. Provide organizations different levels of user privacy to monitor their clients according to their application type.

5.1Remediation

Cloud-based anti-malware should have its own detection paradigm, which supports stronger and faster detection capabilities. In addition to the forensic and retrospective detection with minimum end-users resources consumption, we suggest to use a novel architecture that dedicates distinct cloud resources to monitor each possible violation loophole, then apply the corresponding detection technique and use artificially appropriate detection engine to scan suspicious files. Moreover, cloud should have specific correlation engine that is able to correlate information extracted from different users and to analyze information extracted from the detection engines. To be acceptable to the general users, this solution should support granularity either by providing varied levels of constraints and limitations that allow end-users to preserve their data privacy and minimize overhead on their systems when needed, or to choose aggressive protection constraints to prevent any potential attack attempt depending on the end-user preferences.

6.Conclusion

In this paper, we studied both academic and industry cloud-based anti-malware solutions. This work aims to support and foster the ongoing research and development toward cloud-based antimalware solutions. Our contributions are (1) Comparing two representative industry solutions that adapted cloud computing, (2) Investigating current academic research in the area of cloud-based anti-malware detection, and providing comparison among those proposed solutions in terms of their pros and cons, and (3) Consolidating the good practices from both industry and academic solutions to recommend and enhance the cloud-based anti-malware solutions. At the end, we brief our proposed solution and give recommendations for researchers who are interested to proceed in this area of research.

Acknowledgement

This research is supported in part by the National Science Council of Taiwan under grants number NSC 99-2218-E-011-018 and NSC99-2218-E-228-002.

References

[1] AV-comparative. On-demand detection of malicious software. Technical report, February 2010.

[2] S.K. Cha, I. Moraru, J. Jang, J Truelove, D. Brumley, and D. G. Andersen. SplitScreen: Enabling efficient, distributed malware detection. In *proc. 7th USENIX NSDI*, San Jose, CA, pr 2010.

[3] K. M. David Maynor. *Metasploit Toolkit for Pentration testing, Exploit Development, and Vulnerability Research*. Syngress Media Inc, 2007.

[4] L. Z. ED Skoudis. *Malware: Fighting Malicious Code*. Prentice Hall PTR, Nov 2003.

[5] F-secure. Silent growth of malware accelerates, 2008.

[6] S.-T. Liu and Y.-M. Chen. Retrospective detection of malware attacks by cloud computing. In *2010 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 510-517, Oct 2010.

[7] L. Mortignoni, R. Paleari, and D. Bruschi. A framework for behavior-based malware analysis in the cloud. In *proceedings of the 5th International conference on Information Systems Security*, ICISS '09, pages 178-192, Berlin, Heidelberg, 2009. Springer-Verlag.

[8] Microsoft. Microsoft security intelligence report (july to December 2006). Technical report. May 2007.

[9] I. Muttik and C. Barton. Cloud security technologies. Information Security Technical Report, 14(1), 1-6, 2009.

[10] J. Oberheide, E. Cooke, and F. Jahanian. CloudAV: N-version antivirus in the network cloud. In *proceeding of the 17th USENIX Security Symposium*, San Jose, CA, July 2008.

[11] G. Ollmann. The evolution of commertial malware development kits and colour-by-numbers custom malware. *Computer Fraud & Security*, 2008(9):4-7, 2008.

[12] Symantec. Symantec global internet security threat report, 2009.

[13] W. Yan and N Ansari. Why anti-virus products slow down your machine? In Proceeding of 18th International Conference on Computer Communications and Networks 2009. ICCCN 2009. Pages 1-6, Aug 2009.

[14] W Yan and E. Wu. Toward automatic discovery of malware signature for anti-virus cloud computing. In complex Sciences. Springer Berlin Heidelberg, 2009.