IEEE Open Journal of the Computer Society

Received 18 January 2024; accepted 6 February 2024. Date of publication 9 February 2024;
date of current version 22 February 2024. The review of this article was arranged by Associate Editor Jiankun Hu.

Digital Object Identifier 10.1109/OJCS.2024.3364576

Unveiling the Connection Between Malware and Pirated Software in Southeast Asian Countries: A Case Study

ASIF IQBAL [1] (Member, IEEE), MUHAMMAD NAVEED AMAN [2] (Senior Member, IEEE), RAMKUMAR REJENDRAN [1], AND BIPLAB SIKDAR [1] (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583
[2]School of Computing, University of Nebraska-Lincoln, Lincoln, NE 68523 USA

CORRESPONDING AUTHOR: MUHAMMAD NAVEED AMAN (e-mail: naveed.aman@unl.edu).

ABSTRACT Pirated software is an attractive choice for cybercriminals seeking to spread malicious software, known as malware. This paper attempts to quantify the occurrence of malware concealed within pirated software. We collected samples of pirated software from various sources from Southeast Asian countries, including hard disk drives, optical discs purchased in eight different countries, and online platforms using peer-to-peer services. Our dataset comprises a total of 750 pirated software samples. To analyze these samples, we employed seven distinct antivirus (AV) engines. The malware identified by the AV engines was classified into four categories: adware, Trojans, viruses, and a miscellaneous category termed others. Our findings reveal that adware and Trojans are the most prevalent types of malware, with average infection rates of 34% and 35%, respectively, among our pirated software samples. Notably, our evaluation of AV detection performance highlights variations in sensitivity, ranging from a high of 132% to a low of 30% across all AV engines. Furthermore, upon installing pirated software, the most adversely affected operating system settings are the firewall and user account control configurations. Given the potential for malware to steal information or create malicious backdoors, its high prevalence within pirated software poses a substantial security risk to end users.

INDEX TERMS Software piracy, malware, anti-virus.

I. INTRODUCTION

Malware generally refers to software that is deployed with a malicious intent. Such software is typically used to obtain information that can be used for monetary gains (e.g., login credentials, bank and credit card numbers, and intellectual property), or for launching cyber attacks on network based services and cyber-enabled infrastructure (e.g., the attacks on the power grid in Ukraine in 2015 and 2016 [1]). Malware affects computing systems worldwide, and results in significant financial and productivity losses. According to the annual Evil Internet Minute report from RiskIQ [2], in just one minute on the Internet, $2.9 million is lost to cybercrime, i.e., cybercrimes accounted for a total of $1.5 trillion loss to the global economy.

The development and exploitation of malware is often carried out by criminal organizations [3]. One of the means exploited by cyber-criminals for infecting computers with malware is to propagate it through pirated software or media. It is a well known fact that free (pirated) software is offered to customers as an incentive by various (non-law-abiding) computer vendors and information technology service providers in many parts of the world. Moreover, it is easy to find pirated software in optical discs (DVD) openly or covertly sold by various small shops in many developing countries. Pirated software is also easily accessible on the Internet, e.g., through the use of peer-to-peer networks.

Although "free" versions of software should raise suspicion, many users of such software are not aware (or choose to ignore) that the free software being offered may have undesirable consequences in the form of malware infections. According to a Business Software Alliance (BSA) estimate, the worldwide piracy rate is 39% with a $52.3 billion

commercial value of unlicensed software [4]. However, more alarming is the high rate of unlicensed software being used in legally-inclined organizations. For example, the top 100 colleges/universities and top 50 engineering schools in the USA had a piracy rate of above 99% in 2018 as determined by U.S. News and World Report [5]. Similarly, the top 100 U.K. universities also had a piracy rate of about 100% in 2018 as estimated by Hoovers [5]. In the same year, according to Fortune and Software magazines, the piracy rate for Fortune 100 and Software 100 companies in the US was 75% and 85%, respectively [5]. Similarly, the top 100 manufacturing firms in the U.K. had a piracy rate of 76% in 2018 [5]. Moreover, pirated software usage is also high in businesses and professionals in Southeast Asian countries [6]. Thus, pirated software is prevalent not only in computers that belong to individuals, but also in organizations and may be a contributing factor towards the $25 per Internet minute that major companies are paying because of security breaches [7]. Moreover, a study conducted by Atlas VPN [8] found that during the first quarter of 2022, almost 80% of the malware targeted Microsoft Office vulnerabilities, growing from 61% in third quarter of 2021. As Microsoft Office is one of the most pirated software in use, its pirate user base is most vulnerable to hackers abuse.

