EVALUATING ANTI-VIRUS PRODUCTS WITH FIELD STUDIES

Fanny Lalonde-Lévesque, Carlton R. Davis & José M. Fernandez
École Polytechnique de Montréal, Montreal, Canada

Email {fanny.lalonde-levesque, carlton.davis, jose.fernandez}@polymtl.ca

Anil Somayaji
Carleton University, Ottawa, Canada

Email soma@scs.carleton.ca

ABSTRACT

The evaluation of anti-virus (AV) products is a vital component in helping the industry develop better products that match the evolving malware threats, and in helping users to make informed decisions about product selection. Traditional evaluation methods involve testing in laboratory environments under various threat scenarios, some more realistic than others. In this paper, we present a first study of an alternative method of product evaluation involving real users. We report on the performance of one AV product in a four-month field study involving 50 users, using their own machines in their normal daily business. In addition, we cross-analyse detection data with user behaviour and demographic characteristics in order to determine what factors are conducive to higher risks of infection. We conclude by discussing options that would allow this methodology to migrate to multi-product evaluations, and become a repeatable and viable alternative to traditional lab-based comparative testing.

1. INTRODUCTION

Malicious activity on the Internet is growing at an unprecedented rate. According to the latest report from *McAfee*, Q1 2012 had the largest number of PC-based malicious programs detected per quarter in the last four years [1]. The number and the variety of threats are increasing and attackers are also adapting the techniques employed to infect their victim's computers. In fact, increasingly infections occur because users are enticed into taking an action that leads their computers to become infected, such as opening an email attachment, visiting a malicious website, or even willingly installing a piece of software whose true intention they ignore. On the defensive side, we have gone in the last decades from anti-virus (AV) products essentially based on signature detection to complex security software combining multiple protection techniques.

This increase in complexity of both the threat and AV products has made it more difficult to accurately evaluate the performance of the latter. Typical evaluation methods are based on automated tests in controlled environments. Yet these testing methodologies do not reflect the performance of products in real-life situations because they do not always take into account the influence on performance of many factors such as user interactions, machine configuration and environment, and evolving threats. The variety of different threats, possible configurations, and user actions makes it very difficult to explore the performance of AV products in the lab under all possible combinations of these factors, and so the results of such tests can often be biased. This has been particularly controversial with respect to the selection of threats against which the products should be tested, the so-called *sample selection problem.*

On the other hand, and due to the fact that user actions now play a much more preponderant role in the process of infection, it becomes important to understand how general user behaviour or even user characteristics affect the risk of infection by malware. For example, while it might be intuitive to think that infection by drive-by-download will be more prevalent in users who do more browsing on less reputable websites, it is important to confirm whether this is truly the case. In other words, beyond the impact of the AV alone, we seek to understand the effect of user behaviour on the risk of infection.

To address both of these issues, i.e. performance of AV in a realistic and representative environment and correlating the risk of infection with user behaviour and characteristics, a novel methodology of AV performance evaluation [2] was proposed in 2009 that takes inspiration from the clinical trials used in medical and pharmacological research. This approach involves conducting long-term field studies where the participants use their own computers in normal life, which are protected by an AV product and instrumented with special purpose tools that allow us to determine *a posteriori* the effectiveness of the AV, while also collecting data about the usage patterns of the participants. In this paper, we report on the first study of this kind conducted at the École Polytechnique de Montréal from October 2011 to April 2012. The study involved 50 participants, who were monitored for a period of four months. We describe in Section 2 the methodology used in the study. In Section 3, we describe and discuss some of the preliminary results of the study, both in terms of AV performance and user behaviour. We describe in Section 4 how our methodology could be adapted to conduct large-scale comparative studies with hundreds of users and several different AV products. We conclude in Section 5 by summarizing the results of the study, the lessons learned and directions for future work.

2. STUDY DESCRIPTION

In order to prove the feasibility of our approach, we conducted a four-month proof-of-concept study involving 50 participants. The details of the methodology have been published elsewhere [3], but we provide a summary description here. The study includes monitoring real-world computer usage through diagnostics and logging tools, monthly interviews and questionnaires, and in-depth investigation of any potential infections.

The study had the following goals:

i To develop an effective methodology to evaluate anti-virus products in a real-world environment;

ii To determine how malware infects computer systems and identify the source of malware infections;

iii To determine how factors such as the configuration of the system, the environment in which the system is used, and user behaviour affect the probability of infection of a system.

