Published in Proceedings of Fourth International Conference on Internet Monitoring and Protection, 2009. pp 66-71
http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5076351

Where Only Fools Dare to Tread: An Empirical Study on the Prevalence of Zero-day Malware

Håvard Vegge, Finn Michael Halvorsen and Rune Walsø Nergård
Department of Telematics, Norwegian Univeristy of Science and Technology (NTNU)
NO-7491 Trondheim, Norway
{havardv, finnmich, runewals}@stud.ntnu.no

Martin Gilje Jaatun* and Jostein Jensen
SINTEF ICT
NO-7465 Trondheim, Norway
{Martin.G.Jaatun, Jostein.Jensen}@sintef.no

Abstract

Zero-day malware is malware that is based on zero-day exploits and/or malware that is otherwise so new that it is not detected by any anti-virus or anti-malware scanners. This paper presents an empirical study that exposed updated Micsosoft Windows XP PCs with updated anti-virus software to a number of unsavoury Internet software repositories. A total of 124 zero-day malware instances were detected in our experiment. Our conclusion is that if a user is sufficiently adventurous (or foolish), no anti-virus protection can prevent a zero-day malware infection.

1. Introduction

IT administrators are constantly fighting to keep their systems patched and updated, while malware authors keep churning out more and more malware every day. The time from when a vulnerability is detected by "the good guys" until exploit code is available keeps shrinking, but the deluge of new malware that do *not* rely on new exploits or other fancy mechanisms ensure that there is also a growing lag between the discovery of a new malware specimen and the time of generally updated virus signatures for this malware.

Zero-day is a broad term and can be applied to various areas of information security. Often people associate the zero-day with software vulnerabilities which are not known to the public, and the creation of zero-day exploits. This paper is focused on zero-day malware, that is, malicious software which is not detected by anti-virus programs due to lack of existing virus signatures or other malware detection techniques. Zero-day malware can also – but does not necessarily have to – be based on zero-day exploits.

Although the concept of zero-day exploits (and malware) has been around for years, no major studies or scientific articles seem to have been published on this topic. Most of the related literature available consists of loose web articles with limited details.

1.1. Background

There is no doubt that files from file-sharing networks represent a great risk. According to a study of malware prevalence in Kazaa by Shin et al. [1], 15% of 500,000 downloaded files were infected by malware. Kalafut et al. [2] found that in over a month of data, 68% of all downloadable responses in LimeWire/Gnutella contained malware. In a study by Berns and Jung [3], 70 out of 379 downloads from BitTorrent sources contained malware (18.5%).

Many companies or web sites test different anti-virus software on a regular basis. Two of the biggest actors in this area are AV-Comparatives.org and AV-Test.org. As such companies are comparing anti-virus vendors, their methodology is not the same as in this paper where we are searching for zero-day malware. Still, a proactive/retrospective test performed by AV-Comparatives [4], can give indications of what results to expect. A retrospective test is used to test the proactive detection capabilites of scanners. It gives an idea how much new malware a scanner can detect (for example by heuristic/generic detection), before a signature is provided for the malware.

According to SANS Institute [5], all operating systems and all software applications are vulnerable to zero-day vulnerability discovery and exploitation. This paper is limited to detection of zero-day malware threatening Windows XP and Internet Explorer 7.

*. Corresponding Author

1.2. Paper Outline

The rest of this paper is structured as follows: In Section 2 we describe the method and preparations for our experiment, and in Section 3 we describe how the experiment was carried out. We present our results in Section 4 and discuss them in Section 5. Section 6 concludes the paper.

2. Method

An outline of the method can be seen in Figure 1. The preparations before the exposure phase included setting up a lab environment directly connected to the Internet, and installing operating systems and anti-malware packages on the laboratory computers.

