We Need to Talk About AntiViruses: Challenges & Pitfalls of AV Evaluations

Marcus Botacin[1], Fabricio Ceschin[1], Paulo de Geus[2], André Grégio[1]

[1]Federal University of Paraná (UFPR-BR)
{mfbotacin,fjoceschin,gregio}@inf.ufpr.br
[2]University of Campinas (UNICAMP-BR)
paulo@lasca.ic.unicamp.br

Abstract

Security evaluation is an essential task to identify the level of protection accomplished in running systems or to aid in choosing better solutions for each specific scenario. Although antiviruses (AVs) are one of the main defensive solutions for most end-users and corporations, AV's evaluations are conducted by few organizations and often limited to compare detection rates. Moreover, other important factors of AVs' operating mode (e.g., response time and detection regression) are usually underestimated. Ignoring such factors create an "understanding gap" on the effectiveness of AVs in actual scenarios, which we aim to bridge by presenting a broader characterization of current AVs' modes of operation. In our characterization, we consider distinct file types, operating systems, datasets, and time frames. To do so, we daily collected samples from two distinct, representative malware sources and submitted them to the VirusTotal (VT) service for 30 consecutive days. In total, we considered 28,875 unique malware samples. For each day, we retrieved the submitted samples' detection rates and assigned labels, resulting in more than 1M distinct VT submissions overall. Our experimental results show that: (i) phishing contexts are a challenge for all AVs, turning malicious Web pages detectors less effective than malicious files detectors; (ii) generic procedures are insufficient to ensure broad detection coverage, incurring in lower detection rates for particular datasets (e.g., country-specific) than for those with world-wide collected samples; (iii) detection rates are unstable since all AVs presented detection regression effects after scans in different time frames using the same dataset and (iv) AVs' long response times in delivering new signatures/heuristics create a significant attack opportunity window within the first 30 days after we first identified a malicious binary. To address the effects of our findings, we propose six new metrics to evaluate the multiple aspects that impact the effectiveness of AVs. With them, we hope to assess corporate (and domestic) users to better evaluate the solutions that fit their needs more adequately.

Keywords: AntiVirus Malware Detection Remediation & Attack Opportunity.

1 Introduction

Malicious programs and Web pages are prevalent threats to interconnected systems. Successful attacks involving malware or compromised pages may result in financial losses or damage to the image of Internet users. Thus, combating them requires that individuals and corporations adopt defensive solutions to protect their systems. One of the most deployed defensive solution overall is the antivirus (AV), that have become popular to the point of being a mandatory requirement for corporations obtaining the PCI-DSS security certification [32]. Therefore, evaluating AV's effectiveness and efficiency is essential to allow both system administrators and users to select the best solution for their needs. However, AV's evaluation might not be straightforward.

Current market-oriented AV evaluations adopt an "one-size-fits-all" approach. None of the most popular tests [3, 1] provide results broken down by threat categories. Instead, they provide generic results without considering multiple infection scenarios, such as the specifics of the target user country/relationship with Internet-connected systems, and ignore important features regarding AVs' way of operation. On the one hand, AV's threat detection rate is a widespread metric adopted by most AV evaluations. On the other hand, AV evaluations often neglect the time that an AV solution takes to react to a new threat discovery (AV's response time) and/or AVs stopping detecting a sample after some time (detection regression). Moreover, most evaluations cover uniform scenarios, such as considering single platforms or worldwide datasets as generalization of specific countries and contexts. With a limited view of AV's operation, users and corporations might be biased to choose their security solutions in a way they are not fully security-covered due to the lack of information about AVs particularities. Therefore, users that choose their AVs based on their best results for the general scenario may be less protected in their real-life system's use than if they have chosen an AV more focused in handling the particular threats of those users' scenarios. In addition, AV evaluation results are either diluted along academic research (other goals than users') [51], or not updated even after a decade [35], a period in which AVs have undergone through many changes in their detection engines (see Section 2).

To bridge this understanding gap about how AVs behave in actual scenarios, we conducted a longitudinal evaluation of their behavior, i.e., how AV's detection changes over time when considering the same dataset. We collected daily samples from two representative malware sources: a popular collection of worldwide malware and a regionalized malware collection provided by a Brazilian CSIRT. This allows us to isolate the effect of dataset in the overall AV's behaviors. We repeatedly submitted the collected samples to VirusTotal (VT) AV scans for a period of consecutive 30 days, which allowed us to identify any detection result change, such as in AV's detection rates and labels. As far as we know, we are the first to perform a longitudinal analysis of AVs at a daily-basis granularity. Our experiments considered distinct file formats (binaries and Web pages), platforms (Windows, Linux, and Android), regionalized datasets (BR and World samples), and periods (within an entire year), thus evaluating AVs in their multiple aspects. In total, we considered 28,875 unique malware samples. During the whole observation period, we performed more than one million distinct VT submissions.

Our experimental results show that: (i) understanding phishing contexts is a challenge for AVs, thus malicious Web pages detectors are less effective than their binary counterparts; (ii) detection procedures derived from the generalization of global data are not enough to ensure broad detection coverage, thus particular datasets (e.g., Brazilian malware) are less detected than world-wide malware; (iii) detection rates are not constant, and all AV products presented detection regression effects when periodically scanning the same malware samples dataset; and (iv) AV's long response times to deliver new signatures and heuristics create a significant attack opportunity window within the first 30 days a binary sample was first discovered by us, updating results from previous research work [35].

We propose six new evaluation metrics regarding threat detection and elimination to be considered during AV solutions selection to better account the aforementioned AV's operation drawbacks, which includes the measurement of response time and regression occurrence. We present an exploratory analysis of these metrics applied to end-user and corporate scenarios to highlight how the selected AV solution changes according the defined scenario needs. On the one hand, corporate users weight more AV's response time when selecting an AV because corporate users are likely more affected by zero-days than end-users. On the other hand, end-users weight more AV's detection regression when selecting an AV because end-users are likely more affected by long-term malware campaigns than corporate users.

In summary, our research work's contributions are threefold:

- A longitudinal evaluation of AVs considering their operation in actual scenarios, and highlighting their weaknesses and strong aspects.
- Definition of six new evaluation metrics to characterize AVs in their multiple dimensions (of use and deployment);
- Validation of the proposed metrics, showing how they can be leveraged to identify the best AV for distinct scenarios and users' requirements.

This paper is organized as follows: in Section 2, we present background information on AV operation; in Section 3, we present our methodological approach and the evaluated malware samples; in Section 4, we present evaluation results that characterizes current AV solutions operation; in Section 5, we present our proposed metrics, their interpretation and discusses the best metrics for distinct scenarios; in Section 6, we discuss the impact of our findings and proposals; in Section 7, we present related work to better position our work; finally, we draw our conclusions in Section 8.

2 Background

We propose to evaluate AVs according to their capacity of both detecting and labeling malicious artifacts (e.g., binary files, scripts, URLs, and/or web-pages). However, these capabilities are strongly tied to the way the AV is designed and implemented. Therefore, to better position our results, we try to shed some light on the AV engine's internal working mechanisms.

Historically, AV engines started detecting threats performing pattern matching using signatures, which are sequences of bytes known to belong to malicious samples [26]. In response to AVs measures, attackers started deploying malware variants, samples generated from the same source but presenting distinct byte sequences. This competition caused an arms-race between attackers and defenders since the 90's [34] and still observed in current AV's implications.

Since AVs could no longer keep up with the fast pace required for signature generation on a per-file basis, AVs started to "guess" and label some files as probably malicious through the use of heuristics [40]. A typical heuristic is to flag binaries as malicious when any obfuscation signs are found. For instance, benign files packed with crypters–pieces of code which protect their payloads by encrypting themselves at compilation time and decrypting at runtime–are often detected as malicious given their frequent use also in malware samples distribution [44].

As time went by, binaries became so complex that even heuristic approaches have not been enough to flag malware without leading to false positives [7] (FPs). An AV that detects benign software as malicious becomes impractical since it prevents users from using the applications that the AV was supposed to protect. Therefore, more powerful detection solutions were required to detect complex threats without causing FPs. As such, AV engines started to rely on Machine Learning (ML) and/or on Artificial Intelligence (AI) for their classification and decision procedures [6]. ML/AI may be used, for instance, to flag samples as malicious based on the usage frequency of some assembly instructions [25].

After that, AVs have been implementing a combination of all aforementioned techniques in their detection engines, thus their detection rates and labels are biased by all these factors at the same time. In practice, the labels assigned to the samples may vary according to the internal engine that a solution leverages for detecting them: (i) samples detected by known signatures may present detailed label information (e.g., `W32/Sample-Name`); (ii) samples detected through heuristic approaches may present either the heuristic name (e.g., `W32/Packed`) or a "`generic`" label; and (iii) samples detected via ML approaches might only present detection rates (e.g., `malicious confidence: 90%`), without additional information.