While software piracy is a global concern, its manifestations vary across regions. In several first-world countries, stringent anti-piracy and copyright regulations often create hurdles for obtaining pirated software physically. Despite reasonable software pricing in these regions, the widespread access to high-speed internet still prompts users to opt for illicit online channels to acquire pirated copies of the software. Conversely, in developing regions like Southeast Asia, enforcement of anti-piracy laws might be less rigorous. Software costs are relatively high, posing financial challenges for individuals with lower income levels, thereby making genuine software purchases difficult. Consequently, pirated software, available in physical form, is readily accessible through small shops within computer equipment malls in the region. As discussed earlier, the usage of pirated software poses significant risks due to potential malware infections, endangering system security and potentially compromising sensitive data, leading to substantial vulnerabilities in computer systems and networks.

Given the prevalence of pirated software in developing nations, this paper focuses on conducting a case study centered on Southeast Asian countries. Our study involves an examination of the malware presence within pirated software. Alongside pirated software acquired from online sources, we obtained physical copies of pirated software integrated into pre-built computer systems and bundled in DVD formats. With the acquired dataset of pirated software, our primary aim is to quantitatively investigate the question, *"what is the relation between malware and pirated software?"*. The major contributions of this paper are that it addresses the following questions which then lead to the main objective:

1) Which type of malware is more common in pirated software?
2) Which source of pirated software has the highest prevalence of malware?
3) Can we determine the most sensitive antivirus engine based on its malware detection rate?

To answer these questions, this paper analyzes over 555 samples of personal computers, DVDs, as well as software downloaded from peer-to-peer networks using a suite of seven AV software. Our study revealed that adware and Trojans constitute the most prevalent types of malware, with infection rates averaging 34% and 35%, respectively, among the examined pirated software samples. The infection rate for pirated software obtained from DVDs was the highest, averaging approximately 1.17 (117%) malware instances per software, followed closely by hard disk drives (HDDs), which showed a rate of 96%. In contrast, downloaded samples exhibited a lower infection rate, at 26%. Additionally, our analysis of AV detection performance showcased sensitivity disparities, ranging from a high of 132% (1.32 malware per sample) to a low of 30% across various AV engines. Moreover, upon installation of pirated software, the firewall and user account control configurations within the operating system experienced the most substantial adverse effects.

The rest of the paper is organized as follows: Section II describes the related work in existing literature. Section III presents the analysis methodology used in this paper. Section IV presents the results and discussion. We conclude the paper in Section V.

II. BACKGROUND AND LITERATURE REVIEW

A. MALWARE: CHARACTERISTICS AND CATEGORIES

Malware come in many forms and can infect a wide range of devices such as personal computers, servers, smart phones, printers, and embedded devices [9], [10] etc. Malware may generally be characterized by the following four attributes of its operations [11], [12]:

1) *Propagation:* The process through which malware may be distributed to multiple systems by the adversary or an infected host.
2) *Infection:* The means of installing the malware, e.g., the installation file.
3) *Self-Defense:* The mechanism to stay hidden and evade detection.
4) *Capabilities:* Command features available to the adversary.

Based on these characteristics, in this study, the malware identified by the AV engines has been categorized into three primary classes and one general class, briefly described as follows:

1) *Adware:* This category encompasses relatively less harmful malware. Adware exposes users with unwanted and potentially malicious advertisements, often altering the browser's home page and default search engine settings. The primary objective behind adware is to generate revenue through user clicks on advertisements or, in extreme cases, as a delivery method for more

dangerous malware types like spyware and keyloggers. Spyware stealthily resides within a host system, gathering reconnaissance data to identify vulnerabilities. Similarly, keyloggers, a form of spyware, capture every keystroke on a compromised computer and transmit it to a remote server, often targeting critical credentials such as passwords.