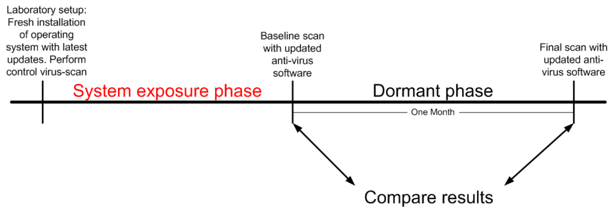
Laboratory setup: Fresh installation of operating system with latest updates. Perform control virus-scan
Baseline scan with updated anti-virus software
Final scan with updated anti-virus software
System exposure phase
Dormant phase
One Month
Compare results

Figure 1. Method to find zero-day malware

The computers were installed with Windows XP, and the latest service pack (SP3) was applied from CD. The computers were not connected to the Internet until proper anti-virus software was installed, and then the first thing we did was to install the latest updates from Windows Update.

Each machine in the laboratory was set up with different anti-virus software (see Table 1), which all automatically updated themselves with the latest virus definitions when they detected an Internet connection. The main purpose of the desktop anti-virus programs was to avoid having our computers infected with already known malware. All the software was installed with default settings, but some changes were made to make the different anti-virus software as similar as possible. All the anti-virus software was set to quarantine infected files if possible.

Computer name	Anti-virus
Gustav	Norman
Ivan	Norton
Katrina	F-Secure
Mitch	Avast!
Andrew	AVG

Table 1. Computer overview.

We installed the popular and free Spybot Search & Destroy[1] on all machines to protect against spy- and adware. While we did not originally intend to do this, it became apparent that it was necessary to avoid the machines becoming so cramped up with spy- and adware that they would be practically unusable for the intended activity.

3. Procedure

As the time schedule below shows, we actively exposed the computers to suspicious web sites and file-sharing networks during a period of two weeks. The computers were then shut down for about a month, before they were turned on in the beginning of November 2008 to perform anti-virus scans and analyses.

September 10th: All computers in the laboratory were connected to the Internet

September 15th: Control virus scans and experiment start-up.

October 1st: Baseline virus scans before computers were shut down for a month.

November 3rd: Final virus scans.

3.1. System Exposure

Monday September 15th 2008 we actively started to expose the computers to web sites, file-sharing systems, etc. We had prepared an initial list of suspicious web sites containing warez, screensavers, codecs, mp3s and other free downloads. The actual number of visited web sites ended up being much larger, as we clicked on many advertisements and visited partner sites. Since web sites from Romania, Hong Kong and Russia were considered most risky by the computer security company McAfee, Inc. [6, 7], we tried to include some sites from those countries as well.

As seen in Table 2 we had also come up with a list of search keywords, which we applied when using file-sharing programs. The list was compiled from the names of the 50 most popular Windows downloads at Download.com [8] on September 15th.

As the timeline in Figure 2 indicates, the system was exposed to web sites and file-sharing networks over a two week period. Some days were spent on one source only, while other days consisted in the use of several sources. *Install* indicates that the same downloaded files were installed on all the computers. The time slots containing *X* means nothing was actively done to expose the computers, but they were still connected to the Internet and both Internet Explorer and the file-sharing clients were running.

The same actions were performed on the five computers in the laboratory almost simultaneously, in order to facilitate as fair as possible comparisons. The following sections describe the procedure and actions that were taken in Limewire, μTorrent, Internet Explorer and the anti-virus software, respectively.

[1] http://www.safer-networking.org

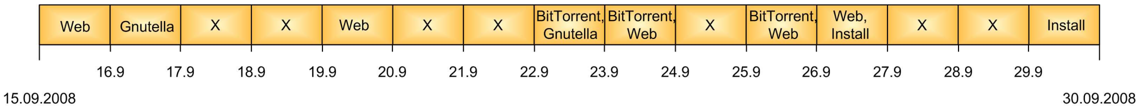

Figure 2. System exposure timeline.

