REAL WORLD ANTI-VIRUS PRODUCT REVIEWS AND EVALUATIONS – THE CURRENT STATE OF AFFAIRS

Authors' note: The original work upon which this paper is based discussed problems and alternatives relating to the evaluation of anti-virus software. It was published with the hope that users and developers would provide us with suggestions for developing evaluation methodologies which would work in the real world. Our goal was to help create viable evaluation criteria which corporate security managers could apply when selecting an anti-virus product. Since the original publication of this paper, we have received suggestions from many anti-virus product vendors, security personnel, magazine evaluators and reviewers and government representatives. This revision reflects the new direction anti-virus product certification appears to be taking in the "real world" today.

Sarah Gordon (sgordon@dockmaster.ncsc.mil)

Richard Ford (rford@commandcom.com)

Abstract:

This paper will discuss frequently encountered errors in the evaluation process relative to anti-virus software selection by examining some of the methods commonly used by corporate and governmental personnel working in the area of Management Information Systems (MIS). In addition to discussing inherent problems, we will suggest alternative methodologies for evaluation. We will examine commercial certification processes, as well as the Information Technology Security Evaluation and Certification (ITSEC) approach, as possible models for anti-virus product evaluation and certification. Finally, we will discuss ways in which the information which is currently available may be used to help select anti-virus software which is both functional and cost efficient.

Introduction

The evaluation of anti-virus software is not adequately covered by any existing criteria based on formal methods. The process, therefore, has been carried out by various personnel using a variety of tools and methods. Some of these tools and methods should be part of the evaluation process; others can provide misleading or damaging information resulting in increased exposure to computer viruses. Areas of the evaluation which are relatively straightforward include the elimination of products which are unsuitable for your environment, the cost of the software, comparison of vendor pricing policies and licensing conditions and assessing compatibility requirements. In all of these areas, you must of course anticipate future growth; for instance, if you are planning to add platforms or anticipate many users taking work home, you will need to rule out software which does not support multiple platforms or which does not allow for acceptable home use pricing packages. Products must of course be well documented and easily configurable. Transparent operation is required, as products requiring large overhead tend to invoke removal or circumvention on the part of the user or administrator. These areas of examination are important; however, there are other aspects of the selection process which are even more critical. You may even depend on evaluations you don't know anything about, as in the first two cases we will examine. Unfortunately, as we will show, there are serious problems with *all* of the evaluations on which people are currently relying.

"It is unfortunate, but a large majority (say 90 percent) of the current anti-virus tests published within the last couple of years are worthless, or even worse than that, purposefully made misleading." [1].

We will examine this claim, beginning with the types of evaluations you may find yourself having to base your decision upon. The following, based on "Real-World Anti-Virus Product Reviews and Evaluation" [2], illustrates that the majority of methods are impractical.

The Provider of Friendly Advice

Managers seriously underestimate the power of "the friendly recommendation" by friends, or colleagues who have "used xyz anti-virus and it worked just great". However, with the limited time and resources many companies have to investigate what constitutes a viable anti-virus solution, the influence of the friend should be duly noted. The inherent problems in relying on the recommendation of friends, even knowledgeable friends, result from both the competence level of the friend and the variance in needs of users. For instance, if the main requirement for "the friend" is that the system provide for a means of circumventing a scan, whereas your need requires non-circumvention, you would be ill-advised to select a package which allowed for easy circumvention. Variables such as packaging, pricing, and speed are all subject to interpretation, and the interpretation will be greatly influenced by the needs of the individual who does the reading.

A much more serious issue is related to claims of performance in the area of actual virus detection. Consider the claim of a friend that "the program worked fine. My system is virus free!". The question here is "How does he know he never had a virus?" If he is using a product which misses viruses, he may think he never has had one when in fact

he has. He may also be relying on what he has heard from a friend of a friend, who really likes anti-virus because it is the one he is familiar with. People are very much influenced by name-recognition. However, do you want to trust the security of your data to a product based on its name? We argue that you should base your decision on the actual performance of the product. Unless 'the friend' happens to be particularly skilled in anti-virus product evaluation metrics and methodologies, it is probably not a good idea to trust his or her advice.

The Employee (or the employee's friend, colleague or Internet acquaintance)

The Employee resembles "The Provider of Friendly Information" in many ways, with the additional attribute of feeling somewhat responsible. Employees become "virus experts" by reading virus information message areas on various on-line services. They may obtain some viruses to "test" the efficacy of software you have, or are considering purchasing.

You need to be concerned about the employee *not* because he/she is acting out of any form of malice; on the contrary many employees feel they are helping you by becoming "experts" in virus testing. However, a thorough understanding of product evaluation is not something an employee can learn in their off hours by "beta-testing" some anti-virus software and recommending it to people because "it caught a lot of viruses".