On the one hand, such heterogeneity complicates homogeneously evaluating AV detection. Therefore, this work proposes metrics to highlight specific AV's operational characteristics to allow more fine-grained evaluations. On the other hand, as such heterogeneity appears in practice, we cannot overlook it in evaluation procedures. Hence, we present an AV landscape considering AV's outputs regardless of the internal operation of their engines.

In addition to multiple detection mechanisms implementation, AVs also update them frequently to keep up with malware evolution. Thus, new signatures should be released for matching newly created samples, new heuristics for detecting malware variants and classifier's definition updates due to concept drift, a natural phenomenon in dynamic and non-stationary environments where characteristics and distribution of data change as time goes by [12]. Therefore, in this paper, we present a continuous evaluation that encompasses AV's update procedures rather than a static view of AV solutions operations.

3 Methodology & Dataset

Design of Experiments. Our experimental approach consisted in submitting all collected malicious artifacts (executable binaries and malicious web pages) to the VirusTotal (VT) service [49] via `Python` bindings for VT's public API [48] and retrieving detection rates and labels for all AV solutions. All retrieved data was stored in a `SQLite` database which was further queried for data discrepancies identification and metrics calculation.

The samples which were reported as first-seen in the VirusTotal service were daily re-submitted for consecutive 30 days. In each re-submission, a new scan, with updated malware definitions, was forced, thus allowing us to track how AV solutions detection evolved (temporal analysis). We also performed non-temporal analysis about time-independent aspects of AV detection, such as sample's labels meaningfulness.

We are aware that comparing AVs using VirusTotal has significant drawbacks [47], mainly because their running AV's version might differ from the ones locally installed on customer's machines. However, using VT is the only way to scale analysis to million submissions as presented in this work. Also, in a significant part of the paper, we are not looking at individual AV solutions, but trying to characterize the behavior of a hypothetical "average" AV solution that ignores AV's specific features. Therefore, we considered this trade-off as acceptable. To mitigate the uncertainty regarding the validity of our findings in the real-world, we confirmed our results by locally running some of the AVs. The confirmation results can be found on Appendix A.

Datasets. We considered four distinct malicious artifacts sources for the experiments proposed in this work: (i) a private repository of country-widespread, specific malicious objects collected by a Brazilian CSIRT's abuse e-mail and sensors network; (ii) the Malshare [30] repository of daily-collected, worldwide malicious objects; (iii) the VirusShare [46] repository as the source of Linux malware samples; and (iv) the VirusTotal service as the source of Android malware samples.

The first two sources provide us with malicious Windows binaries and web pages daily. The continuous malware collection allows us to perform a time-evolution comparison (temporal analysis) of AV's ability to detect the samples present in these datasets. The last two sources provide Linux and Android malware samples without precise timing information. Therefore, we leveraged their samples to enrich our non-temporal AV evaluation dataset, so that we can compare the results of AV operating on distinct platforms and environments.

We continuously captured samples from August/2017 to December/2018. In total, we considered 5,614 worldwide-crawled PE binaries, 3,302 Brazilian-collected PE binaries, 5,929 worldwide-crawled web pages, 4,030 Brazilian-collected pages, 5,000 ELF binaries, and 5,000 Android applications. During the whole observation period, we performed more than 1M distinct VT submissions. Table 1 summarizes the number of samples, malware families, and artifact types in each dataset. Family labels were normalized using AVClass [41].

Table 1: **Dataset Summary.** Malware families labels were normalized using AVClass.

Dataset	Samples	Families	Formats
Brazil PE	5614	23	21
World PE	3302	16	7
Linux	5000	47	6
Android	5000	52	N/A
Brazil Web	4030	N/A	N/A
World Web	5929	N/A	N/A

Most of the experiments focus on the Brazilian and World datasets of PE malware. They trigger the most mature AV's detection engine to be evaluated since AVs have been analyzing this type of file for a while. We selected the Brazilian dataset for this study since it represents the real threat's distribution that a significant part of the Brazilian population faces daily. Therefore, we can better understand the real impact of AV's drawbacks on user's lives. This dataset has been already described in other studies [12] and demonstrated to challenge other malware detection techniques [5]. All samples in this dataset are considered as malicious as they were collected and labeled by the CSIRT team. Most of the samples in this dataset were first submitted to VirusTotal by us, thus indicating a significant level of novelty. The World dataset, in turn, was selected for this study because it does not present the bias of the Brazilian dataset. Therefore, we can attribute any effect observed in both datasets to the AV's drawbacks and not to dataset's characteristics. We have not observed samples intersection between these two datasets.

4 AV Evaluation

We have identified the most common pitfalls in AV evaluations, which are shown in the next subsections. We also present AV detection results to support our discussion on these pitfalls. Although some of them might have been individually pinpointed in previous work, we are not aware of articles/documents discussing all of them together along with updated data about AV detection. We consider this discussion essential since there is a non-negligible research corpus that relies on AVs evaluation/detection rates.

4.1 AVs evaluation results cannot be uniform

AV evaluations often consider the detection rate as the only criteria for assessing effectiveness, thus neglecting other important AVs' operation aspects. In addition, these evaluations often report very high detection rates as their main result, which seems incompatible with user's risk perception in practice [37, 21].

The observed discrepancy is caused by the difference between the characteristics of the datasets used in the evaluations and the scenarios faced by real users in their daily routine. Most evaluations consider completely balanced datasets in regard to malware family distribution (e.g., same number of Trojans, virus, worms, and so on), and only well-known file formats (e.g., they keep standard binaries and discard executable scripts). In practice, however, users are targeted by threats in an unbalanced way, according to the operational context they are part of. For instance, the selection of the best AV in an evaluation that considers a balanced dataset might bias the detection results, since poor detection rates for a given malware family may be masked within the overall detection rate. Therefore, we advocate that AV evaluation reports should break down results according to the multiple AV's aspects (e.g., by families of samples and/or file format detection). Hence, users will be able to evaluate the best AV according to the characteristics of the scenario in which the AV is aimed to operate.

To show the impact of performing this breakdown, we compared the difference of presenting detection results for the samples in our datasets in both ways (consolidated and separated by categories). In Figure 1, we show the consolidated AVs' results for the average detection of standard (non-scripted) Windows, Linux, and Android malware binaries. We discarded scripts and other file formats from these experiments as they present their own drawbacks, as further discussed. Therefore, this experimental variable isolation allows us to spot AV operation in their most favorable conditions.

All datasets presented high detection rates, as in most current AV evaluations. More interesting, this result holds true for all platforms/environments. This happens because we balanced the datasets (using AVClass [41]) in a way that they present the same number of samples of all malware families, and we considered only standard binaries.

To understand the impact of breaking down AV evaluations, let's consider the average AV detection rate for most popular Windows malware families common to the two datasets, shown in Figure 2. The detection rates are not uniform: Trojans have been significantly more detected than bankers, for example. Considering these results, we highlight that while the consolidated evaluation would be able to suggest the best AV for users of a scenario mostly targeted by Trojans, this approach would completely bias AV selection for users in scenarios mostly targeted by banking malware.

In addition to malware family distribution, file formats also affect AV detection rates. This happens due to the fact that not all AV solutions parse the same file formats and, at the same time, their focus is on standard binary formats, such as Windows PE. To demonstrate the impact of including multiple file formats

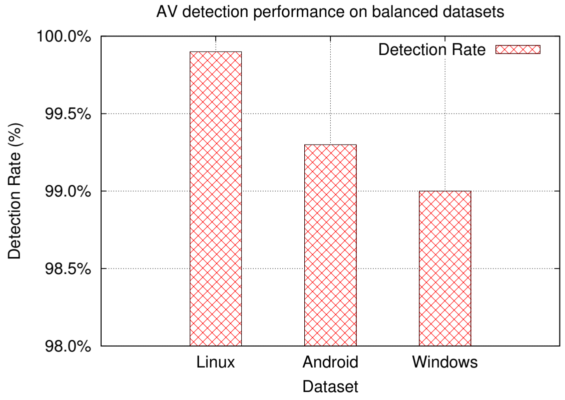


Figure 1: **Consolidated AV results.** Dataset balancing bias the overall detection rate.

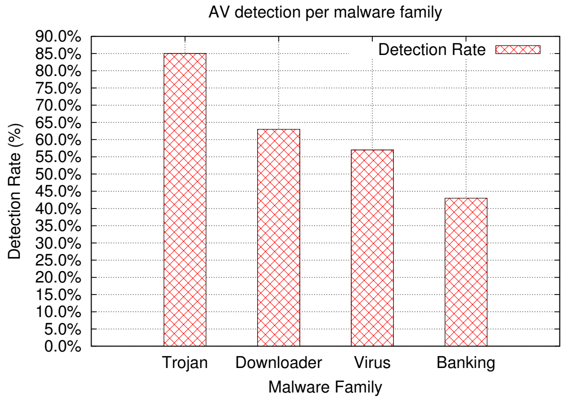


Figure 2: **Detection breakdown by malware family.** Some families are more detected than others in average.

on AV evaluations, let's consider the breakdown presented in Figure 3. It includes all MS-Windows platform-supported file formats, even the ones that were not considered in the previous experiments. AVs are more prone to detect the most popular executable formats (e.g., COM and EXE) than scripted and interpreted formats (e.g., VBE and JARs). Therefore, if an evaluation clearly presents its results separated by file formats, it would allow users to identify AVs unable to detect threats in specific formats, as well as to choose the best solution for targeted scenarios.