The 50 participants were recruited through posters and newspaper advertisements on the Université de Montréal campus (where the École Polytechnique is located). A short online questionnaire was used to collect initial demographic information. Using these profiles, we categorized interested volunteers based on their gender, age group, status and field of work/study. We randomly chose a sample from each category in order to have a diverse and representative sample of users that included students and employees from various fields.

2.1 Equipment

We provided laptops with identical configuration to the participants. The following software was installed: *Windows 7 Home Premium*; the AV product to be evaluated, i.e. *Trend Micro Titanium Maximum Security* (*Trend Micro*'s premium product for home users); monitoring and diagnostic tools including *HijackThis*, *ProcessExplorer*, *Autoruns*, *SpyBHORemover*, *SpyDLLRemover*, *tshark*, *WinPrefetchView*, *WhatChanged*; and custom Perl scripts which we developed.

We used the scripts to automate the execution of the tools as well as for compiling statistical data regarding the system configuration, the environments in which the system is used, and the manner in which the system is used. The data compiled by our scripts included: the list of applications installed and the list of applications for which updates were available; the number and the type of websites visited; the number and the type of files downloaded; the list of browser plug-ins installed; the number of different hosts to which the laptop communicated per day; the list of the different locations from which the laptop established connection to the Internet; the average number of hours per day the laptop was connected to the Internet; and the average number of hours per day the laptop was on.

The AV product was centrally managed on our server. All the AV clients installed on the laptops were sending relevant information to our server about any malware detected or suspected infections as they occurred.

Before deployment, we benchmarked the laptops by running tools and recording the output. The recorded information included: a hash of all files plus information about whether the files were signed; a list of auto-start programs; a list of processes; a list of registry keys; a list of browser helper objects (BHO); a list of the files loaded during the booting process; and a list of the pre-fetch files.

In order to avoid biases in user behaviour and at the same time limit the liability of the university, the laptops were sold to the participants at an advantageous, below retail-market price, with the laptops remaining in the users' possession at the end of the study.

2.2 Experimental protocol

The study consisted of five in-person sessions: an initial session where participants received their laptop and instructions, followed by monthly one- to two-hour sessions where we performed analysis to determine whether the laptop was infected.

To encourage the participants to remain in the study until its end, we paid them to attend the monthly in-person sessions. If participants completed all required sessions, the entire cost of the laptop would be reimbursed, along with an additional compensation. We encouraged participants to configure their laptop as they desired and use it as they would normally use their own computer. The only restrictions applied during the experiment were that the participants could not format the hard drive, replace the operating system, create a disk partition, install any other AV product on the laptop, or remove our software and tools.

Each month, participants booked an appointment via an online calendar system hosted on our server. During these monthly sessions, participants completed an online questionnaire about their computer usage and experience. The questionnaire was intended to assess the participants' experience with the AV product and gain insights about how the laptops were used. Meanwhile, the experimenter collected the local data compiled by the automated scripts. Diagnostics tools were also executed on the laptop to determine if an infection was suspected. If the AV product detected any malware over the course of the month, or if our diagnostics tools indicated that the laptop may be infected, we requested additional written consent from the participant to collect specific data, such as the browser history, the tshark log files (i.e. network traffic data), and the suspected file(s), in order to help us identify the means and the source of the infection.

At their last visit, participants completed an online survey about their experience during the study. The aim of this final survey was to identify activities or mindsets that may have unduly influenced the experimental results. We requested that participants keep the experiment data stored on their laptops for an additional three months, so that if we discovered that further analysis was necessary, we could contact them and seek their permission to collect and analyse the relevant data. Finally, we provided them with a procedure for removing the diagnostic tools and the scripts, as well as the experiment data stored on their laptop.

It is important to note that since the study involved human subjects, the entire project had to undergo strict review by the university's Ethics Review Board. The Board imposed certain restrictions on the study such as limits to the type of (potentially) personal information kept, the length of time data could be kept, the purpose of the research, and adequate remuneration for participation in the study.

2.3 Cost

Due to the involvement of users and equipment, the cost structure of running this kind of experiment is quite different from laboratory experiments and merits discussion. The costs incurred can be divided in the following three categories:

Initial expenses: In order to provide laptops for every user, we bought 50 quality laptops at $375 each, for a total of $18,750.