The reasons such well-meaning expertise is ineffective relate in part to the technical skills required to construct and perform a meaningful test. Can the employee disassemble and replicate samples to ensure the test-set is clean; i.e. that test samples are actually viruses and not corrupted files? Is the employee capable of judging the efficacy of the removal and terminate and stay resident (TSR) modules of packages? What tools does he have at his disposal? Does the employee have a dedicated test machine upon which to perform tests and has he or she studied the subject enough to do the job correctly for you? It is unlikely that most companies have the resources to answer 'yes' to these questions, yet we see company virus representatives talking about their in-house evaluation of products. We suggest that their evaluation is not only inadequate, but it can also be harmful to the integrity of the company data.

The employee who has been granted some official status may be familiar to you as one whom you have designated to do in-house evaluation - a member of the technical support team or a programmer. However, even a technically competent employee is not likely to be able to carry out tests of the quality which you require in order to evaluate a product fully. You must remember that "technically competent" in programming or network administration does not imply "technically competent" with computer viruses.

The Computer Magazine (non-virus/security specific)

The Computer Magazine evaluator /reviewer is in a unique position; he holds a lot of influence over the public, while at the same time usually having insufficient experience in the field to provide accurate information. This frequently leads to reviews which rely on incorrect assumptions. As an example, a well-known computer magazine recently hosted an on-line forum during which the magazine "expert" stated certain boot sector viruses can infect the fixed disk of an otherwise clean machine simply by the user typing the command "DIR" with an infected diskette in the A: drive. Apart from a lack of technical ability and information, a computer magazine is unlikely to have a large and clean collection of computer viruses. Therefore, the reviewer is likely to take one of the following approaches:

- Carry out a test on a very small "collection" of "viruses", gathered from friends or colleagues.
- Approach an anti-virus software developer for a collection of "viruses".
- Obtain a collection of "viruses" from a virus exchange bulletin board system (vX BBS), ftp site, the World Wide Web, or a publicly-available virus collection such as those available on CD-ROM.
- Use a virus "simulator" to test the detection capabilities of products.

Unfortunately, any tests based on "samples" obtained in this manner lead to questionable results. We shall examine the problems with each approach in turn.

Using a small collection of viruses is clearly an unacceptable way to carry out a product evaluation. In order to test a product's detection capability, tests should be carried out against at the very least all those viruses known to be in the wild (ITW). We suggest "The Wildlist", by Joe Wells as a good starting criteria for detection. Testing against only a few viruses will not give an accurate impression of a product's ability to meet the real threat. However, such tests have been done and the results printed. We are even aware of one review which based its final detection results on a test-set of only 11 viruses [3].

The problem with using a vendor's virus collection is equally obvious: bias. A vendor could simply doctor the test-set so that its own product would score well, or release test-sets which will show the product gradually improving with time.

There is, of course, the additional concern of magazine reporters' and journalists' technical competency in not only replication and analysis but in management of virus libraries. It is important to make sure the viruses used for testing are not only real, but that they do not inadvertently escape and cause harm to unsuspecting users, or result in liability to the magazine. We know of several cases where computer viruses were inadvertently released on computer diskettes distributed with computer magazines (although we are *not* aware of any link between this and the testing and reviewing of anti-virus products).

The issues raised by obtaining a virus collection from a vX BBS or the Internet are more subtle. In these cases, the reviewer has no way of ensuring that each sample is actually infected by a virus. Virus collections obtained in this way are frequently badly organized, containing a large number of corrupted or uninfected files. Detection tests carried out against such a collection are not likely to be accurate, and will discriminate against the better products. This is summed up by Tanner [4] in "A Reader's Guide to Reviews", which looks at some of the ways to fix a test made on two fictitious products, Grottyscan and Wonderscan:

> You'll need a test suite. Ideally, you should get it from Grotty Inc. You might find that Grotty Inc. don't have a virus library, in which case, you should find a collection of files that contain viruses and also lots of corrupted and innocent files. That way, if half the files you use are not viruses, the GrottyScan score of 30% doesn't look too bad compared with the 40% that the best product got.

The article continues onwards in a similar vein, and highlights several of the other ways to bias a test, either intentionally or inadvertently.

In the case of a fixed collection, like that available on CD-ROM, there is yet another issue: anti-virus product developers have had unrestricted access to the actual samples against which the test will be carried out. This is a problem because if the scanner manufacturer has access to the test collection, it is a trivial exercise to alter the product so that every sample in the test-set is deemed to be infected, regardless of its state. Although the scanner may detect the samples of the virus stored on the CD-ROM, it may be unable to detect further replications of each sample. This is particularly true in the case of polymorphic viruses, where test results are invalidated if the software developer has copies of the actual samples used during the detection tests. Thus, using a fixed collection of viruses to which every vendor has had access provides little real information about real world scanner performance.

We have observed the development of a disturbing trend: testers using virus simulators to test products. This is unacceptable for several reasons. One of the more popular simulators creates .COM and .EXE files, and provides supplemental Mutation Engine (MtE) samples. The .COM and .EXE files simply print a message to the screen and exit. It is clearly unacceptable for an anti-virus product to detect such activity as viral. Although these files also contain virus signatures (non-functional "fragments" of virus code), anti-virus technology has by necessity evolved in such a manner as to render detection of such simulated "viruses" a useless measure of the product's actual capability. According to a report published by Luca Sambucci, of the Italian Computer Virus Research Institute, tests using simulated viruses are "misleading and in some cases harmful".

