White Paper March 2005

AVERT™

Counting Spyware Detections

Perception and Reality

Table of Contents

The myth of signature counts

Shenanigans!

In the early days of the anti-virus industry, it was common for vendors to brag about their signature counts and withhold their samples from other anti-virus vendors to prevent them from catching up. Eventually, everyone realized that this was causing misleading claims and was performing a disservice to the marketplace. This practice in many ways led to the rise of the independent testing bodies, such as ICSA Labs, VTC, AV-Test.org and so on.

Unfortunately, the anti-spyware industry as a whole has not reached this level of maturity. Many vendors count each executable, data file, registry key, etc. belonging to a package as a new signature. There is also wide variation in the kinds of threats detected. Some detect backdoors Trojans and worms, while others do not. Others have numerous detections for cookies, which have no security impact and debatable privacy implications.

When is a threat counted as unique

The Anti-virus perspective

The anti-virus industry has long taken heat for confusion over virus naming, and there is wide disparity not only in how many detections a product claims to be capable of, but in how threats are named and counted. This discrepancy occurs in the anti-virus world for a variety of reasons:

- Historical conventions
- Lack of time to synchronize names during outbreak events
- Simple stubbornness
- Differences in generic and heuristic detection technology

To understand the last point, imagine the following 4 files consisting only of 6 letter sequences were all viruses:

1. ABCDEF
2. ACCDEF
3. AACDEF
4. AACDFF

Now imagine a group of anti-virus vendors writing code to detect all of these (detection code executed in order shown):

Vendor X:

- Detect A*CD*F as Virus 1

Vendor Y:

- Detect AACD*F as Virus 1
- Detect A*CDEF as Virus 2

Vendor Z:

- Detect ABCD** as Virus 2
- Detect ACCD** as Virus 3
- Detect AACD** as Virus 4

What you end up with is:

Table 1: Hypothetical anti-virus naming concordance

File	*Vendor X:*	*Vendor Y:*	*Vendor Z:*
	Sig Count 1	Sig Count 2	Sig Count 3
ABCDEF	Virus 1	Virus 2	Virus 2
ACCDEF	Virus 1	Virus 2	Virus 3
AACDEF	Virus 1	Virus 1	Virus 4
AACDFF	Virus 1	Virus 1	Virus 4

So in this grossly-simplified case, three vendors with a 200% variation in signature count all detect all 4 samples, but with different names.

The VGrep tool, maintained by McAfee for Virus Bulletin (*http://www.virusbtn.com/resources/vgrep/index.xml*) attempts to help alleviate some of the confusion by cross-referencing vendor's names across a large body of samples.

For a more realistic example, in a recent test at AV-Comparatives (*http://www.av-comparatives.org/seiten/ergebnisse_2004_08.php*), products tested claim anywhere from just shy of 53,000 signatures (Dialogue Science) to nearly 123,000

signatures (Frisk Software). But at the end of the day, almost all of the products tested detected over 300,000 of 323,000 unique samples. So a large difference in number of signatures may relate to only small differences in actual detection capabilities.

Now this level of confusion occurs in an industry that:

- Has been around for nearly 20 years in some form or another
- Has consistent definitions for what is and is not a virus or a Trojan (the definition for virus can be expressed in mathematical terms it is so precise, Adleman, 1988, "An Abstract Theory of Computer Viruses")
- Has a variety of professional bodies (AVPD, CARO, AVED, AVAR, EICAR) to promote cooperation and consistency
- Has routine collection-trading occurring between a vast majority of vendors
- Has a number of well-respected testing bodies (e.g. ICSA, VTC, VB, AV-Test.org, AV-Comparatives.org) with high-quality collections

In fact, the entire basis for rating the detection capabilities of anti-virus vendors rests on the fact that there are people and organizations with canonical, or at least, very complete, collections. In this way, we can make apples-to-apples comparisons between anti-virus products in a way that is unambiguous despite all of the inherent inconsistency in counting and naming.

The Spyware perspective

Now let's switch to the spyware industry, where a recent (September, 2004) internal McAfee survey showed the following "raw" signature counts:

Table 2: Spyware Product signature count

Competitor	*Product*	*# detections*
Aluria	Spyware Eliminator	18625
Lavasoft	Ad-Aware	9637
Computer Associates	PestPatrol	118060
Safer Networking Ltd	Spybot S&D	17679
Spycop	SpyCop	467
Webroot	Spyware Sweeper	31104
Javacool	SpywareBlaster	3183
PC Tools.com Ltd	Spyware Doctor	10684
Giant Company Software‡	GIANT Antispyware	> 100,000
McAfee	VirusScan Enterprise	3175*
McAfee	Antispyware**	384

* counts only Potentially Unwanted Program detections

** McAfee consumer antispyware product

‡ Now Microsoft Antispyware

This shows a difference of 3 orders of magnitude between the highest and the lowest signature counts in the different products. Now consider the performance of those products when they scanned McAfee's APP collection (detections that are neither viral nor Trojan):

Table 3: Spyware product detection over McAfee APPS collection

Vendor	*Number of Detections*	*Time Taken*
McAfee VirusScan Enterprise	11288	0:19:50
Spyware Doctor	135	0:00:24
SpySweeper	951	0:02:54
Adwaresafe	151	0:00:57
Adaware	356	0:03:19
PestPatrol	2601	0:25:00
McAfee Antispyware*	270	0:03:53
Aluria Spyware Eliminator	358	0:12:07
Giant Antispyware	617	0:22:19
SpyBot	0	0:03:08

* McAfee Consumer antispyware product

PLEASE NOTE:

- This is a terrible test and should not be used for real comparisons or competitive evaluation.
- McAfee Anti-Spyware Enterprise (corporate product) was not available for inclusion at the time of the tests.