However, this purchase cost was partially covered by the users that had to buy their laptop at $350 (well below the $475-$500 retail price of the same laptop).

Operating expenses: Technical work was required for the collection of the data during the monthly sessions. We hired a student who worked on this task for a total of 95 hours, at a cost of $1,800. We also bought a cell phone for the study in order to provide a designated telephone number to the participants. The purchase of the telephone and the plan cost us around $200.

User compensation: Each user received a free one-year licence for the AV product installed on his computer. The cost of the licences was covered by the participating AV vendor. Users also received a $50 compensation for attending each of the first three sessions, a $100 compensation for attending the fourth session and a bonus of $150 for attending every session. Therefore, users had the opportunity to receive up to $400, resulting in a maximum cost of $20,000 for 50 users.

Overall, the total cost of our four-month field study with 50 users was to $23,250. This covered the purchase of 50 laptops, the compensation we paid the participants and technical work. Of course, this does not include the time spent by the researchers on the design, supervision and analysis of the study results.

3. RESULTS

3.1 Threat detections by AV

During the four-month study, 380 files were detected by the AV product being evaluated on 19 different user machines. However, some of these files were detected twice or more on the same user machine. Without these repetitions, we obtain a total of 95 detections. Figure 1 shows the distribution of the detections without repetition for each month of the study. We can see that the level of detections is very similar for each month, contradicting the hypothesis that users are most at risk when they first start using their machines.

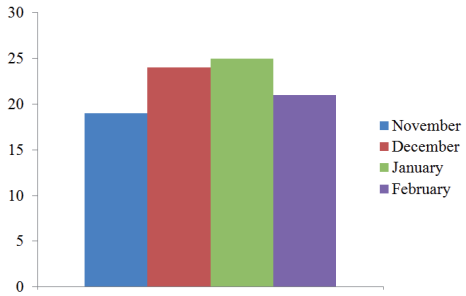

Figure 1: Malware detections by month.

Each of these detections was classified based on the information provided by the AV product. Figure 2 shows the distribution of the malware detections by type. As we can see, almost all detections were classified as trojans, while virus and adware have a relatively weak representation.

These figures are somewhat similar to those reported for overall infections by other AV vendors. For example, the first 2012

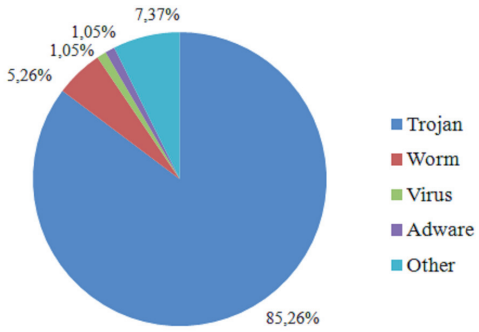

Figure 2: Malware detections by type.

quarterly report from *Panda Security* [4], indicates that trojans account for most of the detections with a ratio of 63.30%, while worms, viruses, adware and other have respectively ratios of 8.39%, 7.90%, 7.81% and 9.60%.

Nonetheless, the differences with our results could be partially attributed to the differences in the classification methods. For example, a file may be classified as a trojan by the AV product being evaluated and as a virus by another product. Furthermore, statistical error could be significant since our results are only based on a collection of 95 malware samples, while those of *Panda Security* are probably based on thousands of different samples.

In terms of detection mechanism, 93 of these 95 detections were made through real-time scan and only two of them through a manual scan initiated by the users. These results are good considering that the AV products should be able to detect a maximum of threats with minimal user intervention.

In terms of product response, the AV product quarantined 78 files, failed to quarantine 10 files and encrypted two files. For the last five detections, the AV product reported them as 'a potential security risk'.

While a detailed analysis of the causes and means by which these threats ended up on the computer still remain, we already know that 17 of them came from external devices.

3.2 Missed detections

The experimental protocol describes in detail the procedure for identifying and classifying during the monthly visits suspicious files that were not detected by the AV. This process of identification and classification is based on user reporting of suspicious machine behaviour, the analysis of logs from the monitoring tools, the results of queries to on-line processes, file and start-up program databases (obtained automatically by scripts that we wrote), and any other relevant piece of information that the technician conducting the review might deem relevant.