In comparative tests we conducted using both simulated viruses and real viruses, we found that while the scanners we tested detected all of the real viruses, only one scanner detected any of the simulated viruses. Tests performed on simulated (fake) viruses do not necessarily accurately reflect the detection capabilities of a product [5]. [Note: The EICAR test file, developed by the European Institute for Computer Anti-Virus Research, should not be considered a simulated virus; rather, it is a program which scanner developers have deliberately chosen to detect. While it is not useful for measuring the comparative detection ratio of products, it may be used to test installation of anti-virus products. It is available from most vendors as well as from http://www.commandcom.com/html/eicar.html.]

The use of simulated polymorphic viruses presents yet another problem. In the most widespread virus simulator available, the Dosen Rorenthal Virus Simulator (this and other simulators are discussed more completely in [5]), the polymorphic viruses supplied *are* viruses, but have extremely limited propagation, infecting only certain designated goat files. Since these "viruses" cannot infect any other executables, the ability of a product to detect them is meaningless in terms of actual protection for the user; a vendor may of course decide to detect them for purely commercial or academic reasons. One possible risk is that these "test viruses" can be modified to be malicious in their action. Thus, many products detect these files "just in case". Such test viruses provide fodder for test libraries, but little else. The creation of computer viruses for any "testing" purpose is both unnecessary and unethical, and the International Federation for Information Processing (IFIP) has issued strong positional statements against such creation.

Assuming that the magazine has managed to gather a number of real viruses without obtaining them from a vendor, a CD-ROM, simulator or unverified source, magazine evaluations rarely test anything other than user interface, configuration issues, and the detection rate of the non-resident scanner. While these factors are important, in no way do they comprise a comprehensive evaluation. Yet, many MIS managers base their choice of anti-virus software on

"Editor's Choice" Awards, or magazine reviews. Such awards are a valuable measure of some aspects of performance, but can be subjective and should not be considered in any way a complete product evaluation.

The Computer Security/Virus Magazine

Reviews published by computer security/virus specialist magazines can provide you with information which may be useful in determining a product's strengths and weaknesses because they have a distinct advantage: the reviewers generally have both experience and a specialized knowledge of anti-virus products. These reviews tend to be well done and informative, focusing on the ability of products to meet published criteria.

Many reviews published in this type of journal attempt to focus on the threat posed in the real world, concentrating on those viruses which are known to be ITW. *Virus Bulletin*, for example, uses the Wildlist to form the "In The Wild" test-set for file viruses. This examination of the real threat, frequently coupled with tests which take into account the product's performance against a number of different infection strategies leads to in-depth reviews of a good quality. Unlike most magazine reviews, the specialist magazines are almost guaranteed to carry out tests against real viruses, and are a source of accurate detection results. Unfortunately, even these reviews have their share of problems. For instance, although having now instituted a totally ITW Polymorphic test suite, *Virus Bulletin* tests on boot sector viruses and polymorphic viruses have in the past included viruses which are not in the wild, leading to some confusion in interpretation of test results. *Secure Computing* published in their May 1996 Lead Review, tests which measured the ability of a program to detect its "Advanced Polymorphic" test suite. The scanners were tested on a collection of polymorphic viruses which were damaged in some way and would not either replicate or execute. Samples which do not replicate are of course not viruses, and while the tests were correctly interpreted, they are also a completely meaningless measure of actual protection.

Another commonly cited problem is that of tester independence. The two most well-known magazines which regularly test anti-virus software (*Virus Bulletin* and *Secure Computing*) have both been associated with producers of anti-virus products: *Virus Bulletin* with Sophos (Sweep) and *Secure Computing* with S&S International (Dr. Solomon's Anti-Virus Toolkit). While there is little evidence of deliberate bias in the review methodology and choice of test-set, these links are worth considering, and are frequently cited by disgruntled product manufacturers. How much bias there is in reviews carried out by such journals is impossible to quantify, but we stress that assuming bias when there is none is just as damaging as not being aware of bias when it is present.

Another problem is the limited nature of the tests. Non-resident scanners are the most commonly tested modules of anti-virus software. The "best" product for a company must be able to operate in a variety of environments, and under several different conditions. Most reviews (particularly comparative reviews) are in reality only measuring one aspect of product performance. Properties which are trivial to measure, such as the rate of false-positives, are often overlooked, and disinfection or detection in memory is rarely if ever tested. Due to time constraints and cost, however, it is not practical for even a specialist magazine to test all aspects of product performance. *Virus Bulletin* has taken some positive steps in this area, however, and is in the process of adding memory-detection and disinfection testing.

Finally, the information given in these magazines is often highly technical in its nature, and it is easy for the reader to suffer from an information glut, obscuring the true strengths and weaknesses of the product. An example of this is the *Virus Bulletin* comparative review of virus disinfection software [6], where the results detailed which parts of the EXE file header had been altered - data which most users would not know how to interpret.