3.1.1. File-sharing Networks. File-sharing networks are known to be a significant source of malware [1], and therefore it was important to expose the experiment to these networks. We used the keywords in Table 2 to search for candidate files.

For the Gnutella network we chose the client Limewire[2], which is the most popular client for this network [9]. Limewire was installed with default settings, but sharing of files was disabled to avoid any legal issues and excessive network traffic. The most popular client for BitTorrent is μTorrent[3] and is well suited for our laboratory as it is lightweight and easy to use.

avg, antivirus, adaware, limewire, frostwire, winrar, winzip, mirc, irc, player, real, media player, zip, free edition, youtube, downloader, irfanview, google, chrome, adobe, firefox, virtualdj, vlc, iso, cleaner, msn, live, nero, divx, spyware, torrent, activex, flash, trillian, norton, mp3, 2008

Table 2. Keywords used for data collection.

3.1.2. Surfing the Web. More and more people get access to the Internet, but it can not be regarded safe, even though you try to avoid obviously suspicious web sites. According to Provos et al. [10], approximately 1.3% of the incoming search queries to Google's search engine return URLs labeled as malicious. In order to expose the laboratory computers to a wide range of threats, it was then natural to visit some potentially risky web sites.

A great deal of all web sites contain adware, viruses and other threats. McAfee has published two reports that show which domains that are most risky [6, 7]. Based on these reports and the use of search engines like Google and Yahoo! together with popular search phrases, we came up with a list of possibly malicious sites. The list is a mixture of popular ordinary web sites and sites which claim to serve downloadable items, such as warez, screensavers and mp3s.

When visiting web sites, the integrated browser Internet Explorer 7 was a natural choice. If plugins like Flash etc. were missing, they were installed on demand.

The following strategy was followed:

- We basically started at the top of our list and visited the web sites one by one.
- Since our goal was to be exposed to as much malware as possible, we acted like a foolish person, uncritically clicking *OK* to everything that popped up.
- If the particular web site had partner sites or other tempting links, we paid them a visit too.
- When visiting warez sites, where it was possible to download for instance software, we typically chose a few of the most popular items and saved them to a directory on the computer for later analysis.

The same procedure was performed on all our computers almost simultaneously. Still we experienced that different pop-up windows showed up on different computers, so minor dissimilarities occurred.

3.2. Offline Search

In addition to the installed anti-virus packages, we obtained two offline anti-virus programs, F-PROT and Avast! BART CD. F-PROT is available free of charge, while Avast! BART was obtained through a trial license. By offline we mean that the host operating system is not booted, the anti-virus software is run from a live CD.

As noted in the time schedule at the beginning of Section 3, offline scans were performed three times during this project.

1) A control virus scan was performed before starting the exposure phase, to verify that the laboratory computers were clean.
2) A baseline scan was performed after the two weeks of exposure.
3) The final scan was performed one month after the baseline scan, in order to compare the results.

4. Results

Our experiment resulted in 124 zero-day malware instances, i.e. malware that was not found by F-PROT or Avast! BART immediately after the exposure period, but was detected after a new scan with updated signatures after the one-month dormant period.

4.1. Zero-day Malware Results

According to the offline anti-virus programs avast! BART and/or F-PROT, the 124 files are all zero-day malware, and none of the files were reported malicious at the baseline

[2] http://www.limewire.com
[3] http://utorrent.com

scan. In addition, all files detected by avast! BART and F-PROT have been uploaded and scanned at VirusTotal[4] where 36 different updated anti-virus engines are present.

As an example, the file *Easy Video Downloader 1.1 2008 fxg.rar* is detected by Avast! BART as *Win32:Trojan-gen {Other}*, while F-PROT did not detect it at all. However, F-PROT is not the only anti-virus engine that lacks a signature for this file. Only 5 out of 36 engines at VirusTotal detected this file to be malware. That means 31 engines, F-PROT included, were lacking a signature at the time of the last virus-scan.