Suspicious files found on the computer were categorized into four categories: dangerous, suspicious, safe and unrated. All files marked as dangerous, suspicious and unrated were subject to a more in-depth analysis. When we suspected one of these

files to be dangerous, additional data were collected with the consent of the user, including the actual browsing history, the suspicious files and other related files present on the computer, etc.

Our analysis led us to identify 20 possible infections on the laptops of 12 different users. In terms of detection method, the most useful tool was *HijackThis*, which was involved in identifying 18 of the suspected infections, with *SpyBHORemover* helping us find one more. The last suspected infection was reported by the user, who called the project manager on the duty cell phone when he suspected that his machine had been infected. All the suspicious files were captured during the monthly visits, except for the user-reported suspected infection. While the logs show the location and filename, the file could not be retrieved as it seems that the suspected malware uninstalled itself between the time the user called in and the following lab visit.

All files captured (19 out of 20) were later scanned with the evaluated AV product in order to see if they would be detected *a posteriori*. Even several months after the end of the experiment, none of them were being detected by the AV product or identified as a potential threat. In addition, we also scanned the captured files *a posteriori* with the *VirusTotal* service, in order to compare the results obtained by various AV products and also to compare theses results with those obtained a few months before. Additionally, we searched the Internet to find as many details as we could for each of these 20 detections. As a result of this analysis, we classified two of the samples as 'clean', seven as 'unwanted software', nine as 'adware', one as 'definite malware', and one is suspected to be malware ('maybe malware').

The adware samples detected were either BHO or toolbars. In all cases, they were not willingly installed by the users. Their effects varied from changing the web browser home page, to redirecting web searches or displaying advertisements. Further analysis will be required in order to determine if these adware programs are indeed malicious, in that they show additional behaviour that might have further consequences for the user than those described (e.g. theft of personal/private information, etc.).

While we have not yet analysed in detail the two suspected malware samples, we know that one of them is a confirmed rogueware. As mentioned previously, the corresponding user contacted us to inform us that his laptop was probably infected. It turned out that the laptop was infected with the fake AV Security Scanner. Windows were showing up on a regular basis to inform the user that harmful software had been detected on his computer and every application started was killed, except for web browsers. In order to get rid of these infections, the user was invited to register and was asked to give his contact and payment information. At that moment, the user suspected that he may be infected and made contact with us. As explained before, since the files disappeared from the computer before the monthly visit, it was not possible for us to verify whether the AV product detects this threat *a posteriori*.

While we have only detected two potentially malicious files, we expect to look further at the data in order to find more infections that we could have missed during the monthly sessions.

3.3 User feedback

Every time a malware sample or threat was detected on a computer, either by the AV product or by us, the user was asked to fill out a short online questionnaire. This questionnaire was intended to seek more information about the infection and the user's reaction.

When asked if they had observed strange behaviour while operating their computer, 40% of the respondents said yes, 55% said no and 5% said they did not know. For those who answered yes, the most often observed behaviours were the slow down of the computer, the occurrence of pop-up windows, and problems with the web browsers such as redirection and changes of home page.

Half of the respondents remembered seeing a prompt window from the AV informing them that a security problem had been detected on their computer. Among these users, 35% said they were concerned about the security of their computer, 30% felt secure, 20% said they were annoyed at the interruption, and 15% felt confused.

3.4 User profiling and behaviour

One of the hypotheses that we tested is whether an increase in certain types of user behaviour leads to a higher probability of the users' system being infected with malware or spyware/adware. We also wanted to determine whether user demographic characteristics had any bearing on incidence of infection. To test these hypotheses, we identified several variables that could potentially have a bearing on incidences of infection, and we measured these variables during the course of our four-month study. In total, more than 50 variables were considered and analysed. Table 1 presents statistics computed for 10 of these variables that seem to have a stronger correlation with incidences of infection. Many of these behaviour statistics concern number and types of websites visited. These statistics were obtained from the AV product, through the reporting data sent to the update server we ran for the participants; in other words, the statistics rely on the classification of visited websites provided by the AV vendor.