Even with these problems, the virus and security specific publications offer possibly the best analysis of the detection capabilities of anti-virus products.

The Independent Professional Evaluator (IPE)

There are some independent reviewers who posses the expertise to conduct a meaningful review. One good example of such a reviewer is Rob Slade, a frequent contributor to Virus-L and the Fidonet Virus echo and author of several books on computer viruses. His reviews illustrate a major difficulty experienced by others who are attempting to carry out reviews: lack of resources. However, in Slade's case much of this is made up for by his experience and expertise. While Slade represents all that is best about the IPE, there are many self-appointed experts who have neither his experience nor expertise. There is no easy way to discriminate between those who are qualified to carry out such a review and those who are not. One only has to recall the glut of virus "gurus" who appeared during the "Great Michelangelo Scare" to see the problems which you will have deciding how much reliance to place in independent reviews of software.

Another notable reviewer (and founder of the Italian Computer Anti-Virus Research Institute), Luca Sambucci, has provided independent testing to computer magazines since 1992. His anti-virus tests are thorough and competent; however, he has not released a result for almost one year. He still conducts tests, and is primarily concerned with

scanner-based detection. He includes explanations of test terms in his test documentation, and gives developers the opportunity to comment on the tests -- as part of the actual test document. Although Sambucci's tests are good, it is difficult to pick his results out from those of the other self-appointed experts without considerable expert knowledge.

The signal-to-noise ratio surrounding the IPE can be observed by monitoring the electronic traffic which accompanies reviews by other independents. Generally the complaints revolve around the lack of performance by a specific product and the qualifications (or lack of them) of the IPE. The publication of qualifications of testers is an important aspect of a complete evaluation and is critical in the area of product certification. The need for this is built in to the very fabric of the Trusted Computer Security Evaluation Criteria (TCSEC): 'Certification should be done by personnel who are technically competent to assess the system's ability to meet the security requirements according to an acceptable methodology' [8]. Thus, without an in-depth knowledge of the IPE's qualifications and history, you should assign little (if any) weight to his results.

The Commercial Evaluator

Probably the most well-known commercial evaluators in the USA are Patricia Hoffman (*VSUM*) and the National Computer Security Association (*NCSA*). Currently there are serious problems with both of these evaluation services, although since the earlier study we have observed some of these problems have been addressed. In particular, *NCSA* has made significant revisions to its test methodology and criteria. The following list of problems, therefore, will be followed by a notation of the changes adopted by *NCSA*.

In both cases, "certification" is not in fact a thorough testing of the entire product, but a test of the scanning engine, carried out by running the product on a large collection of files which the evaluator claims are infected. In other words, the only property of the product to be evaluated is the non-resident virus scanner's ability to detect viruses. No tests are made on other critical areas of the product, particularly, the real-time protection offered or virus disinfection.

An epidemiological overview of viruses shows that although there are over 8000 viruses known for the IBM PC or compatible, there are less than 300 ITW (that is, actively spreading on PCs). A list of such viruses is maintained by Joe Wells. By collating statistics provided by over 30 contributors from many different countries, Wells tracks those viruses which are spreading. Participants in the list include all the major anti-virus software developers, and several independent researchers. The list is broken down into two parts: an upper list, for viruses which have been seen by two or more participants, and a lower list, which is made up of those viruses seen by only one participant.

Analysis of Wells' list shows that the real threat to computers is posed by less than 300 different viruses; if a computer were protected with a scanner which detected just these viruses, well over 99% of the total threat would be covered [9]. Thus, any intelligent test of anti-virus software must weight the detection of these wild viruses *significantly* higher than detection of other non-wild (Zoo) viruses. In essence, tests of Zoo viruses such as those performed by *VSUM* and *NCSA* provide almost no information on the suitability of a virus scanner for a real-world application.

Such tests, within certain limits, do give the reader quantitative information. However, they are highly limited in their applicability to anything approaching formal certification. Certifications like this fail to provide a fully functional baseline for several reasons; foremost among them the only information given is the overall detection rate of the scanner. No information is given about how well the product performs against the threat which users face in the typical office environment. In an extreme case, it would be possible for a product which could not detect any virus which is in the wild to still be certified. [One test which it is valuable to apply to any evaluation of anti-virus software is to examine how a simple batch file which identified every file it was presented with as infected would fare using the test methodology. Under any test which just measures overall infected file identification, such a batch file would get the highest possible score - a result which is obviously misleading.]

The tests by these commercial evaluation/certification services also do not take into account products which have "review" modes, although this problem is in the process of being reviewed by the Anti-Virus Product Developers (AVPD) Technical Committee, a vendor organization composed of technical representatives of member companies. The problem of review modes is a thorny one to solve. Consider a product which changes the way in which it operates when it detects more than a certain number of viruses on any one scan, loosening the criteria which it uses to identify files as infected. Such a scanner would do well on a test carried out against a large number of infected files. However, its detection rate in the test would not reflect its detection rate against the real threat, as usually one would be relying on the scanner to scan incoming diskettes, when the product would apply its stricter criteria for detection.