The only thing the above results prove is that it is VERY easy to run a meaningless test. Why is it a terrible test?

- It includes samples collected by only one vendor (McAfee) and is HEAVILY biased in VirusScan Enterprise's favor, as this is the collection used to verify that we detect what we are supposed to in terms of Potentially Unwanted Programs (PUPs) before every DAT release.
- It is heavily biased AGAINST products that rely on more than a file to trigger detection. For example, SpyBot detects PUPs ONLY when the properly installed package (including registry entries, et. al) is present on a system. A dumb file collection will not trigger anything. SpyBot actually is a decent anti-spyware product against live PUPs, which cannot be determined from this test.
- It does not include detection for cookies and registry entries, which may make up large portions of the signature databases for some products.
- It does not include backdoor Trojans and other "traditional" malware, which often find their way into anti-spyware products. For example, nearly 70% of the files listed in Pest Patrol's Pest Encyclopedia (*http://research.pestpatrol.com/search/browse.aspx*) of 25,000+ pests are in categories that AVERT generally treats as Trojan horses. In other words, nearly 3/4 of their signature base may include items already detected by McAfee VirusScan Enterprise.
- It does not include a collection of over 200,000 unique dialer programs that are detected by less than 100 signatures in the McAfee DAT files.

But wait, there's more!

But these aren't the only challenges in trying to make some sense of the anti-spyware numbers game. Most viruses and Trojans are self-contained pieces of code – they usually consist of a single file, or even just a tiny piece of code inside another file. While polymorphism and parasitism make the virus picture a little more complex, and can prompt heated disagreements between anti-virus vendors over the correct family name, they are in many ways, less complex than PUPs.

Many PUPs are complete software packages. They include installers, uninstallers, read-me's, EULA's, data files, support DLLs, shortcuts and other paraphernalia common to Windows applications. In our early experience with the spyware collected by the McAfee Anti-spyware (MAS) consumer team, we've seen the following:

- There are something like 14,000 files (only about 5,000 are left after removing data files like txt, jpg, registry, etc. files) contained in the MAS collection, belonging to the only 400 or so unique detections that the product contains. So whereas the DATs have 3,000 signatures to detect 11,000 files (about 3 files per signature), MAS only needs 400 to cover about 14,000 (about 35 per signature).
- A single MAS detection will often contain files detected under 5-10 completely different names in the DATs.
- There is much more code reuse in PUPs, such that the same exact binary might exist in 15 or more individual PUP packages. Even worse, the same binaries might exist in packages that do NOT have PUP characteristics at all, and which we do not want to detect in this context.

In a nutshell, there is **absolutely no correspondence between the number of signatures in a database, and the effectiveness of that product against any given set of threats.**

There is very little consistency about how detections are counted amongst anti-spyware vendors, and no one is in a position to call them on it, because no one has a comprehensive collection. AVERT estimates that there are approximately 7,000 -10,000 real unique PUPs to detect, counting them more or less the way we do in the anti-virus space. So let's assume that anti-virus-centric vendors only see half of the complete picture, and double it, and add some variability for naming conventions, and differing levels of generic detection amongst vendors. Anything more than about 20,000 should raise alarm bells as possibly being inflated.

Clouding the picture

There are potentially a number of ways to report detections in a host-based anti-spyware product:

- By number of unique detection names
- By detection name and variants
- By number of files and registry entries detected
- By number of files and registry entries removed

On some level, none of these methods is more intrinsically correct than any other. However, comparing a report from a vendor that uses the first method with one using the last will generate apparently lopsided results even if they detect and remove the exact same objects!

Some products detect registry keys that exist by default on Windows systems, which may inadvertently get counted as "misses" by other products.

Some products will report the same object multiple times. In one test, a single DLL has been seen listed 50 times in a single report. Many anti-spyware products will report registry keys several times, once for each hive via which they may be addressed, for example:

- HKEY_CLASSES_ROOT\ProgID
- HKEY_LOCAL_MACHINE\Software\Classes\ProgID

Some products will report each subkey or registry value present when deleting a parent key, causing a single "known bad" object to yield a dozen or more items in a repair log.

We have seen cases where different anti-spyware products reported anywhere from 5 to 96 "items" while detecting and repairing a single adware package with virtually identical results, i.e. they all removed the same files and registry entries.

In other words, there is no correspondence between the number of objects reported by two products and their effectiveness.

So how should products be compared?