As indicated in this table, we also divided the participants into two groups. The first group consisted of the 23 participants whose laptops were subject to a threat. This 'at-risk' group includes both users whose AV reported a threat and users for which we suspected a missed detection. The other 'low-risk' group consisted of the remaining 27 participants whose laptops were not subjected to any detected threats. We computed the mean, maximum, minimum and median for both groups. As Table 1 indicates, for all the variables, the mean for the at-risk group was greater than the mean for the low-risk group. Similarly, for all the variables except two (total uptime and number of files downloaded), the maximum value for the at-risk group was greater than the maximum value for the low-risk group.

We also compiled demographic information about the participants, such that we might be able to determine whether participants with given demographic characteristics have a higher incidence of infection. We computed the mode for the selected demographic variables for the 50 participants, and also separately for both groups.

				'At-risk' group (threat detected)		'Low-risk' group (no detected threat)	
Independent variables	**Mean (Median)**	**Std. Dev.**	**Correl. Coef.**	**Mean (Median)**	**Max (Min)**	**Mean (Median)**	**Max (Min)**
Number of hosts contacted	6722 (4679)	6128.266	0.2883	8617 (7074)	20302 (0)	5108 (3078)	19501 (9)
Total time online (hours)	50.83 (33.30)	47.602	0.3248	67.41 (58.84)	203.29 (3.88)	37.7 (26.15)	134.79 (3.86)
Total uptime (hours)	358 (366)	164.631	0.1734	388 (381)	603 (97)	332 (292)	691 (40)
Browser history entries	3955 (2508)	3974.540	0.4223	5755 (3582)	16772 (693)	2421 (2136)	6553 (32)
Number of untested or dangerous sites visited	746 (349)	1277.340	0.2592	1101 (512)	8730 (85)	443 (320)	1367 (5)
Number of adult sites visited	72 (3)	220.943	0.2415	129 (6)	1216 (0)	23 (1)	273 (0)
Number of software/file download sites visited	47 (19)	74.531	0.4604	84 (76)	423 (5)	16 (12)	50 (0)
Number of streaming media sites visited	159 (73)	244.288	0.4319	272 (84)	1055 (15)	63 (37)	230 (0)
Number of games sites visited	45 (14)	80.169	0.1607	39 (32)	381 (0)	34 (8)	295 (0)
Number of files downloaded	489 (296)	550.941	0.1378	545 (303)	2352 (19)	442 (286)	2644 (12)

Table 1: Statistics for selected variables related to user behaviour (per session values).

For the variable related to gender, the participants sample set consisted of 30 males and 20 females; so, the proportion of male to female was 1.50:1. Figure 3 shows that the first group consisted of 14 males and nine females, i.e. a ratio of 1.56:1; whereas, the second had 16 males and 11 females, i.e. a ratio of 1.45:1. While the group subject to threats had a slightly higher male-to-female ratio, this difference is small and not considered statistically significant. In other words, there does not seem to be a significantly higher risk of exposure for males vs. females.

For the age group variable, the mode was the same (18-24 years old) for all three groups (at-risk, low-risk and whole population), except that for the at-risk group, 25-30 had the same frequency as 18-24. The proportion of 18 to 24-year-olds to other age groups in the overall population was 18 to 32 (0.56:1), whereas for the at-risk group it was 8 to 15 (0.53:1) and 11 to 16 (0.69:1) for the low-risk group. It would thus seem that, contrary to what might have been expected, the younger group (18 to 24-year-olds) is slightly less at risk of infection.

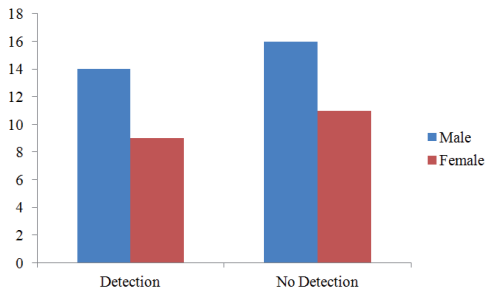

Figure 3: Gender distribution by group.

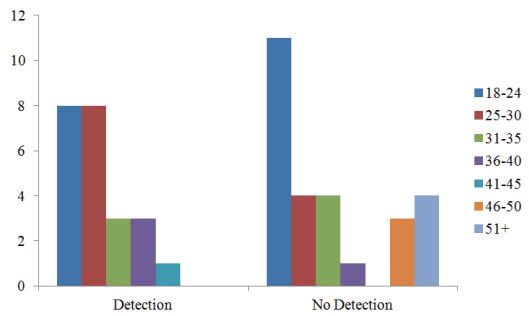

Figure 4: Age group distribution by group.