Finally, there is the question of who has access to the test-set. If software developers are allowed unrestricted access to the actual samples used for the certification, an unscrupulous vendor could change its scanner so that it identified every file in the test-set simply by carrying out a search for a hexadecimal scan string. As the vendor's only interest is finding files in the test-set, the search pattern would not even necessarily be taken from the virus: it would just need

to be something capable of identifying that particular file. In the case of polymorphic viruses, this would result in the scanner detecting the samples in the test-set, but no other replications of the same virus. However, denying the developer *any* access to the test-set raises questions about the quality of the test-set: are the files in it actually infected? How much can the test results be relied upon if there is no peer review of the test samples? [3, 7]

In 1995 the *NCSA* certification scheme [then under the direction of one of the authors, RF.] was altered to reflect new, more stringent criteria. A 100 percent detection rate of ITW viruses using the Wildlist as the criteria for such viruses was implemented, with a two month lag time in testing to allow vendors sufficient time to implement detection, taking into account Beta test and shipping cycles. Developers were disallowed access to any samples used in actual testing in the Wildlist portion of the tests. Developers who were members of the AVPD were given access to *replicants* of samples should their product fail to detect them during a certification test. This has the dual benefits of ensuring that the samples are actually fully functional viruses while disallowing the possibility of the developer implementing detection for the file rather than the virus. As a commercial certification, the PC version of the *NCSA* scheme found acceptance as a minimal criterion by which users could judge effective detection rates of scanner portions of anti-virus software. According to *NCSA* Spokesperson Pam Martin, "Any certifications performed by *NCSA* are performed strictly for end users. There is no attempt to mimic or supplant the ITSEC or TCSEC. Both of these look at multiple functions to determine a security level. Anti-Virus applications are only one of several parts of a total system, which would be evaluated under these more formal programs."

The *NCSA* scheme has not been without problems. A certification scheme for the Apple Macintosh platform which was prematurely promoted had no documented test methodology or criteria; we are told it has been discontinued. *NCSA* "Approval" was briefly promoted as a less stringent form of testing, requiring products to pass certain limited tests. This has also been discontinued and the information regarding the "Approval" has been removed from their WWW Site. *NCSA* has provided statements relative to meeting certain limited test criteria for at least one company; the claims have been publicly disputed by industry experts, and we have found the claims to be technically invalid.

However, the PC portion of the scheme developed during 1995 remains viable. Some anti-virus experts have voiced concern over the direction of the scheme, as it is no longer under the direct supervision of an anti-virus specialist. However, Joe Wells, developer of the Wildlist, has agreed to act as an off-site overseer to the testing methodology and maintainer of the virus library. Wells is a recognized industry leader in the field of anti-virus research. The future direction of the scheme remains to be seen; however, according to Martin, "*NCSA* is working with Joe Wells, and the AVPD, to determine any modifications in direction for the current testing scheme. *NCSA* has received requests to perform more formal false alarm testing, to test "TSR" type background protection, and to test repair capabilities of products. Any future changes will be discussed with AVPD before implementation, and would be implemented with a several month lead time." It is the opinion of these authors that anti-virus tests should be performed by specialists with considerable experience in testing. While Wells' qualifications are excellent, the fact remains he is not on-site. This could present problems in test administration and interpretation.

Secure Computing Checkmark, from West Coast Publishing, claims to be a quick, up-to-date, and inexpensive scheme which product developers may use to show independent verification of detection abilities of products. It is hoped that the scheme will provide developers with a way to support detection claims by referring to their independent third-party tests, and provide users with a way to know products meet a minimally acceptable criteria for virus detection. The author of the scheme, Paul Robinson, editor of *Secure Computing*, states that the purpose is to add value back into the industry and to provide benchmarks in the context of evaluating claims. "As reviewers and testers we need to be very transparent. This extends to methodologies; we are telling people exactly how we are testing what we are testing, there is no room for impurity in the test." The scheme is still under development, and appears from the information available to promote the testing of products using documented methodology and criteria. Currently, plans include using the Wildlist as a source for selection of ITW samples; however, identification of included viruses does remain at the discretion of the *Checkmark* administrator. The testing list is to be made available three months prior to the test. Testing is planned quarterly, and will be made of the scanner portion of products only. Vendors will pay an evaluation fee. The fee varies depending on the number of platforms evaluated. The scheme appears to be developing along the same lines as the new *NCSA* scheme in that no vendor will be given exact samples of missed viruses, but rather replicants.

One of the benefits of this scheme is that the methodology is clearly documented and has been distributed to interested parties. However, as the scheme is still in its draft phase, it remains to be seen how widespread acceptance of the standard will be. The documents relating to the scheme furnished to the authors show promise, but only time will tell which direction the final scheme will take.