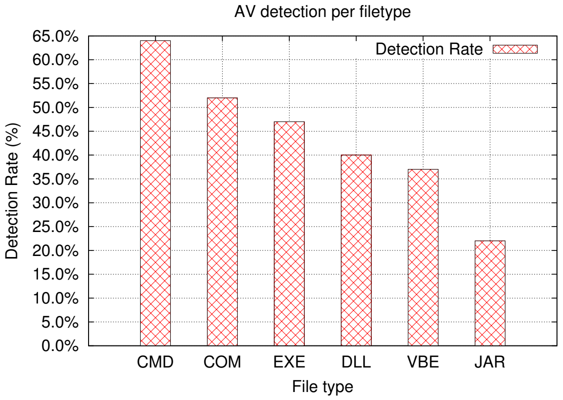


Figure 3: **Detection breakdown by file format.** Although standard binaries are reasonably detected, scripted and interpreted threats pose detection challenges for current AVs.

The aforementioned results highlight the need of considering the operational scenario in AVs evaluation. In Figure 4, we illustrate this finding in practice by comparing average detection rates for two distinct datasets: (i) world samples collected from malshare, which contains 65% of Trojans, mostly distributed as standard PE files; and (ii) Brazilian samples collected from a CSIRT that attends the entire country, composed by 75% of banking malware, distributed in diverse file formats. The overall detection rate for the Brazilian scenario is biased by the low AV performance on detecting

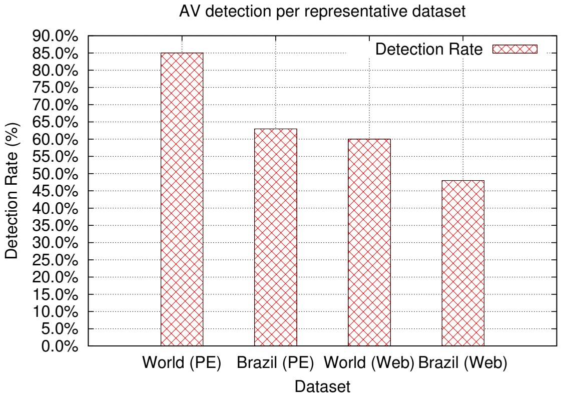


Figure 4: **Detection rates per representative datasets.** The Brazilian dataset is less detected than the World dataset due to the high number of banking malware. Web pages are less detected than Windows executable files.

banking malware and diversified file formats, thus reinforcing the need for considering particular scenarios when conducting AV evaluations.

4.2 AVs respond differently to different types of threats

AVs present different detection rates for distinct threat types in addition to presenting different detection rates for different malware families and file format (as shown in the previous subsection). Figure 4 shows that AVs are less effective in detecting malicious pages than detecting binaries, which holds true for both World and Brazilian dataset.

The detection rate difference in both threat types is explained by the distinct risks that they pose to the system. On the one hand, binaries are focused on directly causing harm to the victim's systems. On the other hand, malicious web pages are mostly focused on indirectly deceiving users into clicking into a malicious link, either for advertisement or for then download a malicious payload.

These distinct operation modes require that AVs deploy distinct strategies for the detection of these threat types. Most system binaries are insensitive to the infection context and detectable through static/dynamic analysis procedures (banking malware are a noticeable exception to this rule [18]). Unlike them, malicious Web pages are mostly not: they are usually sensitive to the infection context, mainly due to phishing Web pages [43], and require that AVs understand their context to recognize their maliciousness. Considering the results presented in Figure 4, AVs are still not able to fully handle this type of threat due to this huge context understanding challenge.

4.3 AVs have a response time

AV detection rates can also vary due to other factors than family balancing, file formats, and threat types. The most significant factor affecting AV detection is the time that has passed since the release of a new sample, its identification, followed by its detection by the AVs after malicious definitions updates.

To evaluate the impact of time on AV detection results, we selected the samples first reported by us to the VirusTotal (VT) service (i.e., samples reported for the first time in VT's database after our submission, according to VT's API queries) and repeatedly submitted them to scans by a period of consecutive 30 days. Figure 5 shows the AV detection rates for multiple datasets in two distinct periods: (i) the first day in which the samples were submitted; and (ii) in the last day when the same samples were submitted to the same AVs, when these were already updated with new malware definitions. We notice that detection rates can vary up to 10% from the initial submission to the final detection in the last day. Such detection rate variation has been observed in all datasets. Therefore, we advocate that the response time metric should be considered by AV evaluations.

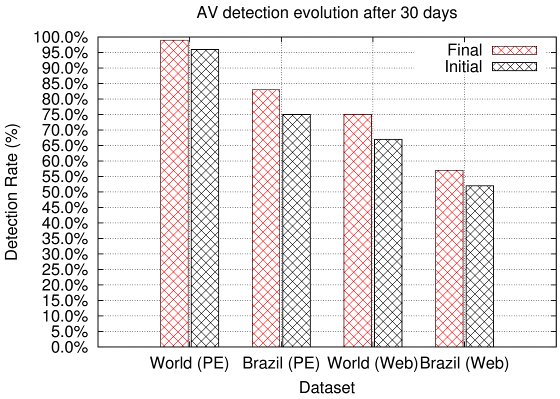


Figure 5: **Time effect over AV detection rates.** Detection rates can vary up to 10% according to the observation period.

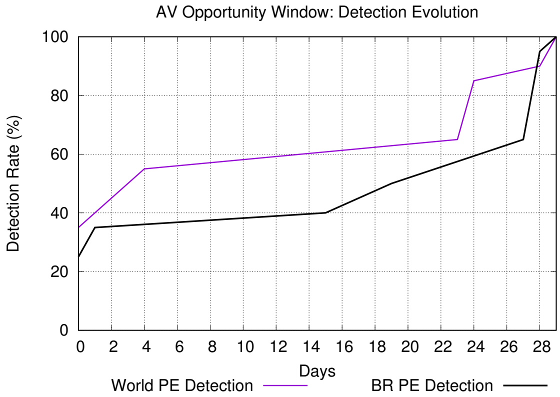


Figure 6: **AV detection evolution.** The long response time create a significant attack opportunity window.

Apart from being a pitfall on evaluation, the time that an AV takes to react to a new threat also directly affects AV's detection effectiveness. AVs taking a long time to react create an attack opportunity window in the meantime, i.e. a period in which users are vulnerable to the new malware sample as the AV has not yet updated malware definitions to detect it. To evaluate how long AVs take to react to new threats, we selected the subset of samples detected in the 30^{th} day and evaluated how their detection by AV solutions evolved over this time period.

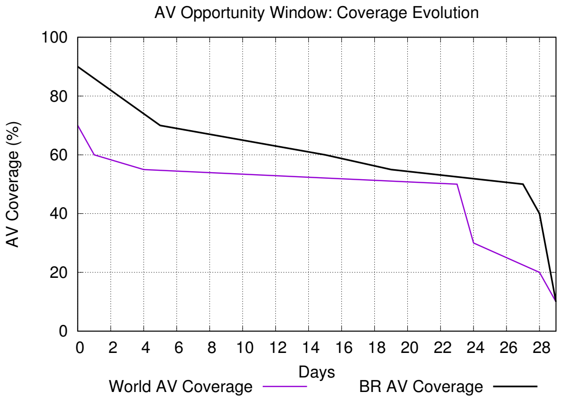


Figure 7: **AV coverage evolution.** Not all AVs are able to keep up with the same detection rates as the times goes by.

In Figure 6, we show the fraction of samples detected by at least one AV solution at a given day (`Detection` curves) and the fraction of AV solutions which agree on detecting all the detected samples at a given day (`Coverage` curves). Less than 50% of samples are detected at day 0–when they were collected and first submitted–on both scenarios, which indicates users are vulnerable to newly created threats even when using an AV solution. Ideally, the attack opportunity window should be as short as possible to reduce user's exposition. In this sense, the hypothesized full protection (100% detection) was achieved only after 29 days on both World and Brazilian scenarios, which is a significant opportunity window for attackers. In fact, whereas AVs quickly detect a fraction of the samples, they slowly increase their detection coverage. This either indicates (i) the existence of a class of samples which is harder to detect, or (ii) the insufficient scalability of existing detection mechanisms to cover the whole context of the threat. We can observe the occurrence of such effect in the World PE detection curve: 55% of the samples were detected within the first 4 days, but solutions took 19 days (until the 23rd day) to detect an additional ≈10% of the threats (up to 65%).