For the 25-30 age range, the proportion of this age group in the population was 12 to 38 (0.32:1), while it was 8 to 15 (0.53:1) for the at-risk group and 4 to 23 (0.17:1) in the low-risk group. This would seem to indicate that participants in the 25 to 30-year-old age group were more prone to infection than the younger 18-24 crowd or even the population at large. In summary, people in the 25-30 age range seem to be the most at-risk segment, while 18 to 24-year-olds are less at risk than the rest of the population.

There are many explanations for this correlation, some of which might have to do with correlation of age with computer know-how or usage pattern. At this time, we are still conducting analysis on correlation of detections with other demographic characteristics, such as computer expertise, and user behaviour metrics such as those reported by the user during monthly surveys. We hope that further analysis of these variables might shed more light on the root factors affecting infection.

4. FIELD STUDIES FOR COMPARATIVE AV TESTING

This single-product field study methodology might be suitable for AV vendors seeking to understand how their products perform in real-world situations in combination with the user, and might help identify which aspect of the product (user interface, detection, remediation, etc.) could be improved. Furthermore, it can help understand what characteristics of user behaviour lead to higher risks of infection, which could help lead to more focused and effective user training and indoctrination policies and programmes, or even help insurance companies assess relative risk in IT insurance policies.

Nonetheless, the reality is that a large portion of AV performance tests performed today are aimed at identifying which AV products perform better than the others, whether it is to allow users to make a more educated choice of product or to help AV companies determine R&D and marketing strategies. Given the advantages in term or realism and availability of user data of field tests such as the one described here, the natural question is whether a *comparative* field study of AV products can be conducted.

One of the major issues in traditional (lab-based) comparative AV testing is to ensure that all AV products are evaluated under exactly the same conditions. Not only should they be tested in the same environment, but they should also be exposed to the same threats at the same time. While it is relatively easy to guarantee these conditions when tests are performed within a controlled environment, it is a challenge in the case of field studies. The exposition of the product to threats cannot be controlled as it is user driven. On the other hand, field studies are inherently unbiased in that threats applied to each AV are independently 'chosen' by users, who have no vested interest in the test results. Furthermore, the law of large numbers guarantees that for sufficiently large populations each product will be exposed to a large enough sampling of threats to make the results statistically significant. In order words, and to eliminate bias in user-driven threat selection, such comparative field studies should be conducted with a large enough population and over a long enough period of time. We postulate that in order to generate statistically significant results such a study should include at least 200 participants and last at least four months.

We suggest here two different models that could be used to conduct such a comparative field study. The first model is similar to the study described here, in that it implies that users will have to be physically present at monthly lab visits, while the second model makes use of remote access to obtain the necessary monitoring data. Based on these two models, we present a sketch of field studies for comparative AV testing based on each one. The common characteristics of both scenarios are the following:

Duration: Four months.

Population: Minimum 200 users running *Windows 7* or *Windows 8* on their home machine with Internet access.

Sample: Minimum 200 users recruited based on their demographic profile (age, gender, status) in order to have a representative sample. As for the previous study, users will have to consent to the installation of monitoring tools on their computer. They will not be allowed to replace the operating system, install any other AV product on their computer or delete our tools. Users will be offered free technical support and they will be allowed to drop out of the study at any time.

Product: Four major AV products will be subject to the evaluation. Each AV product will be assigned a minimum of 50 users, chosen in order to achieve maximum uniformity of demographic characteristics for each AV product.

Equipment: Unlike in the proof-of-concept study, no specific equipment will have to be purchased because the users will use their own personal computer. However, minimum specifications will be required for the memory, the hard drive space and the processor.

Model 1

As for our previous study, users will have to attend five in-person sessions each lasting one hour: an initial session and four monthly sessions. The initial session will allow us to install the AV product to be evaluated, as well as our scripts, software and diagnostic tools. It will also allow us to benchmark the computer while the users will have to sign the consent form and complete an initial questionnaire. During the monthly sessions, users will have to complete an online questionnaire, while we will collect local data compiled on their computer. Diagnostic tools will also be executed on their computer to determine if infection is suspected. If the AV product detects any malware over the course of the month, or if our diagnostic tools indicate that the computer may be infected, an additional written consent form will be requested from the user to collect additional data that will help us identify the means and the source of the infection.