The Academic Evaluator

Another useful source of information is the Academic Evaluator. Good examples of the type of tests carried out by such evaluators are those by Vesselin Bontchev, formerly of the Virus Test Center (VTC) at the University of Hamburg. The principal advantage with these tests is that the test metrics and methodology are clearly stated. The results are generally presented in a scientific manner and the reader is left with little doubt about how they were obtained [10, 11]. While the tests are another useful and accurate source of information they are limited in scope. Tests seem to be mainly concerned with overall detection rates. Little or no mention is made of detection of those viruses which are known to be ITW, although this information is usually available to those who are prepared to extract it from the raw test data. One potential flaw is that these tests may be carried out by students, who have limited resources and who are performing work in an academic (learning) environment.

The New ITSEC Approach.

The ITSEC was issued within the European Community in the summer of 1991, as an attempt to provide formal internationally-recognized standards for the evaluation of IT products for use within governments. In the UK, the market for evaluated products has been driven by Government procurement policies, especially in the defense industry. The ITSEC concerns relative to anti-virus product evaluation differ from the United States TCSEC. Whereas TCSEC specifies development assurance criteria, ITSEC requires certification and accreditation activities which assess how the product matches the operational environment; i.e., how the product meets the real world threat posed by computer viruses. While there is yet no formal methodology available on paper, the UK ITSEC Anti-Virus Working Group (AVWG) was kind enough to send us information on the status of the project.

Each ITSEC certification requires that products of a particular Functionality Class meet a certain Security Target, which consists of either a Systems Security Policy containing a statement of the security objectives, threats and necessary countermeasures for the system, or a Product Rationale, which contains a list of a product's security features, the intended method of use and the intended environment with its associated threats. The traditional ITSEC approach may be thought of as a "snapshot" of the developer and the product at any one time. Thus, only the version of the product which is evaluated by the Commercial Licensed Evaluation Facility (CLEF) is certified; certification lapses with the very next version of the software released. Anti-virus software evaluation requires a more dynamic approach.

Furthermore, the traditional ITSEC approach includes an examination of the development environment. Current work seems to indicate that in the case of an anti-virus software package it should be possible to extend this examination to include such issues as how well the company is able to maintain its product. It is not sufficient for a company to demonstrate its ability to detect a certain percentage of all known viruses in any one version of its software: it must be able to show that it has appropriate procedures in place to track the threat, and alter the product accordingly to meet it. Involved in building the certification guidelines are vendors such as Sophos (Sweep), S&S International (Dr. Solomon's Anti-virus Toolkit), McAfee (VirusScan), Authentec (Alan Solomon); magazines *Virus Bulletin* and *Secure Computing*; and The BSI (German ITSEC Certification Body). Currently, the evaluation process is in the developmental phase. The main areas with which the process is concerned are Standard Documentation, Threat Assessment, Virus Attack Techniques, AVWG Virus Collection, Comprehensive Virus Collection, "Advice Documentation", and Certificate Maintenance Scheme.

Standard Documentation relates to the development of ITSEC documentation which defines minimum security functionality and related information such as functionality class, security target and suitability analysis. These are largely product independent and will be provided by the AVWG. The documents will then be evaluated by a CLEF and approved by the Certification Body (CB) for use in subsequent anti-virus product evaluations. These documents are in final drafting phase at this time and the CLEFS are now being selected.

In the original version of this paper, we discussed the need for product performance to be measured not only by running detection tests on virus collections, but by testing each product's ability to defend against the different attack mechanisms already observed as well. This obviously requires the maintenance of a library of virus attack techniques, and a collection of samples which utilize each of these techniques. As we explained, this is far better than current evaluations, where without specialized knowledge it possible to "certify" a product which provides no protection against a particular attack technique. Attack techniques should include memory-resident operation and disinfection problems.

The ITSEC attempts to address this area in anti-virus product evaluation by proposing to measure the product's performance against the threat not by running and maintaining a large collection of all viruses, but by testing extensively against those viruses which are known to be ITW, and also against a range of different attack strategies. Thus, the tests should reflect not only the product's ability to defend against those viruses which are ITW, but also against the known threat (by evaluating the product's ability to defend against the different techniques used by viruses) and the future threat (by evaluating the developer's ability to track a rapidly changing threat and update the product to deal with it). Currently, the plan is to feed the assessments into the evaluation process, using reports of

incidents, Joe Wells' Wildlist figures, and other available report information. This solution can lead to possible problems as new threat types may be as yet unanalyzed, and the virus itself is not ITW. There is no guarantee as to the time sequence that a virus may be found to exist, be found in the wild, obtained and analyzed by an evaluation or certification service, and its threat type documented. This is illustrated by the recent spate of macro viruses, where there was a noticeable lag between the discovery of the virus (that is, the creation of the threat type), and the implementation of detection and prevention on the part of some developers.

A Virus Attack Techniques Encyclopedia (VATE) has been developed under contract by the AVWG. This is intended to detail all known techniques used by viruses. It is a dynamic document. The VATE will be used to direct more detailed analysis and testing of products; it is a limited distribution document.