The comparison of scenarios indicates that the World scenario responds faster than the BR one. This may be explained by the particularities exhibited by the regionalized scenario. Conversely, the time taken to detect all samples is similar in both scenarios, suggesting that this detection evolution is more related to the need of analyst's intervention to detect new threats than to dataset's specific characteristics.

The attack opportunity window is eliminated in the 29th day when considering all AV solutions. However, some users have been still unprotected in the end of the period because not all solutions detected all threats. Figure 7 shows the AV's Coverage for the evaluated samples, i.e., the fraction of AVs that detected the number of samples previously shown in Figure 6. In the first days, the majority of AVs agree on detecting the same few samples: 70% and 90% for World and Brazilian datasets, respectively. As time goes by, each AV solution detects a distinct set of samples. Only 10% of all solutions agreed on detecting all samples in both scenarios in spite of their contextual differences.

The break-even point between detection and coverage, i.e. when both curves intercept each other, is around 55% for both World and Brazilian scenarios. However, whereas the break-even point is achieved in only 4 days for the World scenario, it takes 21 days to occur in the Brazilian scenario. This difference shows the average protection offered by AV solutions in a general manner while as-yet not fully updated to cover the newly launched threats. In practice, the low correlation between different AV detection rates has already been pointed as an actual problem in many scenarios, such as in the Android platform [31].

4.4 AVs are not good at labeling samples

AVs ideally should also enable users to take the proper countermeasures to mitigate the effects of malware infection in addition to detecting malware samples. Thus, the proper labeling of samples is a very important step to allow users to respond to distinct threat infections (according to malware specific aspects). For instance, the infection by downloader malware samples require users to check computer's filesystem for stored malicious artifacts. In turn, banking malware infections require users to get in contact with financial entities to notify the incident. Besides that, some machine learning models are based on ensembles, in which each classifier is trained using different malware families [24]. Therefore, for these solutions to work right, it is important to label a sample in the right family to keep each model updated according to the family sample's changes.

In the context of this work, we consider AV labeling capabilities as an essential feature for AV solutions as it can be used as a proxy for measuring AV's understanding of the detected samples. In other words, we consider that the more qualified the assigned the labels are, the better the AV is able to recognize the malicious context regarding that given threat. In practice, however, some AV vendors might claim that a good labeling capability is considered only a desired but not mandatory AV feature since AV's primary goal is to detect the malware samples.

AV labels should be standardized by CARO, which defines that *"the full name of a virus consists of up to four parts, delimited by points ('.'). Any part may be missing, but at least one must be present"* [10]. The expected parts are respectively the following: (i) malware family name; (ii) malware group name; (iii) major variant name; (iv) minor variant name. Additionally, the label might present optional modifiers (extensions) representing any vendor-specific information (e.g., "packed with UPX"). The presence of label extensions usually means that the AV has deep knowledge about the identified threat.

To check AVs' ability on labeling samples, we considered the labels assigned to the samples belonging to the following datasets: (i) samples detected in the last day of the observation period, such that we have at least one assigned label per sample to evaluate; and (ii) samples belonging to the long-term Linux and Android datasets, such that we have payload diversity to evaluate labels in a broader manner. For this experiment, we considered only self-contained executable files because the AV solutions available in the VT service often do not label web pages and scripts[1]. In Figure 8, we show our evaluation results regarding AV's label assignment compliance to the CARO guidelines.

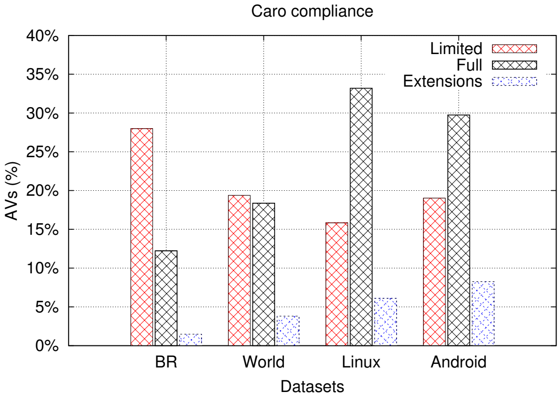


Figure 8: **CARO compliance.** Most samples comply with the minimal standard, but their labels are not informative enough.

We discovered that most samples comply with the CARO standard (sum of all bars labels from Figure 8). However, for 20% of the cases, this is achieved in a minimal way (Limited bar label), presenting only the minimally required amount of information (a single part label). Full information (Full bar label) is not available for the majority of cases, thus important sample's characteristics such as variants and groups are often unknown. The intermediate level of information provided by most labels is compatible with the use of heuristics, which, in the end, are unable to provide full information (see Section 2).

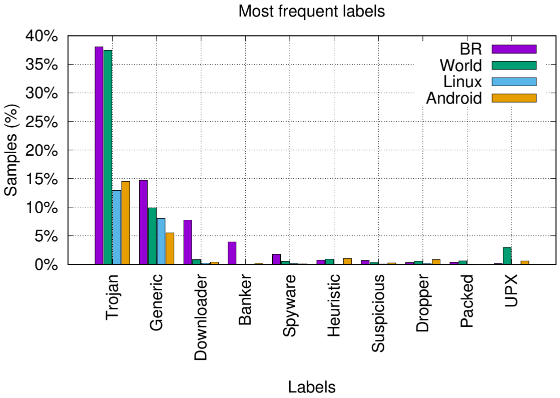


Figure 9: **Label quality.** Heuristic labels, such as `generic`, do not allow users to take the proper countermeasures in case of infection.

AVs providing additional information (Extension bar labels) are even rarer (less than 10% of all cases). Thus, sample's characteristics such as packing and context information are hardly ever provided. This parsimonious number of extended CARO labels is explained by the significant effort required from AV's analysts to study the samples in detail. This manual task is only performed on a small number of samples according to AV vendor's demands. These results indicate that AV companies need to enhance their labeling procedures in an overall way, thus providing stronger support for incident response procedures. Face to the costs of allocating more human resources to perform manual analyses, the development of more in-

[1] See an example at `https://tinyurl.com/yxexo7za`

formative automated procedures should be prioritized.

AVs should be able to provide some meaningful label information to enable incident response even when not providing full label information. To evaluate whether AVs are able to provide such information in practice or not, we checked all labels assigned to the samples in our datasets. In Figure 9, we show the top assigned labels.

On the one hand, we notice that the majority of samples are labeled as `Trojan` in all datasets. This is compatible with the popular infection mechanism used by malware authors of deceiving users into installing modified, malicious versions of legitimate applications through phishing and/or fake advertisements. On the other hand, these most assigned labels, such as `generic` and/or `suspicious` types, do not allow users to take proper countermeasures. This phenomenon derive from the use of heuristic (`heur`) approaches, such as detecting the packer instead of the sample's payload itself. It explains the samples labeled as `Packed` and `UPX`, a packer name that does not provide enough information about the sample content.

4.5 AVs often stop detecting samples

The distinct strategies adopted by the AVs and their response time cause a significant variation in the number of detected samples over time, in addition to the detection opportunity window and label issues. Signature addition/removal and/or heuristic changes over time cause extra samples to start being detected, but unfortunately, some other samples stop being detected simultaneously. We evaluated detection regression–when a sample stops being detected–by observing the detection rate for the subset of samples which were reported to be detected in the last day of the observation period, as shown in Figure 10. Notice that in this experiment we discarded the samples that were not detected since by definition there is no regression effect for them.

We observe that the detection rate decreases several times during the study period. This effect causes, for instance, `World` users to suddenly become vulnerable to 4% of threats in a day (from day 11 to 12). We highlight that this behavior is not related to samples locality, because `Brazilian` and `World` curves presented similar characteristics, decreasing and growing mostly at the same time, which indicates that the same cause might be at play, such as AV relying on heuristic detectors.

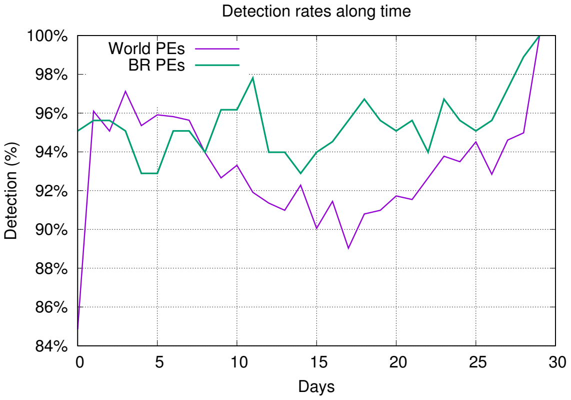


Figure 10: **Overall regression effect for `World` and `Brazilian` PEs.** Some samples belonging to the dataset stopped being detected during the evaluation period such that the overall detection rate decreased in some days before AVs achieving the final detection rate in the end of the observation period.