From the VirusTotal results, we were able to check whether other anti-virus engines detected the same files as Avast! BART and F-PROT. The results from F-Secure and Symantec were further analyzed because they are big vendors of anti-virus solutions and they also had good descriptions of the different malware types, as apposed to F-PROT who did not offer any information about the malware types on their web site. It was also impossible to obtain information about when Avast! had incorporated specific signatures to their database.

A lot of the malware is also considered zero-day according to both F-Secure and Symantec. If F-Secure reports to have added the signature at some date in October, it means they did not detect the malware at the time of the baseline scan (which was done on October 1st). Our results show that out of the 124 zero-day instances, 60.5% were also zero-day malware with respect to F-Secure. Some of the files were not even detected at all, and we conclude that these are zero-day malware with reference to the specific anti-virus software. All files were gathered during September, which means they have been around for over one month. It is disquieting that a large number of anti-virus engines do not detect these files to be malicious, even though they have been in the wild for such a long time.

4.2. Zero-day Malware Sources

As we can see from Figure 3, which is based on the 124 zero-day malware infected files, most of the zero-day malware comes from the use of BitTorrent. While zero-day malware from the use of BitTorrent is estimated to 47% zero-day malware from the Gnutella network on the other hand constitute only 7%. The reason for this big difference can be explained by the difference in their search mechanisms, as discused further in Section 5.

The part labeled *Other* in Figure 3 constitutes 42% of the diagram; this includes files that were not present in the limewire, torrent or web download folder but rather on the desktop, in the Temp folder, the Temporary Internet Files folder, the Program files folder or different system files folders, to mention some locations. These are files whose source cannot be determined for certain. Some of them may come from surfing the web, by clicking on different pop-ups and ads, and some may be created when we installed some of the downloaded files. The part in the figure labeled *Web downloads* indicates that 4% of the zero-day malware comes from files downloaded during surfing the web.

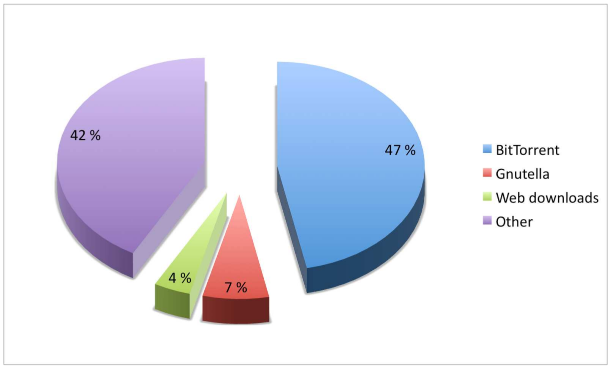

Figure 3. Zero-day malware sources based on the 124 zero-day malware infected files.

Source	**Downloaded files**	**Zero-day malware**	**%**
BitTorrent	∼ 400	58	14.5%
Gnutella	∼ 6000	9	0.15%
Web downloads	∼ 80	5	6.25%
Unknown source	?	52	?

Table 3. Approximate percentage of zero-day malware from different sources, based on all downloaded files.

Table 3 is an attempt to estimate what percentage of the downloaded files in BitTorrent, Gnutella and from the web, respectively, contained zero-day malware. It is important to notice the difference between the percentages in Table 3, which is based on all downloaded files, and the percentages in Figure 3, which is only based on the 124 zero-day malware infected files. Due to our procedure, where desktop anti-virus software removed malware specimens as they were detected, we do not have the exact number of downloadeded files, and the percentages should only be used as a rough estimate. Through BitTorrent the number of download files was actually about 400. We had 40 keywords which on average resulted in 10 downloads apiece. The Gnutella number is more difficult to estimate, since files from the different computers were gathered to a common pool, but the conclusion is unchanged. The percentage of zero-day malware was significantly higher in the BitTorrent sources than in the Gnutella network. Again, this can be partly explained by the difference in the search mechanisms.