2) *Trojan:* Trojans masquerade as legitimate programs but contain malicious instructions that can compromise or harm the user's computer. Activation of a Trojan typically requires user interaction. For instance, Trojans might disguise themselves as security patches or antivirus programs. Users can inadvertently activate Trojans when prompted by fake antivirus pop-ups on infected websites, leading them to download and run these malicious programs. Common types of Trojans include mailfinders, DDoS Trojans, Banking Trojans, and Ransomware Trojans, among others [13].
3) *Virus:* A computer virus is a malicious program capable of self-replication by modifying legitimate computer programs and injecting its code. Viruses are particularly hazardous as they infect other files, making the infected files carriers of the virus. To eliminate viruses, most antivirus engines delete the infected file instead of removing the virus itself. The term 'virus' is often misused, commonly applied to various types of malware. However, viruses specifically refer to malware that self-propagates by embedding its code into legitimate files.
4) *Other:* For samples not labeled by the identifying AV engine, we conducted online searches using their names. If an appropriate category was found matching the description, we placed them accordingly into the previously discussed categories. For those that did not yield any matching category, they were included in this miscellaneous 'Other' category.

Traditionally, cyber-criminals have focused on the sheer volume of malware. The objective was simple: reap as much reward as possible by casting a big net and infect as many computers as possible. However, with more awareness into cyber threats, organizations have started spending more on cyber defenses. This caused a reduction in returns for cyber-criminals, and resulted in a shift towards business-focused and budget-conscious strategies. Thus, cyber-criminals have shifted from a volume focused strategy to a more targeted approach. One such approach has been the bundling of malware with pirated software [14]. Computing equipment infected with malware at the point-of-sale or subsequently downloaded/installed by users who are willing to trade cost for security by opting for pirated software, either knowingly or unknowingly, provide cyber-criminals with an easier means of acquiring compromised targets.

B. PREVIOUS WORK

Some of the existing works on malware highlight the impact of technical, demographic, and socio-economic factors on the prevalence of malware. For example, the authors of [15], used demographic factors such as age and gender and behavioral factors such as application usage patterns to identify high risk subjects. Similarly, the authors of [16], [17] evaluate the likelihood of malware incidence based on network traffic. They analyzed the security logs of an enterprise to characterize the likelihood of malware incidence among the enterprise personnel. The authors in [18] investigated the security risk associated with using free software acquired from online public repositories. They highlighted that not only some downloaded installers were vulnerable to content tampering, but also found that 30% of all analyzed samples were infected with malware. The impact of human behavior in the spread of Internet based malware was studied in [19]. In [20], authors employed a spoofed New Zealand IP address to examine online advertisements on movie piracy websites for malware infections. Their findings revealed that over 97.24% of the high-risk ads discovered on movie piracy websites contained malware. Similarly, in [21], an investigation into the security threats associated with accessing multimedia content through piracy websites was conducted. The study, focusing on 50 websites accessed in the Philippines, highlighted that Filipino consumers visiting piracy sites face a significantly higher risk of malware infections, i.e., 16.66 times greater on torrent sites and 21.66 times greater on streaming sites compared to mainstream websites. Furthermore, in [22], authors studied a sample size of 5000 individuals across the Asia-Pacific region to analyze the relationship between advertisements placed on piracy websites and malware. Their study revealed that typical users visiting these sites were exposed to various malware, ranging from ransomware and Trojan horses to other advanced persistent threats. They advocated for the implementation of regulatory measures to mitigate their influence and render such sites more challenging to operate. To summarize, understanding the correlation between malware and piracy remains crucial in enhancing cybersecurity measures and formulating effective strategies to mitigate the risks posed by such illicit online activities.

Other existing works on malware focus on known samples of malware and investigate their propagation and evasion strategies. For example, the authors of [23] evaluated the detection evasion capabilities of different malware samples. Similarly, the recent 2020 SonicWall's cyber threat security report [24] evaluated the incidence of various types of malware globally. In contrast to these studies which either study the human behavioral aspects of cyber-security or the incidence of various malware globally, this paper focuses on quantifying pirated software as a source of malware.