Product manufacturers must of course include detection for all viruses, whether or not they are found ITW, because the mere existence of a virus constitutes a threat to users. For this reason, it may be prudent to have both entire libraries and attack strategy suites. The AVWG currently is in the process of establishing a virus collection to support the evaluation process. There is no intention to make this comprehensive, as they have neither the staff nor the expertise to maintain a comprehensive collection. Rather, the collection will contain ITW viruses and examples of viruses illustrating the range of attack techniques covered by the VATE. The anticipated number of viruses is 100-1000. Advice on generation of test suites is still being received. The source of comprehensive virus collection to be used during evaluations is under discussion within the AVWG at the time of writing.

In addition to formal ITSEC documentation, the AVWG recognizes the need for a considerable volume of supporting documentation. There will be the current characterization of the threat (In the Wild list, VATE and virus test suites); general advice to evaluators on how to do product testing; information on special cases; the interpretation of test results; criteria for acceptance. Some of this may be incorporated into the existing UK Manual of Evaluation (UKSP05). Advice documentation to vendors may be included into the UK ITSEC Developers guide (UKSP04). The advice documentation is presently being written, but cannot be completed until the formal requirements such as Functionality Class and Security Target are finalized.

In summary, the functionality tests related to virus detection would be comprised of tests of four types:

1. Common Viruses (determined from AVWG threat tracking)
2. ITW Viruses (determined from AVWG threat tracking, Joe Wells' In the Wild list, other information from the AV community)
3. Virus Attack Techniques (from the VATE)
4. Tests against a "comprehensive" virus collection approved by the AVWG.

An increasing level of rigor would be applied and associated with the commonality of the virus or observed technique, i.e. weighted testing. The current plan is to perform tests with 1&2 listed concurrently and cumulatively and to require a 100% score to pass. The current strategy for zoo testing is 90% for a passing score, based on industry input.

The evaluating body would operate in close contact with the developer of the product currently under evaluation. This means that developers will have to demonstrate that not only are they up to date with the current threat, but that sufficient procedures are in place to monitor the threat as a function of time and update the software to match it. This "vendor evaluation" is something which almost all other evaluations of anti-virus software do not include, and is one of the biggest benefits of the proposed AVWG ITSEC approach. It is also one of the areas which appears to meet with the most resistance within the USA. Another concern which has been cited [12] is regarding the sharing of information between CLEFs: "Even though the UK require that all techniques and lessons learnt from evaluations be documented at the end of an evaluation and made available to the UK evaluation community, it is felt that CLEFs prepare this information from a position of non-disclosure of information which is of a proprietary interest to them. There is concern in the US that UK evaluation, by virtue of their commercial nature, do not encourage the sharing of evaluation techniques amongst the evaluation community".

Finally, there are problems with issues of legal liability. Whereas German law demands someone be liable for failure in certified products, the United States makes specific disclaimers assuming no responsibility. Drawing from Borrett [12], we find "the political implications of legal liability for Europe and North America merits further investigation. In the interim, it may suffice to place an appropriate caveat alongside any US evaluated products which appear in UK Certified Product List publications."

Additionally, it is very difficult to estimate the cost of an evaluation without actually submitting a product: the amount of work needed to be done could vary with the claims made by the developer and the precise nature of the

anti-virus software. Unfortunately, it is still too early for a precise estimate of the costs: until a functionality class has been formally defined. The ITSEC/AVWG hopes to have the evaluation process functional by the end of 1996.

Summary of the Problems

Thus, we have shown that none of the groups above can perform anti-virus software evaluations which fit all the needs of those who are attempting to make a purchasing decision.

Aside from the problems which are unique to each tester, we have discussed several difficulties which are shared between almost all anti-virus software reviewers, testers, evaluators, and certifiers:

- Choice of virus test-set. Does the evaluator have the technical skills necessary to maintain and sort a large virus test-set? Using a scanner to determine infected/non-infected state of files is clearly unacceptable. Viruses must be replicated, and first generation samples are unacceptable. The problems of maintaining a clean, well-ordered virus test-set are discussed further by Bontchev [13]. Creation of the test suite includes the minimum of the following (some taken from [14]):
    - Replication of live boot viruses on all media (5.25 360k diskettes, 5.25 1.2 MB diskettes, 3.5 720k diskettes, 3.5 1.44 MB diskettes, HD master boot record and HD DOS Boot sector).
    - Replication of live file viruses including COM files consisting of normal files, files beginning with JMP instruction, COMMAND.COM, file with many NOPs, files infected multiple times; EXE files consisting of normal files, files with 0 and multiple relocations, Windows applications, compressed files etc.
    - Replication of polymorphic viruses of low polymorphism consisting of 10-10,000 replicants and high polymorphism consisting of at least 10,000 samples (100,000 is not unheard of).
    - Replication of companion viruses, macro viruses and multi-partite samples onto appropriate hosts.
- Time involved. Generation of the test suites described above is dynamic, as new viruses are found daily. Additionally, testing is another time consuming process. Testing includes but is not limited to cleaning of memory and media, checking of system integrity, infection of the victim files and/or boot sectors, checking replication potential of the replicants, scanning and report generating.
- Bias. Is the evaluator in any way associated with one of the products which is reviewed? Were the samples obtained from a particular vendor?
- Which aspects of the product have been tested? Were the test results weighted, and if so, how?
- Which tests measure the efficacy of the disinfection routines, the efficiency of memory scanning or the problem of false positives, user interface and documentation; how were they conducted and how were the results interpreted?
- Has the product been tested for compatibility with your system/network and are additional tools provided?
- Has company support/tech support been evaluated? Areas of company support which should be evaluated are response time via telephone and electronic media, completeness of information provided and follow-up.

