

WHY ANTI-MALWARE PRODUCT TESTING MATTERS

The Anti-Malware Testing Standards Organization

David Harley, Small Blue-Green World

Dennis Batchelder, Microsoft Corporation

Dodi Glenn, PC PitStop

John Hawes, Virus Bulletin

Mark Kennedy, Symantec Corporation

Peter Stelzhammer, AV-Comparatives

Righard Zwienenberg, ESET

ABSTRACT

The internet and our computing devices are now irretrievably intertwined with the modern world. Threats to our devices and data have become a fact of life, and at the heart of just about all of these threats is some form of malware. At personal, business and nation-state levels, we all rely on anti-malware solutions to mitigate the dangers we face and to protect the privacy and integrity of our information and communications. As those solutions evolve and improve to combat the existing battery of weapons deployed by bad actors, so those bad actors tune and improve their attack techniques, leading to an unending arms race.

Anti-malware solutions have grown and diversified, both within individual products and across the market space. Individuals, businesses and institutions need to know which solutions will provide them with the best possible protection, and in all but the largest and most well-funded organizations (i.e. those competent and well-enough resourced to implement their internal product evaluation processes) one of the most important factors influencing that choice is publicly available test data. Anti-malware testing is a highly complex and difficult business, and poor testing can lead to purchasing choices based on inaccurate or even misleading information. By not having the right protections in place, consumers are risking their personal security and privacy, and businesses are risking their own security and privacy and that of their users, customers and partners.

AMTSO was formed in 2008 as a coalition of leading anti-malware testers, vendors and academics to provide a form of industry self-regulation for product testing. In the past eight years, the quality of anti-malware testing for traditional solutions has increased dramatically. AMTSO's work on agreeing and documenting sound methodologies, and that documentation being made freely available to aspiring testers and the general public, has significantly contributed to this rise in quality. However, the ongoing evolution of malware and anti-malware means that there remains much work to be done to ensure testing is relevant as well as properly designed and conducted, giving us the most accurate and reliable data and allowing us to make the best choices to protect ourselves and the organizations of which we are a part.

A significant area of ongoing work is the inclusion in advanced tests of weightings for regional and impact-based significance, based on accurate, global telemetry data sourced from as wide and diverse a field of data providers as possible. Another area of major interest is exploring the inclusion of real-world efficacy data gathered by anti-malware providers themselves.

We will analyze the effects of including such weightings and data compared to existing standard practice for tests and show how much is changed when the additional data are taken into account, as part of AMTSO's ongoing mission to improve testing as a whole.

These improvements to the relevance of test results should help reduce the current trust gap between consumers, who often see conflicting test results as an indication of at best incompetence and at worst collusion with the vendors of security solutions. We will also look at the requirements of statistical significance in tests, with data on just how many iterations are required to ensure proper coverage of a hugely diverse and rapidly changing and growing threat landscape. Once we can see more clearly how well our security solutions are able to protect us from the latest attacks, we will all be in a better position to judge our exposure to risk from malware

INTRODUCTION

Malware, by AMTSO's definition, includes all forms of software or other electronic data designed to, or otherwise capable of, infiltrating and/or damaging a computer system or network, including computer viruses, worms, APTs, trojan horses, spyware, and ransomware. These threats can and do infringe upon the privacy of individuals and organizations in a number of ways, such as stealing credentials to bypass authentication procedures, exfiltrating sensitive data, and so on. Anti-malware solutions work in various ways to combat malware, and the most effective products have moved on from simple pattern-matching in static files to include advanced heuristic techniques, behavioral and anomaly analysis, big data analysis and multi-point activity tracking. A wealth of additional modules, components, layers and techniques all promise to reduce the risk of something malicious sneaking past.

Evaluating and comparing all these methods and products is a highly specialized business, with the slightest error in test design easily biasing results or rendering them useless. The traditional analogy is with car safety testing. Imagine a car safety review that concludes that Car A has the best airbags, and is therefore the safest. Unfortunately for the consumer who relies on that review and purchases Car A, while the airbags may in fact be superb, the review may not mention that it has neither seatbelts nor anti-lock brakes. Tests focusing on specific protective layers and ignoring others were one of the pain points which led to the founding of AMTSO, and for the most part, at least in the leading test labs, such issues have been dealt with.

With the fast flux of the modern malware landscape, and the modern market for protective solutions, we must now imagine a car safety test where all aspects of the car can be remotely updated or entirely reconfigured by the manufacturer from moment to moment, where the road surface can switch from flat tarmac to quicksand, deep water or a massive snow-capped mountain without warning, and where the crash test dummy itself regularly changes its shape and physical properties.