The purpose of a detection test is to determine which of a group of products can locate the most "bad stuff" effectively and efficiently. There are several pre-requisites to making a valid comparison:

- All the products tested agree on what "bad stuff" is. At the very least, only samples that all products agree on categorically should be included. Unless all products are intended to remove Trojan horses, or peer-to-peer file sharing programs, they should not be in the test set.
- The sample set should be as large as possible. The samples in the set should come from a well-defined period of time, and they should be verified by an expert in the field. Ideally, the reviewer's sample set should be a superset culled from industry sources, independent experts, and the reviewer's own research to avoid unfairly biasing the test.
- Where the sample set must be limited, the samples should be chosen according to some meaningful criteria – for example, prevalence, potential risk or payload, difficulty of removal. A small set of poorly-chosen samples will mask both positive and negative aspects of the products under test through sheer chance.
- Both false positives AND false negatives should be tested. It is very easy to write detection routines that catch every sample, but create numerous false positives or performance problems.
- Success or failure criteria should be based on independent measurement of file, process or registry changes, NOT on the performance of some "reference" product.

Because of the lack of consistency and definitions in the anti-spyware market as a whole, most tests to date have been poorly designed and implemented. Sample sets have been small and arbitrarily-chosen. The samples used and their effects on the system have been poorly documented. Measurement has often consisted of listing

how many items the product reported, whether or not they were even related to the samples in question.

Next Steps

Industry Cooperation

Currently the anti-spyware industry is engaged in a large game of poker, where no player can see the others' cards. Everyone is bluffing with their customers in the hopes that no one calls them on it, but this situation is untenable. McAfee will begin trying to forge the same kind of alliances among the reputable members in the anti-spyware community that we currently have in the anti-virus community. There is already some limited collection-swapping of PUPs occurring between several large anti-virus vendors.

We will look to expand this effort so that we can begin to get a clearer picture of the competitive landscape, and truly begin to measure our progress on a scale that has some bearing in reality. Naturally, there is some risk in this approach insofar as other companies will also have access to our collections. Through several years of virus collection-swapping, there have been no major shake-ups resulting from this practice; companies that were on-the-ball before still are, and the less-well-organized companies are still behind. We expect the same to be true in the anti-spyware market. In any case, we will be in a better position to absorb new content given the tools and techniques we've developed in the anti-virus space, than anyone else.

Improved Independent Testing

Finally, the anti-spyware industry needs to encourage improved measurement and testing of spyware. In an environment where flawed tests determine vendor superiority, there is no rational way to determine how to improve or to measure our improvement. We need to determine which test methodologies are most likely to yield accurate and meaningful results, and work with independent test organizations to implement those techniques. We will need to help independent reviewers get over their fear of legal repercussions, and build useful collections.

Until this happens, however, we are still in the Wild West of testing. Customers, ill-informed reviewers, partners and OEMs are going to be running poorly-conceived ad hoc tests that do more to cloud the situation than enlighten it. Reviewers should:

- Wherever possible, use prevalence-based data to guide their test sets. Pulling PUP prevalence data from customer reports, support logs, vendor prevalence reports, literally anything has more validity than testing against whatever N PUPs happen to be sitting on someone's Mom's computer, or happen to get installed while visiting a few dodgy web sites.
- Many inexperienced reviewers are going to run a test where they install a bunch of arbitrary PUPs, then run an anti-spyware product, then run a second product, and bludgeon the first product's vendor with whatever they missed. The results of the Spyware Warrior review above indicate that vendors tend to detect ALL or NONE of a particular package, but that all of them miss a significant number of packages. Unless a knowledgeable security researcher has confirmed whether all misses are relevant (and not false positives) and the test was also run in the reverse order, the data produced as a result is meaningless.
- When a product false negative, false positive or mis-repair is reported, the relevant samples should be available to the vendor to reproduce or refute the claim. Since there is no standing legal or industry-wide definition of what is "spyware", any deviations between products may be intentional. At the very least, the vendor should have the opportunity to fix any reported problems for future versions.

Conclusion

The anti-spyware market looks very similar to the way the anti-virus market looked ten years ago. There is great opportunity and also great risk. The market is beginning to show signs of maturity. Legislative and enforcement activity in the US and abroad could completely redefine the playing field in mid-stride. And the behavior of the organizations creating PUPs could make the security community's job significantly easier or harder. But this environment of frequent change and high risk/reward is one we are very used to.

About AVERT

McAfee AVERT is one of the top-ranked anti-virus and vulnerability research organizations in the world, employing researchers in thirteen countries on five continents. McAfee AVERT combines world-class malicious code and anti-virus research with intrusion prevention and vulnerability research expertise from the McAfee IntruShield(R), McAfee Entercept(R) and McAfee Foundstone(R) Professional Services organizations. McAfee AVERT protects customers by providing cures that are developed through the combined efforts of McAfee AVERT researchers and McAfee AVERT AutoImmune technology, which applies advanced heuristics, generic detection, and ActiveDAT technology to generate cures for previously undiscovered viruses.

McAfee, Inc. 3965 Freedom Circle, Santa Clara, CA 95054, www.mcafee.com

6-sps-avecounting-001-0305