It is worth mentioning that the files downloaded from the web are typically from the suspicious sites identified partly by the Google and Yahoo search engines. It will be very

[4] http://www.virustotal.com

wrong to think that 6.25% of files from the world wide web contain zero-day malware, but what the table indicates is that zero-day malware exist in all of these areas. Also note that a significant number of files with malware were found elsewhere on the computers; we can only assume that these files were downloaded by spyware contracted during the experiment.

5. Discussion

The term zero-day malware is widely used and the existence of such malware is well known. However, very few can refer to actual numbers that document the prevalence of zero-day malware. In our experiment, 124 unique files were identified to be infected with zero-day malware. The procedure focused on exposing the computers in the laboratory to a broad range of suspicious material and generally acting as an ignorant Internet user: Installing programs, visiting ads and clicking OK to everything that popped up was part of the exercise. Although a normal user would probably not manage to expose his or her computer to the same amount of suspicious material in the short timeframe used in this experiment, a normal user has a much longer exposure period (i.e. countinuous and never ending). This illustrates that the risk of getting infected by malware that is not detected by anti-virus protection is alarmingly high.

New malware that the anti-virus engines do not have a signature for is likely to escape detection by a desktop anti-virus solution. Proper behavior on the Internet can only protect users to a certain extent. If they visit the wrong web site or download a file with a zero-day malware, however, they will probably *not* be protected from infection.

In a threat summary for the second half of 2008 [11], F-Secure reports that one million detection signatures were added during the year - a number of hitherto unseen magnitude. The acceleration in introduction of new malware instances can likely be explained by the use of obfuscation techniques such as polymorphism and metamorphism. Such techniques have successfully been demonstrated to aid malware in evading detection by commercial virus scanners [12].

A deep analysis of the infected files is time-consuming and considered out of the scope for our experiment. Thus, we did not attempt to determine whether any of the 124 zero-day samples were just different obfuscated instances of the same origin. An indication is however given by looking at the malware descriptions on F-Secure's web site. Many of the infections seem to be just new types of malware or new instances of already known types, but they still fall under our definition of zero-day malware.

Concerning prevalence of zero-day malware in the different infection sources we based our experiment on, the sources using BitTorrent generally seem to contain less malware than Gnutella, although the amount of zero-day malware were in fact higher in the former. This may be related to the different search mechanisms in the two P2P technologies. Only BitTorrent provides the ability to search for newly uploaded files. As such, downloading newly added files is easier in BitTorrent, and the aspect of new malware is one of the most important characteristics of zero-day malware. If malware creators manage to distribute a new malware instance that the anti-virus vendors do not currently detect, the possibility of successfully infecting large number of hosts is a lot higher.

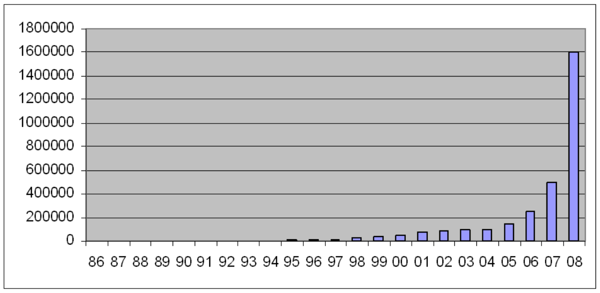

Figure 4. Accumulated number of malware signatures in F-Secure's database from 1986 to 2008 [11].

In this experiment we have demonstrated that our method is adequate to perform a retrospective measurement of the prevalence of zero-day malware. Although our approach was primarily based on using two offline anti-virus scanners to perform the baseline scan and the final scan, the use of VirusTotal illustrated the benefits of including even more anti-virus tools in the process to get more accurate results. The use of VirusTotal helped to verify that the findings were actual malware, and not false positives.