The behavior shown in Figure 10 represents the overall effect, which means that the detection rate grows for some samples and decreases for others. We also evaluated the regression effect for individual samples, as shown in Figure 11. The `Regression` bar label refers to the percentage of samples that had their detection rates decreased at least once by at least one AV solution. The `Restoration` bar label

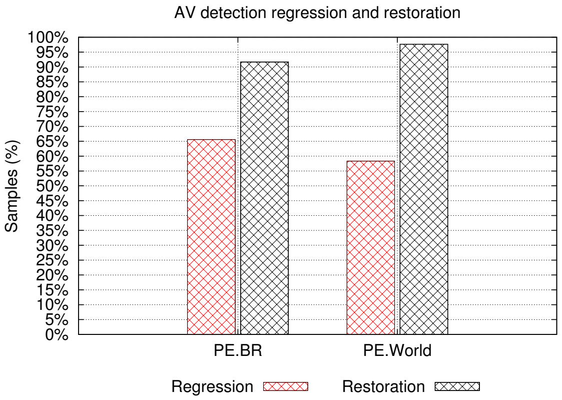


Figure 11: **Regression effect for individual `World` and `Brazilian` samples.** Most samples presented detection regression during at least one day during the observation period. Most of the samples that presented detection regression recovered from this effect, presenting a higher detection rate in the last day than the detection rate presented in all previous observation days.

refers to the percentage of samples that suffered `Regression`, but recovered their detection rates – i.e. their detection rate on the 30^{th} day is equal or higher than the detection rate in any other previous day.

For both scenarios, regression occurs at least once in more than 50% of the samples, which may be associated with the use of aggressive heuristic approaches by some AVs. The final detection rate had been recovered in more than 90% of the cases, i.e., it returned to the original or higher detection rate value.

Regression also affects the assigned labels in addition to the detection rate. The assigned labels change according to the method leveraged for sample detection in each period of time. To evaluate the impact of label regression, we considered the labels assigned to the samples during the 30 days period. We identified that 53% of all considered samples changed their label at least once. Moreover, all AV solutions presented label regression for at least one sample. On average, regression affected each sample in four distinct AVs.

In Table 2, we present representative examples of label changes. Some label changes (e.g., line 4 and 5 of the table) may be considered positive (✓) consequence of AV's updates, since they provide users with more informative descriptions of the detected threats. Other label changes (e.g., lines 3 and 6 of the table) represented information loss, since the original labels were replaced by less descriptive versions. Similarly, labels derived from machine-learning detectors (e.g., lines 1 and 2 of the table) might present a regression effect according to the classifier's accuracy in each time period. Therefore, AV evaluations should be performed considering temporal variations and not considering data of a single day that might not reflect the final decision of the evaluated engine.

5 Metrics & Scenarios

We used all the knowledge gathered on the previously discussed AVs drawbacks to propose new evaluation metrics for AVs. The main novelty of these metrics is that they consider the multiple aspects regarding AVs' way of operation. We also show how these metrics can be weighted according to the needs of distinct scenarios (e.g., domestic and corporate users) to allow AV selection in a more fine-grained way.

5.1 Proposed Metrics

We introduce below our proposed evaluation metrics, as well as the way to interpret them. We propose these metrics because they evaluate the impact of the AV drawbacks presented in the previous sections. We consider that these are significant drawbacks of AVs and that these drawbacks are often overlooked in most AV evaluations. The proposed metrics are the following:

Table 2: **Label Regression.** Whereas in some cases labels become more informative over time, in some cases labels regress to generic.

AV	Day	Label	Day	Label	Enhancement
A	1	'malicious_confidence_100%	2	malicious_confidence_80%	✗
A	12	malicious_confidence_60%	13	malicious_confidence_90%	✓
B	3	Trojan-Banker.Win32.BestaFera.amju	4	HEUR:Trojan.Win32.Generic	✗
B	19	UDS:DangerousObject.Multi.Generic	20	Trojan-Downloader.Win32.Banload.aasyh	✓
C	4	Win32:Malware-gen	5	Win32:Dropper	✓
C	16	FileRepMalware	17	Win32:Malware-gen	✗

- **Attack Opportunity Window (AOW)**: With this metric, we evaluate how much time AV solutions take to generate signatures for new threats. This metric enables us to quantify how exposed a user is even when using an AV software (during the initial detection hiatus).

- **Detection Regression (DRE)**: With this metric, we are able to identify when previously detected threats stop being detected by an AV product. It allows us to evaluate whether users become or not exposed to the same threat after it had been first reported by the AV vendor.

- **Final Detection Rate (FDR)**: With this metric, we calculate the overall detection rate of newly captured samples at the end of the 30-day period. This metric allows us to evaluate user's protection in the long term.

- **Initial Detection Rate (IDR)**: With this metric, we calculate AV's detection rates at day zero, i.e., in the first submission after the sample's collection. This metric allows us to evaluate how users are protected by AV solutions regarding newly reported samples.

- **Label Meaningfulness (LME)**: With this metric, we evaluate how useful labels are regarding taken countermeasures. This metric is important because generic detection labels do not expedite cleanup.

- **Label Regression (LRE)**: With this metric, we evaluate how labels change over time. Such information is relevant, since label changes may require modified countermeasures.

5.2 Evaluating Scenarios

Based on how the proposed metrics may impact in an AV choice, we present an exploratory analysis of how the proposed metrics may impact AV selection procedures when leveraged for evaluating scenarios presenting distinct security needs. To do so, we considered the metrics that distinct user groups would value most. Notice that this does not mean that these are the only important metrics or that all users of that group would consider for their protection. Instead, we encourage the reader to reason about which are the best metrics for their scenario. In our exploratory analysis, we considered three distinct users groups and hypothesized their needs as follows:

1. **Domestic Users**, which are more likely targeted by the same well-known samples over time, thus being affected by AV's final detection rates (FDR) and regression effects (DRE). These are important metrics for domestic users since they do not want their AVs to stop detecting a known sample.

2. **Corporate Users**, which are usually targeted by 0-days, thus being affected by

AV's initial detection rates (IDR) and interested in a small attack opportunity window (AOW).

3. **Incident Response Teams**, which are more interested in (i) performing infection cleanups, thus requiring good AV labeling capability (LME), and (ii) avoiding label regression (LRE) to allow a targeted incident response. We highlight that CSIRT reliance on AV labels has been reported in many real cases [36, 17], although these teams might also adopt additional code inspection approaches [20] (e.g., sandbox execution).

To show that distinct metrics should be used for each scenario instead of a universal criteria, we selected the best AVs to fulfill the requirements of the three aforementioned usage profiles. For the sake of simplicity, we present data regarding only the three AV solutions with the highest detection rates for the samples in our dataset. We also limited our evaluation to the subset of all samples which were effectively detected by all the top 3 AV solutions at least once during the observation period, thus discarding overall detection rate as a significant metric. For metric computation, we assigned values to each AV criteria ranging from 0 to 10, where 10 means 100% detection and no opportunity window and 0 means 0% detection and a 30-day opportunity window.

Figure 12 shows the overall comparison among the three considered AVs, thus allowing us to identify which AV outperformed the other in which criteria. We observe in Figure 12a that the AV1 is the best for assigning labels to samples, which turns it into a well-suited solution for CSIRTs. We observe in Figure 12b that the AV2 may not be as good as AV1 for sample labeling, but it detected malicious samples first (a desirable feature for corporate environments). We observe in Figure 12c that the AV3 also does not perform well on samples labeling, but it is the one that presents fewer detection regression occurrences, which turns it into the most suited for domestic users. In summary, apart from the fact that all AVs were able to detect all samples in some period of time, we discovered that each one is the best for each specific scenario. Thus, we highlight the importance of evaluating AVs using more user-targeted metrics.

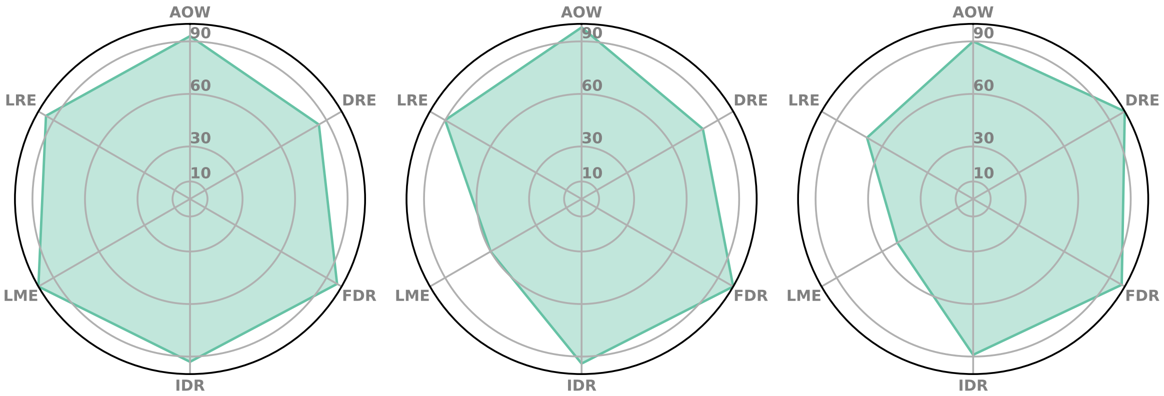
6 Discussion

In this section, we revisit our findings to discuss their implications, contributions, and limitations.