In summary, the problems with anti-virus product evaluation are many. The ITSEC approach provides some suggestions as for how we can adapt and use their fundamental approach to evaluating products, but, as we have seen above, even this is not a complete system.

Conclusion

We have examined the current evaluation methods applied to anti-virus software, and demonstrated that at best they only cover some of the areas which a complete evaluation of a product should cover. We believe that the current plans for anti-virus software evaluation in the ITSEC will address many of these issues, and that when the system is fully operational it will provide the prospective purchaser with some guarantee of software functionality, and moreover some measure of the developer's commitment to continue to meet a rapidly changing threat. We note that the ITSEC methods are not a cure all, and that even if plans of the AVWG are implemented, there are still areas which do not appear to be satisfactorily addressed.

While we recognize the problems of the ITSEC, we believe that the underlying methodology is sound, and that by drawing from the positive addition of new forms of functionality testing and product assessment, we are hopeful that in the near future anti-virus product evaluators of all types will have a more solid knowledge base from which to draw.

We believe that not only is it impractical to perform all aspects of product evaluation in-house, but that doing so can be directly damaging, as it is possible to select a product for entirely the wrong reasons. Thus, the reader is urged to use a wide variety of sources of information. Much of the information outlined above can be obtained at little or no cost; by understanding the strengths and weaknesses of each different evaluation you are in a position to extract figures which are relevant to determining which product is most suitable for your company.

It is still necessary to cull information from a number of sources to select a product which not only fulfills the functionality which is required by your policy (speed, transparency, cost), but also provides an adequate defense against the threat (virus detection). This can only be done by carefully considering your anti-virus policy and creating a list of requirements which your chosen product must fulfill. The first criterion remains "how well does the product detect viruses you are likely to encounter".

Keep in mind, that as the user of any anti-virus product evaluation service, you should be encouraged to contact the evaluator to get any relevant information not contained within the review [7]: only by recognizing the strengths and weaknesses of existing product evaluation schemes can we hope to use the currently-available information to our advantage when attempting to choose the "right" product for your environment.

Bibliography

[1] In Laine [3].

[2] Richard Ford, Sarah Gordon. *Real-World Anti-Virus Product Review and Evaluation,* IVPC 95 Conference Proceedings.

[3] Kari Laine. *The Cult of Anti-Virus Testing*. EICAR 1994 Conference proceedings.

[4] Sarah Tanner. *A Reader's Guide to Reviews*, Virus News International, November 1993.

[5] Sarah Gordon. "*Is a Good Virus Simulator Still a Bad Idea*?" Preprint.

[6] Virus Bulletin. *Disinfection: Worth the Risk?,* September 1994.

[7] Sarah Gordon. *Evaluating the Evaluators*, Virus News International, July and August 1993.

[8] NCSC, *Introduction to Certification and Accreditation,* Rainbow series, NCSC-TG029, January 1994.

[9] Richard Ford. Private communication. 1995.

[10] Marko Helenius. *Anti-Virus Scanner Analysis by Using The 'In the Wild' Test Set*. EICAR 1994 Conference proceedings.

[11] VTC. Anti-virus scanners test protocol. Virus Test Center, University of Hamburg, Germany.

[12] Alan Borrett. *A Perspective of Evaluation in the UK Versus the US*. Proceedings 18th National Information Systems Security Conference. Baltimore, Maryland. October 1995.

[13] Vesselin Bontchev. *Analysis and Maintenance of a Clean Virus Library*, Virus Bulletin Conference Proceedings , Amsterdam, 1993.

[14] Vesselin Bontchev, Klaus Brunnstein, Wolf-Dieter Jahn. *Towards Antivirus Quality Evaluation*. Virus Test Center, Faculty for Informatics. University of Hamburg, Germany. From the Proceedings of the 3rd EICAR Conference, Munich Germany. December 1992.

About the Authors:

Sarah Gordon is Security Analyst for Command Software Systems, where she works in Research and Development, maintaining the virus library. Dr. Richard Ford is Technical Director for Command Software Systems, and works in the area of product testing. Both Ms. Gordon and Dr. Ford have extensive experience in testing anti-virus products and have published numerous articles on computer viruses and other computer security topics. They may be reached respectively at sgordon@dockmaster.ncsc.mil and rford@commandcom.com.

Acknowledgements: Megan Alexander, Command Software Systems.