This is basically what is required of the modern anti-malware tester. "Products" are for the most part closer to services these days, relying on cloud systems to provide the rapid response needed to keep tabs on the latest threats and to power the number-crunching of pooled data from install bases numbering into the hundreds of millions. This means that the thing you are testing is never the same from moment to moment. Malware emerges at a lightning pace, as those creating it target new audiences and new channels, or simply tweak their wares to avoid detection, so that today's major threat is tomorrow's old news. Even if its components are carefully selected, the test-bed of malware samples used in a test can also only reflect a single moment of reality.

Testing rapidly-changing products against an ever-evolving threat landscape can only ever show performance at the specific point in time at which the test took place - not only is the target moving, but the platform we are shooting from is moving too - so it is of vital importance that those

snapshots – or each of those snapshots in a longitudinal test sequence – show the most accurate picture possible of how well products are protecting us against the threats that matter the most.

CAN PREVALENCE OR GEOGRAPHY INCREASE TEST RELEVANCE?

In traditional anti-malware testing, all samples have generally been considered to have equal importance. Testers acquire samples (which may be executable or document files, live URLs on the internet, or simply snippets of exploit code), and measure how many of them are spotted by the solutions they are testing. If 90% of them are detected, the product is considered 90% effective.

However, some of those samples may have only impacted tiny numbers of people in the real world, while others could have targeted millions. So a product which picks up 90% of all samples, but misses the most widespread and dangerous ones, will appear more effective than one which only spots that most significant 10%.

Studies conducted by AMTSO member test labs have clearly shown this bias, with the ranking of some products changing dramatically when the scores are weighted according to the prevalence of samples. Measuring that prevalence is itself no easy matter, of course, with the main source of data being the developers of anti-malware solutions themselves, who track and monitor which threats are targeting their users, in which regions and at what scale. To achieve a truly accurate picture, data from multiple sources must be acquired and correlated, again no easy task given the diversity in the format and regularity of various sources.

Further, file prevalence alone is not useful without taking into consideration the context of family prevalence in the ecosystem. A single executable file may only be seen by a single user and thus seem of minimal significance, but it may represent part of a highly widespread attack which serves each visitor to a given URL with a file showing slight differences at the binary level but identical behaviors. In that case, it should be treated as a representative of that group as a whole and should share the prevalence of the entire family.

Reliance on a single vendor's telemetry, especially if that telemetry was localized or from a small customer base, can skew results for one or more vendors in a test. Geography is another important consideration, because some malware may target only one region of the world and might not be detected by, or matter to, every vendor if they don't have a strong customer base in that region. For more accurate modeling using prevalence and geographic data, the anti-malware industry needs more vendors and academics to submit telemetry data, and those data need to be derived from consistent reporting methodologies.

CAN REAL-WORLD RESULTS INCREASE TEST RELEVANCE?

In traditional anti-malware testing, comparative tests are performed in a laboratory setting that tries to reproduce real world conditions as closely as possible. This approach requires thoughtful selection of samples to remove bias in favor of some vendors and geographies, and it requires the test consumer to factor the lab's testing approach into what the targeted consumer group is experiencing in the real world.

Another approach that could be taken is to measure how anti-malware vendors handle malware in the real world. Most vendors collect data on how their products perform, including what they miss initially but detect post-infection. Sharing such data and allowing testers to analyze and verify it could be another useful metric for measuring product efficacy.

This would require either independent observers or self-reporting by the anti-malware vendors or their customers, which are not easy tasks to accomplish. An alternative route to similar efficacy measures from the real world might be “field trials”, where systems in everyday use by real people are periodically analyzed for signs of infiltration.

ADVANCING TESTING QUALITY

We believe that over the past eight years, the continued collaboration of leading stakeholders in the anti-malware space has led to the greatest advancements in improving the quality of anti-malware testing as a whole. In some ways, AMTSO has allowed these stakeholders to step outside of their silos to collectively tackle extremely complex testing issues, including testing of products which promise protection from highly targeted “APT”-type threats. Thus, the world’s top anti-malware testers are sitting at the table with not only the top anti-malware vendors from the US, but from Europe, China, India, and Russia, as well as academics from all over the world. AMTSO represents a critical community of cutting-edge researchers who are collectively tackling some of the most complex topics in anti-malware.

Most recently, AMTSO has developed a Real-Time Threat List (RTTL), providing a common platform for sharing telemetry on files, one of the core aims of which is to meet the need for complete and accurate data to support prevalence-weighted and region-specific testing. Vendors should be incentivized to use the RTTL to share their samples and telemetry to ensure that files and families affecting end users are properly represented in tests. Testers should be incentivized to use the RTTL to select which samples to test against. The inclusion of this data will result in more accurate anti-malware test scores and a more informed public that can make better choices about protection products.

AMTSO was founded to ensure that anti-malware products are tested in a way that is relevant and objective, so that end users are provided reliable information about the solutions they employ. Quite simply, we believe that testing matters to security, and security matters to people. Anti-malware solutions remain at the core of any protection profile, and while we remain agnostic as to the type of solution that is employed, a solution is always necessary, all solutions are testable, and all tests must be objective, relevant and fair.