Recommendations for AV evaluations. We expect that our findings could be seen as feedback information to enhance AV evaluation procedures. More specifically, we advocate that:

- **AV evaluation results should be broken down.** AVs present different detection results according to the considered malware family and the considered file format. Therefore, AV detection results should not be presented as an average of all results, since it would mask the AV limitation on detecting a particular type of threat. Instead, AV results should be presented broken down according to each family and/or file format. It would allow one to identify AV's weak and strong points and correlate it to the requirements for the targeted operational scenario.

- **AV evaluations should consider multiple datasets.** Given the differences on the detection of each threat type, AV selection should not be carried by looking to a generalized result. Instead, they should consider datasets which resemble the scenario in which the AV is supposed to op-



(a) **AV1.** Recommended for incident response teams.

(b) **AV2.** Recommended for corporate users.

(c) **AV3.** Recommended for domestic users.

Figure 12: **AV's operational aspects**, considering the six metrics proposed.

erate. We showed the need for considering distinct scenarios to evaluate AV solutions via the comparison of Brazilian with Worldwide samples. In our tests, Brazilian samples were less detectable than worldwide counterparts. Therefore, Brazilian users choosing an AV solution that best performed in the global scenario might have been overlooking the best solution for their particular scenario.

- **AV evaluations cannot be a snapshot.** AVs are dynamic mechanisms. As time goes by, signature addition/removal, ML models updates, and/or heuristic changes cause extra samples to start being detected, but, unfortunately, some other samples stop being detected at the same time. Given more time, samples might recover their detection rates. Therefore, AV evaluations should be conducted in a time-longitudinal way instead of being limited to a single observation day. Time-limited observations might bias results with regards to the detection rate obtained in the single day and not identify the AV final decision.

AV development gaps & challenges We also expect that our findings can be seen as a set of suggestions aiming at enhancing current AVs. More specifically, we advocate that:

- **AVs need to enhance their malicious web pages detection capabilities.** Our evaluation results indicate that AV performing significantly worse on detecting malicious Web pages than malicious binaries. It suggests that AVs need to improve their malicious Web pages detection capabilities. We discovered that malicious Web page detection became harder due to contextual issues: phishing pages, for instance, besides presenting malicious objects, are language-dependent so as to deceive users into clicking in the malicious links. In this sense, the use of natural language processing for such tasks is an open research question that could improve AV detection capabilities.

- **AVs need to respond faster to new threats.** Our evaluation results also showed that there is a significant attack opportunity window, i.e, a period in which AV users become vulnerable because their solutions are still not able to detect newly launched samples. It happens because AVs do not yet have signatures (or adequate heuristics) for malware samples in the sample's first appearance day since they will be developed by human analysts after the malware discovery. The time taken to unveil the sample, develop a signature, and distribute it to AV clients constitute the opportunity window. To face this delay, automated learning mechanisms should be developed and/or improved, thus reducing the need (and the significant required time) for humans to develop malware signatures. Notice that we do not claim that AV companies are not making their best to respond to the incident. Instead, our claim is that there is also a long path of technical challenges to be overcome.

- **AVs need to provide more significant labels.** Evaluating AVs' labeling is as important as evaluating the AV's detection capabilities since a good label allows for more oriented incident response procedures. Our results, however, suggest that AVs are not very good at labeling samples, presenting many generic and heuristic labels that do not allow gathering any sample information. We highlight that the development of effective automated learning procedures should be pursued since we understand that most generic labels derive from heuristic procedures. Such development would allow AVs to provide users with information about the sample's characteristics in addition to just detecting it.

On the Adoption of the Proposed Metrics. We expect that our proposed metrics might help anyone interested in the security provided by the AVs (e.g., users, companies, AV vendors) to better evaluate them. However, due to the required knowledge to to model a given user's needs and faced threats, we suppose that the metrics are more likely to be adopted by corporate users. Companies with mature security practices often have dedicated security teams able to model security needs in a very comprehensive manner.

We believe that these metrics might be made accessible to end-users via the intermediation of AV benchmarking companies, that might incorporate these metrics in their evaluation while leveraging their knowledge to highlight the most important aspects to the users. We are aware that the adoption of the proposed metrics implies that more complex explanations should be presented to the users. We can hypothesize that avoiding to explain the complexity of AV solutions is one of the reasons for the current AV evaluations to be presented in a generalized manner.

Finally, we do not expect our proposed metrics to be the only one considered by the evaluations. These should still consider the already popular metrics such as accuracy, precision, recall, and so on. In particular, the evaluations should always consider the False Positive (FP) rate, as AVs should not prevent users from running legitimate applications. FP rates have already been adopted by some AV evaluations [2] and we expect them to consider our metrics in the same manner.

Regional and cultural differences. Our evaluation results show that AVs do not present the same effectiveness on detecting all types of samples. Hence, samples from particular datasets, such as country-specific ones, are less prone to be detected than generic samples. Unfortunately, most AV evaluations do not distinguish sample's source and mix detection rates for samples from all localities into a single, non-weighted detection average rate. In this case, a user may choose an AV solution that best performs in the global scenario but that is not the best suited for his particular one. It highlights the need for considering distinct scenarios when evaluating AV solutions.

We hypothesize that Brazilian samples might have been less detected than their worldwide counterparts in part due to country particularities, as shown for Brazil in other contexts [18, 8]. We believe that this information may be used to both enhance protection in localized scenarios also help general researchers on identifying trends and attackers' behavior.

Finally, in addition to the characteristics that we found particular to the Brazilian scenario, particular malware characteristics have also been identified in other contexts, such as in China [50, 28]. Therefore, we advocate for more country-specific analysis both to understand their impact as well as to develop more targeted AV solutions.

If not Brazil? Our experiments considered the effect of Brazilian malware samples on AV detection. This raises the concern of how much of the AV result is affected by it. Although we have also considered a dataset of worldwide samples to show that the AV's behaviors are similar in both, it is natural to hypothesize that if another country malware dataset was chosen the results would be different. Whereas we are sure that the overall rates would change, we believe that the overall AV behavior would remain the same. This is because our evaluation is not about the dataset, but mainly about how AV evaluations (badly or not) operate over them. We showed that the BR dataset is different from the global dataset mainly because the BR one has a distinct distribution of filetypes and malware classes. Whereas a distinct country would present another distribution, the key point is that no country-representative dataset would be equally-balanced as typical malware evaluations are. Thus, our claim in this paper is for more realistic evaluations. We are aware that considering unbalanced dataset might also introduce bias. For instance, a malicious stakeholder might bias the dataset to favor its preferred company and/or product. In this context, the consideration of Brazilian samples played a key role, since we are able to claim that a dataset balanced like that is found in actual scenarios. Therefore, we claim that real-world data (from any country) is a good criterion for evaluating whether a good dataset is adequate for an AV experiment or not.

The future of AV solutions. Our evaluation results showed the existence of significant AVs operational gaps, such as excessively long response times. This way, an attack opportunity window is opened within the first 30 days after the release of a new sample. It does not imply that AVs must be discarded as security solutions, but that their weaknesses need to be addressed. We believe that a paradigm shift is required to reduce AV's response time, such as making them adopt more proactive detection approaches instead of current reactive operational mode. In this sense, we believe that research aiming to predict exposure [42] is a possible path towards overcoming the response time reduction challenge.

Limitations & future work. In this work, we highlighted the differences between a country-specific dataset (Brazil) and a heterogeneous dataset (World samples). Our goal was to emphasize the need for more personalized AV solutions. As complement to our results, further research work might characterize other country-specific datasets and respective AVs detection rates in these scenarios. Also, our time-series analyses were limited to a period of 30 consecutive days. We have established this limit based on our own previous experience, which showed us that this period was enough to highlight most of the characteristics that we were interested in. However, additional AV detection drawbacks might be observed by enabling longer observation periods, which is also left as a future work. Finally, our experiments only considered the detection of isolated web-pages. We acknowledge that procedures considering the entire website and/or domain might result in distinct detection results.

7 Related Work

In this section, we present closely related work about AV selection and evaluation to better position our work.

AV Product Selection. AV evaluation is often understood as a way of choosing a product to buy, instead of the best solution for some scenario. In this sense, many websites, such as AV-Comparatives [1] and AVTest [3], present AV benchmarks to evaluate detection rates, memory footprint and CPU usage. However, despite evaluating these important characteristics, these evaluations do not say much about AV efficiency, ignoring aspects such as the existing attack opportunity window, label inconsistencies and/or variant resistance [19], evaluation gaps that our work intends to fill. In addition, such evaluations are focused on individual AV products, whereas we also focus on evaluating AV products in a general way, thus identifying the current state of AV detection solutions. Another AV selection pitfall is that users often do not have enough technical knowledge to make an informed decision, thus their decisions towards picking an AV solution tend to be centered on advertisements and relation's recommendations than proper cost-benefit analyses. This problem becomes even more significant when we consider the impact of diversity [16], which is observed even in organizations that present well structured decision criteria [45]. Therefore, this work proposes metrics to better evaluate AV solutions in their multiple aspects.