Our procedure for exposing the laboratory machines to potentially malicious content was focused on making the experiment as close to a real world scenario as possible, and that implies combining web surfing with file downloading and file-sharing activities. In our results the prevalence of zero-day malware for the different infection sources are indicated. However, because of our procedure we cannot state the origin of each detected zero-day malware instance with exact certainty. A stricter and more firmly defined procedure would have to be defined and followed if the goal were to get more accurate measurements.

6. Conclusion

We have presented an empirical study where we have exposed updated Microsoft Window XP PCs with different up-to-date anti-virus packages to numerous locations we suspected of containing zero-day malware. After a two-week exposure period, our computers had contracted a minimum of 124 malware specimens that were not detected by our anti-virus packages during (or at the end of) the period.

The prevalence of zero-day malware implies that anti-virus software which primarily relies on signatures does not provide sufficient protection. Coupled with the exponential growth of new malware variants, our findings indicate that the anti-virus vendors already have major problems with keeping the signature lag within acceptable limits.

7. Further Work

Due to time constraints, we have not gained any further knowledge on the prevalence of zero-day *exploits*. We suspect that this would require a more extensive lab setup, and a longer dormant phase. It is possible that more complete results could be obtained by automating the exposure process, e.g., by using web crawler technology.

It would also have been interesting to perform a more thorough investigation of the 124 zero-day malware instances in order to discover the exact nature of those files. We see great promise in tools like ANUBIS [13], and hope to delve further into these aspects in future research.

Acknowledgments

This paper is based on the results of a minor thesis at the Norwegian University of Science and Technology (NTNU).

References

[1] S. Shin, J. Jung, and H. Balakrishnan, “Malware prevalence in the kazaa file-sharing network,” in *IMC ’06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2006, pp. 333–338.

[2] A. Kalafut, A. Acharya, and M. Gupta, “A study of malware in peer-to-peer networks,” in *IMC ’06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2006, pp. 327–332.

[3] A. Berns and E. Jung, “Searching for Malware in BitTorrent,” University of Iowa, Tech. Rep. UICS-08-05, April 24 2008, http://www.cs.uiowa.edu/~ejjung/courses/169/project/publish/AndrewBerns_presentation.pdf.

[4] A. Clementi. (May 31, 2008) Anti-Virus Comparative No.18: Proactive/retrospective test. AV-Comparatives. [Online]. Available: http://www.av-comparatives.org/seiten/ergebnisse/report18.pdf

[5] SANS Top-20 2007 Security Risks. SANS Institute. [Online]. Available: http://www.sans.org/top20/

[6] D. Nunes and S. Keats, “*Mapping the Mal Web*,” *McAfee SiteAdvisor*, March 12, 2007, http://www.siteadvisor.com/studies/map_malweb_mar2007.html.

[7] S. Keats, “*Mapping the Mal Web Revisited*,” *McAfee SiteAdvisor*, June 4, 2008, http://www.siteadvisor.com/studies/map_malweb_jun2008.pdf.

[8] Most popular Windows downloads. Download.com. [Online]. Available: http://www.download.com/3101-2001_4-0.html?tag=mncol;sort

[9] E. Bangeman, “Study: Bittorrent sees big growth, limewire still #1 p2p app,” *ars technica*, April 21 2008. [Online]. Available: http://preview.tinyurl.com/3recfp

[10] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose, “All Your iFRAMEs Point to Us,” in *Proceedings of the 17th USENIX Security Symposium*, 2008. [Online]. Available: http://www.usenix.org/events/sec08/tech/full_papers/provos/provos.pdf

[11] F-Secure IT Security Threat Summary for the Second Half of 2008. F-Secure. [Online]. Available: http://www.f-secure.com/2008/2/index.html

[12] A. Moser, C. Kruegel, and E. Kirda, “Limits of static analysis for malware detection,” *Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual*, pp. 421–430, Dec. 2007.

[13] (2008) ANUBIS. Last visited February 19th 2009. [Online]. Available: http://anubis.iseclab.org