The closest existing works to ours are [25] and [26]. In [25], authors analyzed the malware incidence in the pirated software that comes bundled with newly purchased computers. The studied samples were purchased from 11 different countries. In addition, this study evaluates the types of malware present along with source of the infected files. In [26] the authors used Symantec AV telemetry data to study the prevalence of malware internationally and highlighted pirated software as a source of malware. However, this study focuses

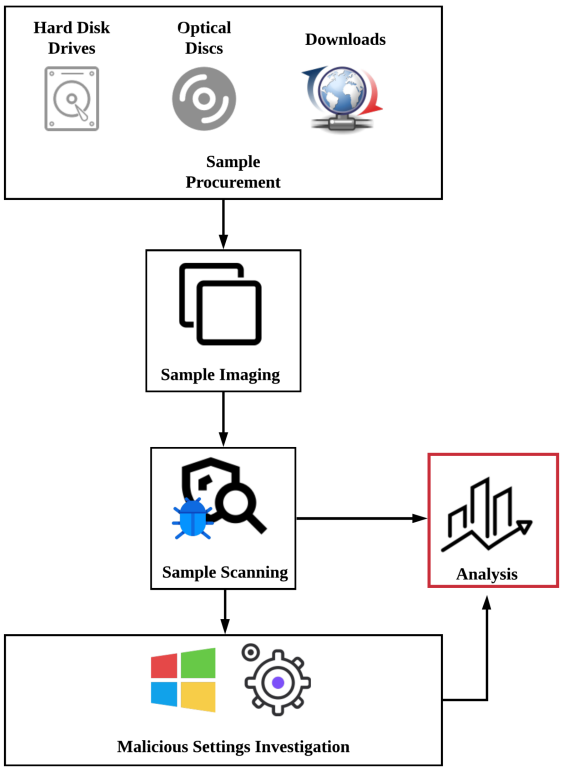

FIGURE 1. Methodology.

on pirated software downloaded from the Internet and peer-to-peer networks only without quantifying the incidence of malware in these samples. This paper focuses on quantifying the incidence of malware in pirated software obtained from three different sources, i.e., the Internet, HDDs in newly purchased computers, and DVDs. Moreover, we also check the effectiveness of seven different AV engines in detecting malware in the pirated software samples obtained from these sources.

III. METHODOLOGY

To quantitatively establish the link between pirated software and malware, this paper carried out extensive analysis of samples collected across various countries in Southeast Asia. The various steps carried out for the analysis in this paper are shown in Fig. 1 and described below.

A. SAMPLE PROCUREMENT

While software piracy remains a global issue, obtaining physical copies of such software is often challenging in most first-world countries due to the strict enforcement of anti-piracy and copyright laws. In addition, the widespread availability of high-speed internet in these nations grants users easy access to pirated software online. However, the situation differs in developing countries, such as those in Southeast Asia, where anti-piracy and copyright regulations are not rigorously enforced, and high-speed internet access is limited.

TABLE 1. Sample Procurement

Country	HDDs	Optical discs
Malaysia	10	30
Philippines	10	29
Indonesia	10	31
Vietnam	10	30
Korea	15	2
Thailand	15	43
Sri Lanka	10	–
Bangladesh	10	–

Consequently, the market for physical forms of pirated software, like DVDs, thrives in these regions. Furthermore, the exorbitant costs associated with essential software, including operating systems (OS), productivity tools, and enterprise software, drive users in these countries towards piracy as a cost-effective alternative. Therefore, for our case study, we concentrated our efforts on procuring physical copies of pirated software in the Southeast Asian countries detailed in Table 1.

We considered three distinct sources of pirated software in our study. These sources included HDDs extracted from new personal computers preloaded with pirated software by the seller, DVDs, and pirated software obtained through peer-to-peer file sharing applications on the Internet. It's important to note that, for the purpose of acquiring HDDs, we procured both desktop and laptop computers. Detailed information regarding the HDD and DVD samples can be found in Table 1. In total, we acquired 90 HDDs, 165 DVDs, and 300 pirated software download samples. Each HDD came preinstalled with a range of software, including a Windows OS, an office suite, and, in some instances, other software tools like photo and video editors, PDF readers, and antivirus software. None of these software components were genuinely licensed, unless if freeware. The software samples obtained through DVDs and downloads were categorized into three groups: (i) operating systems, (ii) productivity and enterprise software (such as office applications, Adobe suite, photo and video editors, etc.), and (iii) various other software types, including games and antivirus software. To ensure that our study encompassed widely-used pirated software, we selected popular pirated software based on download count and active seeds, representing commonly sought illicit applications.