AV Evaluation. Evaluating AV solutions is a hard task because most of their internal working mechanisms are closed source solutions and with limited configuration possibilities. Given this limitation, overall AV evaluations are required to develop specially-crafted samples to trigger individual AV components [38]. Therefore, most evaluation reports focus on specific factors affecting AV working, such as detection regression, when a sample stops being detected after some time [15]. In this work, we adopt an approach based on metrics to evaluate the occurrence of detection evaluation pitfalls, including detection regression.

Another challenge is to evaluate the labels assigned to multiple samples by the AVs. This evaluation requires applying criteria such as consistency and completeness [33] to evaluate the results. This allows one to identify when and how often distinct AVs do not agree on naming strains. This evaluation is important because the use of inconsistent AV labels may even decrease AV classification accuracy [9]. Whereas theoretically AV labels should be standardized by CARO, in practice, non-standard extensions are often implemented by vendors. Although some work focus on unifying AV labeling [29, 22, 41], these approaches are not practical for end-users. In this work, we evaluate the real impact of inconsistent labeling.

Given the challenges of directly assessing AV's capabilities, many academic results in the literature have their root in security work targeting other goals. For instance, an epidemiological study of malware that compromise enterprise systems [51] ended up identifying that users are targeted by threats in an unbalanced manner, and the AV they considered provided different responses for each scenario. In this work, we systematized the evaluation for multiple scenarios and presented results that extended from a single AV to multiple ones (Section 4.1 and Section 4.2). Similarly, during the evaluation of a cloud-based AV proposal [35], the authors pointed to the existence of an attack opportunity window related to the age of the malware sample. While they presented results grouped on periods of three months from a period of time of almost a decade ago in their work, we present results of today's malware on a daily-basis in ours (Section 4.3).

Recent Advances on AV Research. AVs are continuously evolving to keep up with new malware threats. This continuous evolution also affects the scope of AV evaluations, as more tests are required to exercise all AV's capabilities and features. For instance, whereas cloud-based AVs have been proposed [13], there is no real-world, specific AV evaluation to assess cloud-based AVs operation particularities. Similarly, whereas most AVs are AI-powered [23], there are few initiatives to assess their drawbacks in real cases. We consider that conducting this evaluation is extremely important as AI has already been proved to have significant weaknesses in academic scenarios that might also occur in actual scenarios [11]. We consider that establishing clear assessment metrics, such as the one here proposed, might help on overcoming AV's key challenges, such as reducing false positives [39]. This is essential for a solution to operate in real scenarios, with complex datasets, such as mailboxes of large companies [14]. The next-generation of AVs will also have to face the challenge of generating more understandable indicators of compromise [27]. We consider that the label quality metric hereby proposed might be a first step towards this direction. The next-generation of AVs, however, must not be limited to operate on typical binaries, such as the one presented on this study, but might also cover other cases, such as social media threats [4]. This evolution will also require specialized evaluation for effectiveness assessment.

8 Conclusion

In this paper, we investigated the problem of evaluating AVs in actual scenarios. To do so, we presented a longitudinal study of AV detection rates on samples daily collected from multiple malware sources and then submitted to VirusTotal by a period of consecutive 30 days. We showed the panorama of current AVs operation and identified that: (i) understanding phishing contexts is a challenge for AVs, making malicious web pages detectors less effective than their binary counterparts; (ii) generic detection procedures have not been enough to ensure broad detection coverage, incurring in lower detection rates for particular datasets (e.g., Brazilian malware) than for worldwide malware; (iii) detection rates are constantly changing, and all AVs exhibited detection regression effects even for periodic scans of the same malware dataset; and (iv) AVs long response times to deliver new signatures and heuristics offer a significant attack opportunity window within the first 30 days in which we discovered a malware sample.

To overcome existing evaluation drawbacks on these identified gaps, we proposed six new metrics for AV evaluations. These metrics consider AV's multiple aspects and operational contexts. We believe that this work may help users as well as security professionals to make proper choices regarding the best AV for each scenario and/or needs. We also hope that this work fosters smart discussion on how AV internals are really implemented, as well as instigates authors in conducting further research following our methodology either to evaluate security solutions and to describe their datasets in detail.

Acknowledgments

This project was partially financed by the Brazilian National Counsel of Technological and Scientific Development (CNPq, PhD Scholarship, process 164745/2017-3) and the Coordination for the Improvement of Higher Education Personnel (CAPES, Project

FORTE, Forensics Sciences Program 24/2014, process 23038.007604/2014-69).

References

[1] AV-Comparatives. Independent tests of antivirus software. https://www.av-comparatives.org, 2018.

[2] AVComparatives. Spotlight on security: The problem with false alarms. https://www.av-comparatives.org/spotlight-on-security-the-problem-with-false-alarms/, 2018.

[3] AVTest. Antivirus & security software & anti-malware reviews. https://www.av-test.org, 2018.

[4] S. Bell and P. Komisarczuk. Measuring the effectiveness of twitter's url shortener (t.co) at protecting users from phishing and malware attacks. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW '20, New York, NY, USA, 2020. Association for Computing Machinery.

[5] T. Beppler, M. Botacin, F. J. O. Ceschin, L. E. S. Oliveira, and A. Grégio. L(a)ying in (test)bed. In Z. Lin, C. Papamanthou, and M. Polychronakis, editors, *Information Security*, pages 381–401, Cham, 2019. Springer International Publishing.

[6] X. bin Wang, G. yuan Yang, Y. chao Li, and D. Liu. Review on the application of artificial intelligence in antivirus detection systemi. In *IEEE Conf. on Cybernetics and Intelligent Systems*, 2008.

[7] BitDefender. How important are false positives in measuring the quality of an antimalware engine? http://oemhub.bitdefender.com/importance-of-false-positives-for-antimalware-engine-quality, 2015.

[8] M. Botacin, A. Kalysch, and A. Grégio. The internet banking [in]security spiral: Past, present, and future of online banking protection mechanisms based on a brazilian case study. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ARES '19, New York, NY, USA, 2019. Association for Computing Machinery.

[9] D. Carlin, A. Cowan, P. O'Kane, and S. Sezer. The effects of traditional antivirus labels on malware detection using dynamic runtime opcodes. *IEEE Access*, 2017.

[10] CARO. A new virus naming convention. http://www.caro.org/articles/naming.html, 1991.

[11] F. Ceschin, M. Botacin, H. M. Gomes, L. S. Oliveira, and A. Grégio. Shallow security: On the creation of adversarial variants to evade machine learning-based malware detectors. In *Proceedings of the 3rd Reversing and Offensive-Oriented Trends Symposium*, ROOTS'19, New York, NY, USA, 2019. Association for Computing Machinery.

[12] F. Ceschin, F. Pinage, M. Castilho, D. Menotti, L. S. Oliveira, and A. Gregio. The need for speed: An analysis of brazilian malware classifers. *IEEE Security & Privacy*, 16(6):31–41, Nov.-Dec. 2018.

[13] D. Deyannis, E. Papadogiannaki, G. Kalivianakis, G. Vasiliadis, and S. Ioannidis. Trustav: Practical and privacy preserving malware analysis in the cloud. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, CODASPY '20, page 39–48, New

York, NY, USA, 2020. Association for Computing Machinery.

[14] L. Gallo, A. Botta, and G. Ventre. Identifying threats in a large company's inbox. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, Big-DAMA '19, page 1–7, New York, NY, USA, 2019. Association for Computing Machinery.

[15] I. Gashi, B. Sobesto, S. Mason, V. Stankovic, and M. Cukier. A study of the relationship between antivirus regressions and label changes. In *2013 IEEE 24th Inter. Symp. on Software Reliability Engineering (ISSRE)*, 2013.

[16] I. Gashi, V. Stankovic, C. Leita, and O. Thonnard. An experimental study of diversity with off-the-shelf antivirus engines. In *2009 Eighth IEEE Inter. Symp. on Network Computing and Applications*, 2009.

[17] GIAC. Chad robertson. `https://www.giac.org/paper/gcfa/4799/indicators-compromise-memory-forensics/115906`, 2013.

[18] A. R. A. Grégio, D. S. o. Fernandes, V. M. Afonso, P. L. de Geus, V. F. Martins, and M. Jino. An empirical analysis of malicious internet banking software behavior. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 1830–1835, New York, NY, USA, 2013. ACM.

[19] M. Guri, G. Kedma, A. Kachlon, and Y. Elovici. Resilience of anti-malware programs to naive modifications of malicious binaries. In *2014 IEEE Joint Intel. and Sec. Informatics Conf.*, 2014.