Initial expenses: Because this model is an adaptation of our previous study, the only initial expenses involved should be the improvement of our scripts and tools. Considering one month of work from a technical assistant, the initial expenses should be $5,000.

Operating expenses: Technical work will be required, estimated at a total of five hours per user for the sessions, one hour per user for technical support, and one hour per user for administrative work. With a total of 1,400 hours for 200 users, and a $20 compensation, the total expenses should be around $28,000.

User compensation: Each user will receive a free one-year licence for the AV product installed on his computer. We expect the cost of the licences to be covered by the AV companies. Users will also receive a $20 compensation for every session they attend, for a total of $100 per user. With a 200-user study, the total cost should not exceed $20,000.

The final cost for a four-month field study with 200 users based on this model should be around $53,000. This model is less expensive than the one we previously used for our study because we do not need to provide laptops for the users. It also allows researchers to meet the users and have direct access to the computers. On the other hand, it has two significant disadvantages: 1) it is labour-intensive, due to the necessity of meeting the users and inspecting the machines in person, and 2) it introduces an undesirable geographic bias in that all users would have to reside in proximity of the university or testing lab (i.e. same city or region).

Model 2

As the second model should entirely be operated remotely, it presents more technical challenges. At the beginning of the study, users will have to be directed to a secure website where they will be able to download a package. When executed on their computer, the package should remove any anti-malware products previously installed, and then install the AV product to be evaluated, as well as our scripts, software and diagnostic tools. Once installed, our scripts will benchmark the system and send the information to a designated server. On a monthly basis, the user will have to complete an online questionnaire, while our scripts will send the data compiled during the previous month to our server.

Even if this collection process is automated, technical analysis of the data received will be required, in order to determine if the computer is infected or suspected to be infected. If the AV product or our tools indicate that the computer may be infected, the user will be directed to an online additional consent form, in order to allow us to collect supplementary data that will help us to identify the means and the source of the infection. This additional collection of data would be remotely triggered either by the user or by remote access to the user's computer. Since all data will be sent remotely to our designated server, data will have to be encrypted in order to preserve the identity and the security of the users but also the security of our infrastructures.

Initial expenses: As this model presents more technical issues to address, we expect that a minimum four-month development period will be required to develop and test the package, the tools, the scripts, etc. The initial expenses should be around $20,000.

Operating expenses: Technical work will also be required in this model: two hours per user for the monthly analysis of the data, one hour per user for technical support, and one hour per user for administrative work. With a total of 800 hours for 200 users, and a $20 compensation, the total expenses should be around $16,000.

User compensation: Users will be offered a free one-year licence for the AV product installed on their computer. We expect the cost of the licences to be covered by the AV companies. Users will also receive a $50 gift certificate (for online purchases, usable worldwide) if they complete the entire study. With a 200 users, the total cost should not exceed $10,000.

The final expenses for a four-month field study with 200 users based on the second model should be around $46,000. For these specific conditions this model is less expensive than the first design because we do not need to meet the users. While this approach does not give us direct access to the computer being used (and hence is more susceptible to miss rootkits and other types of advanced persistent threats), it has the advantage that users can be located anywhere in the world because they do not need to attend sessions in person.

Figure 5 shows the comparative expenses for each model based on the number of users. While the second model is less costly for a study with 200 users, it is not the case for studies of fewer than 150 users where the first model is cheaper. If we remove the initial expenses that only apply on the first time, the second model is more advantageous. In other words, Method 1 might be quicker (less development time) and cheaper for a proof-of-concept comparative AV field study involving just over a hundred users. On the other hand, the extra one-time investment and longer lead time (due to toolset development) of Method 2 will be more than offset in the long run, if the use of field studies as a means to conduct comparative AV tests becomes widely accepted in the industry and in academic computer security research.

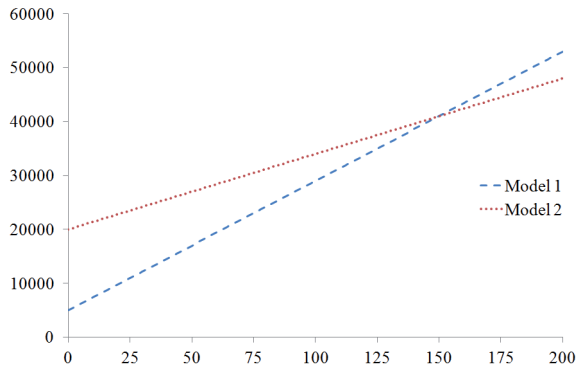

Figure 5: Expenses in dollars as a function of number of study participants.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we describe the first field study of AV performance evaluation conducted with real users in non-laboratory conditions. The study follows a methodology proposed by some of the co-authors [2] in 2009 and is akin to the clinical trial methodology used in medicine.

This first proof-of-concept study shows that this methodology constitutes a viable alternative to conduct traditional, lab-based AV evaluation, with a cost commensurate with that of other types of user studies in computer security and even computer science at large. Furthermore, we have presented two different cost models for developing this methodology into a multi-product, repeatable process that could be competitive with current industrial product comparative evaluation methodologies.

Most importantly, however, this methodology offers performance results that are far more significant and less prone to controversy, due to the realism of the testing environment and the independence of the threat selection process, which is totally and solely dependent on the user's behaviour and daily use of his computer.

Furthermore, and from a scientific point of view, while the results of this limited proof-of-concept study are not statistically significant, and cannot be generalized to the whole population of home computer users, it does offer a glimpse of the kind of analysis that could be performed with a large-scale study. In particular, we were able to conduct preliminary correlation analysis to try to establish whether demographic factors such as age and gender, or behavioural factors such as overall usage and visits to suspicious or risky websites influenced the probability of exposure to threats and potentially infection.

Nonetheless, there are many limitations to and lessons learned from this study. First, the results are not statistically significant; we knew this would be the case from the start, but we wanted first and foremost to evaluate the feasibility of such an approach. Furthermore, as the study progressed, we discovered some flaws in the experimental protocol such as not collecting information about the number and types of external devices connected (e.g. from the registry) and not recording hashes (e.g. MD5 or SHA1) of files as they were being tagged/detected by the various monitoring tools; this last feature would have allowed us to potentially recover the 'missing' sample that was missed by the AV and conduct *a posteriori* performance evaluation. On the user behaviour side, it has been suggested to us [5] that it would have been interesting to classify users according to psychological characteristics such as 'risk averseness' in order to generate a better understanding of the personal causes leading to infection; there are several psychological tests that would have allowed us to evaluate this for each user, had we thought of it and had requested the Ethics Review Board for permission to conduct them.

In summary, we hope to continue our research and conduct larger-scale experiments in the future to address some of these issues and obtain more definite answers to the underlying scientific and industry-motivated questions. We have to that end presented two main models to conduct field studies applied to comparative AV testing. The first one, which involves meeting users in person, is more costly but allows the researchers to have direct access to the computer and can be started right away. This model should be considered as a transition to the second one, which leverages remote access to monitoring data. The second model presents many technical issues and a higher investment cost. However, we believe it presents a more viable and interesting solution in the long term, for both the AV industry and the scientific community.

ACKNOWLEDGEMENTS

The authors wish to thank Ann Fry for her contribution to this project. This project was funded by the Internetworked Systems Security Network (ISSNet), a Strategic Research Network of Canada's National Science and Engineering Research Council (NSERC), MITACS, and *Trend Micro*.

REFERENCES

[1] McAfee Labs. McAfee Threats Report: First Quarter 2012. http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q1-2012.pdf.

[2] Somayaji, A.; Li, Y.; Hajime, I.; Fernandez, J.M.; Ford, R. Evaluating Security Products with Clinical Trials. Proc. USENIX Work. on Cyber Security Experimentation and Test (CSET), 2009.

[3] Lalonde-Lévesque, F.; Davis, C.R.; Fernandez, J.M.; Chiasson, S.; Somayaji, A. Methodology for a Field Study of Anti-Malware Software. Proc. Workshop on Usabe Security (USEC), 2012.

[4] PandaLabs. Quaterly Report January – March 2012. http://press.pandasecurity.com/wp-content/uploads/2012/05/Quarterly-Report-PandaLabs-January-March-2012.pdf.

[5] Clementi, A. AV Comparatives. Private communication, July 2012.