[20] T. J. Holt, A. M. Bossler, and K. C. Seigfried-Spellar. *Cybercrime and Digital Forensics: An Introduction*. Routledge, 2017.

[21] A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. The psychology of security for the home computer user. In *2012 IEEE Symposium on Security and Privacy*, pages 209–223, May 2012.

[22] M. Hurier, G. Suarez-Tangil, S. K. Dash, T. F. Bissyandé, Y. L. Traon, J. Klein, and L. Cavallaro. Euphony: Harmonious unification of cacophonous anti-virus vendor labels for android malware. In *IEEE/ACM Inter. Conf. on Mining Software Repositories (MSR)*, 2017.

[23] N. Kaloudi and J. Li. The ai-based cyber threat landscape: A survey. *ACM Comput. Surv.*, 53(1), Feb. 2020.

[24] A. Kantchelian, S. Afroz, L. Huang, A. C. Islam, B. Miller, M. C. Tschantz, R. Greenstadt, A. D. Joseph, and J. D. Tygar. Approaches to adversarial drift. In *AISec 2013*, 2013.

[25] P. Khodamoradi, M. Fazlali, F. Mardukhi, and M. Nosrati. Heuristic metamorphic malware detection based on statistics of assembly instructions using classification algorithms. In *Inter. Symp. on Comp. Arch. and Digital Systems (CADS)*, 2015.

[26] J. Koret and E. Bachaalany. *The Antivirus Hacker's Handbook*. Wiley Publishing, 1st edition, 2015.

[27] Y. Kurogome, Y. Otsuki, Y. Kawakoya, M. Iwamura, S. Hayashi, T. Mori, and K. Sen. Eiger: Automated ioc generation for accurate and interpretable endpoint malware detection. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, page

687–701, New York, NY, USA, 2019. Association for Computing Machinery.

[28] S. L. Lim, P. J. Bentley, N. Kanakam, F. Ishikawa, and S. Honiden. Investigating country differences in mobile app user behavior and challenges for software engineering. `https://ieeexplore.ieee.org/abstract/document/6913003`, 2014.

[29] Y. Liu, Y. Zhang, H. Wang, J. Xu, and J. Li. Research on standardization of the android malware detection results. In *2016 IEEE Int. Conf. on Net. Infrastructure and Digital Content (IC-NIDC)*, 2016.

[30] Malshare. Malshare. `https://malshare.com/`, 2018.

[31] I. Martín, J. A. Hernández, S. de los Santos, and A. Guzmán. Poster: Insights of antivirus relationships when detecting android malware: A data analytics approach. In *Proc. ACM Conf. on Comp. and Communications Security*, 2016.

[32] G. Mateaki. Pci requirement 5: Protecting your system with anti-virus. `https://www.securitymetrics.com/blog/pci-requirement-5-protecting-your-system-anti-virus`, 2017.

[33] A. Mohaisen and O. Alrawi. Av-meter: An evaluation of antivirus scans and labels. In S. Dietrich, editor, *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 112–131, Cham, 2014. Springer International Publishing.

[34] C. Nachenberg. Computer virus-antivirus coevolution. *Commun. ACM*, 1997.

[35] J. Oberheide, E. Cooke, and F. Jahanian. Cloudav: N-version antivirus in the network cloud. In *Proceedings of the 17th Conference on Security Symposium*, SS'08, pages 91–106, Berkeley, CA, USA, 2008. USENIX Association.

[36] R. Obialero. Forensic analysis of a compromised intranet server. `https://www.sans.org/reading-room/whitepapers/forensics/paper/1652`, 2006.

[37] S. Oyelere and L. Oyelere. Users' perception of the effects of viruses on computer systems – an empirical research, 2015.

[38] D. Quarta, F. Salvioni, A. Continella, and S. Zanero. Toward systematically exploring antivirus engines. `https://conand.me/publications/quarta-crave-2018.pdf`, 2018.

[39] D. Sacher. Fingerpointing false positives: How to better integrate continuous improvement into security monitoring. *Digital Threats: Research and Practice*, 1(1), Mar. 2020.

[40] D. J. Sanok, Jr. An analysis of how antivirus methodologies are utilized in protecting computers from malicious code. In *Proc. Annual Conf. on Inf. Sec. Curriculum Development*, 2005.

[41] M. Sebastián, R. Rivera, P. Kotzias, and J. Caballero. Avclass: A tool for massive malware labeling. In F. Monrose, M. Dacier, G. Blanc, and J. Garcia-Alfaro, editors, *RAID*, 2016.

[42] M. Sharif, J. Urakawa, N. Christin, A. Kubota, and A. Yamada. Predicting impending exposure to malicious content from user behavior. In *ACM CCS*, 2018.

[43] P. Soni, S. Firake, and B. B. Meshram. A phishing analysis of web based systems. In *Proceedings of the 2011 International Conference on Communication, Computing &#38; Security*, ICCCS '11,

pages 527–530, New York, NY, USA, 2011. ACM.

[44] V. G. Tasiopoulos and S. K. Katsikas. Bypassing antivirus detection with encryption. In *Proc. Panhellenic Conf. on Informatics*, PCI '14, 2014.

[45] N. B. Vasilyevna, S. S. Yeo, E. S. Cho, and J. A. Kim. Malware and antivirus deployment for enterprise it security. In *Symp. on Ubiquitous Multimedia Comp.*, 2008.

[46] VirusShare. Virusshare. `virusshare.com`, 2018.

[47] VirusTotal. Av comparative analyses, marketing, and virustotal: A bad combination. `https://blog.virustotal.com/2012/08/av-comparative-analyses-marketing-and.html`, 2012.

[48] VirusTotal. Public api version 2.0. `https://developers.virustotal.com/reference`, 2018.

[49] VirusTotal. Virustotal. `http://www.virustotal.com/`, 2018.

[50] H. Wang, Z. Liu, J. Liang, N. Vallina-Rodriguez, Y. Guo, L. Li, J. Tapiador, J. Cao, and G. Xu. Beyond google play: A large-scale comparative study of chinese android app markets. `https://arxiv.org/pdf/1810.07780.pdf`, 2018.

[51] T.-F. Yen, V. Heorhiadi, A. Oprea, M. K. Reiter, and A. Juels. An epidemiological study of malware encounters in a large enterprise. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1117–1130, New York, NY, USA, 2014. ACM.

A Experiments with local AVs

Current AV's are complex software pieces and present multiple operation modes. This includes real timing monitoring methods, cloud-based scans, and other multiple features. The AV's versions running on VirusTotal are only limited versions of local AV installations. More specifically, VirusTotal often provides only command-line versions of AVs that are triggered only on-demand. This difference raises concerns with regards to the validity of our findings when considered the actual scenario of a user using a local version of an AV solution. To increase our confidence in the reported results, we cross-checked the results obtained using VirusTotal and using local AVs. Due to scaling issues, we cannot repeat all experiments previously presented and/or test all AVs available on VirusTotal. Therefore, we limited our checking procedures to a subset of them. We opted to repeat the experiment shown in Section 4.2 (using the same dataset). We selected the three most popular AVs in the online software repositories rankings that we visited for this experiment: ESET NOD32 12.0, Kaspersky 20.0, and Symantec Norton 360. They were all installed using their default configurations.

The first significant difference between VirusTotal AV's versions and the local ones is that some samples started being detected as soon as we added them to the test machine due to the real-time monitoring features. This behavior was observed in all AVs. Apart from this behavior, no significant difference was observed. Figure 13 shows the detection rates for the distinct malware classes upon a manually triggered file scan. We notice that although the detection rates in fact increase a little bit from the VirusTotal's version to the local ones, the overall picture remains the same: distinct malware classes present distinct detection rates. Thus, we are confident that the conclusions presented along the entire paper hold true

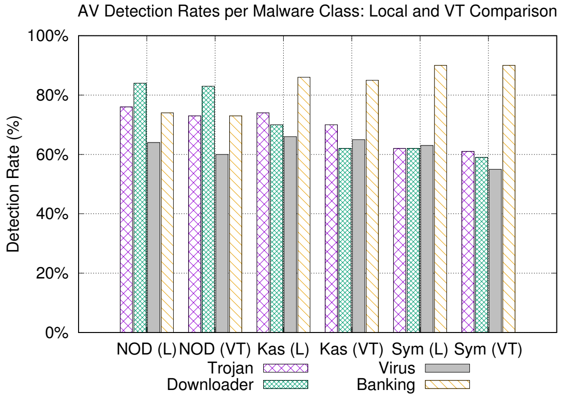


Figure 13: **Comparing VirusTotal's and local's AV versions.** Although the detection rate increased a bit, AVs kept presenting distinct rates for each malware class.

in actual scenarios. We acknowledge that this experiment does not mean to be the definitive conclusion of whether VirusTotal is reliable for malware evaluations or not. Instead, we claim that it helps to increase our confidence in the average results reported in the paper.