The HDD and DVD samples were obtained by independent contractors hired by the authors. These contractors acted as ordinary customers seeking computing hardware and software without specifically requesting pirated software. Our focus was on traditional personal computer retail settings, where customers engage in discussions about their computing needs before making purchases. In such scenarios, sales personnel often provided incentives to boost their sales revenue by offering pirated software, such as operating systems, office applications, and game bundles. The sellers of the samples were

randomly chosen, mostly comprising small shops located in shopping complexes that specialize in computer hardware and software, as well as standalone shops in street markets. Notably, we did not acquire any samples from globally or nationally recognized stores, or directly from manufacturers, as they typically do not install pirated software in their sales. Furthermore, the procurement of HDD and DVD samples took place between late 2015 and early 2016, while the software downloads were completed in the 3rd quarter of 2022. Additionally, our samples do not overlap with those from [25].

B. SAMPLE IMAGING

To preserve the original samples of HDDs during the analysis, we first created disc images. A sector-by-sector copy of each HDD was created using Microsoft's `Disk2vhd` software tool. All partitions of the HDDs were selected for creating the disc images. `Disk2vhd` creates virtual hard disk clones of the actual physical disks which can then be loaded in a virtual machine for scanning the HDDs for infections, malware, and other forms of tampering. Thus, any breakout or inadvertent modifications to an HDD sample during the sample scanning and investigation stages can be avoided using sample imaging. The creation of sample images also facilitates the use of a separate untouched copy of the original sample with various scanning software in the sample scanning stage.

Our acquired dataset consists of a total of 555 individual sources comprising 90 HDDs, 165 DVDs, and 300 downloads. Each HDD came preinstalled with an OS, an office suite, and few software tools, none of which were genuinely licensed. A total of approx. 230 software were found in these 90 HDDs, with an average of approx. 2.6 application software per HDD. Here, we consider the entire installed OS (as well as the office suite) as distinct application sample as it is installed from a single source. In case of the DVDs, apart from OS discs, the rest contained productivity, enterprise, or bundled software, making up a total of 220 software samples from 165 discs (average of 1.33 software per disc). In contrast, each of the 300 downloaded samples was a standalone pirated software. In total, we had a total of approx. 750 pirated software samples for analysis.

C. SAMPLE SCANNING

To detect various forms of malware within the acquired HDD, DVD, and download samples, we used a combination of free and paid AV engines to address both corporate and personal usage scenarios. The paid AVs encompassed McAfee, Ikarus, and Norton, while the freeware category included AVG, BitDefender, Kaspersky, and Windows Defender. Apart from Windows Defender, the remaining freeware AVs also offer paid subscription options. Based on yearly surveys [27], [28], these software solutions have consistently emerged as the top and most widely utilized security applications in both corporate and personal contexts.

To hide the true identity of an AV during the analysis, we have randomly labeled the AV engines as $A_1, A_2, \ldots, A_7$. Note that a separate copy of a sample was used with each AV to ensure that any inadvertent changes to the image by one AV engine does not affect the results of other engines. Thus, we make sure that a sample scanned by each AV engine for a given sample is same as the original and identical. Moreover, the HDDs have preinstalled software, whereas, the DVDs and downloads contain software installation binaries. To make the comparison fair, we used the following rules for each AV:

1) The latest definitions and updates are installed before scanning.
2) The AV engine was configured to scan all files and directories.
3) The option for automatically removing an infection was turned off. The detected infections were copied and saved for further analysis.
4) The details of each scan such as type, name, and location of a detected infection were recorded.

Consider an HDD infected with malware X; the AV may detect multiple files in the HDD infected with the same malware X, treated as a single case of malware detection. Therefore, if an AV identifies the same malware in 10 different files, it is counted as a single positive instance. This principle applies to installation binaries in DVDs and downloads as well. For DVDs and downloaded samples, we conducted scans using the AV engines without prior installation. Since a single software sample may contain multiple associated installation files, there is a possibility that a single (or multiple) malware may infect multiple files, counted as a single (or multiple) detection. This phenomenon is illustrated in Fig. 3, where the cumulative average detection rates for DVDs reached approximately 117%, and in Fig. 4, where A_3 exhibited the highest average detection rate of 132%. These findings highlight that, on average, more than one malware detection occurs per software sample.

D. MALICIOUS SETTINGS INVESTIGATION

Some low risk malware such as adware may change some critical operating system (OS) settings, thereby paving the way for more harmful malware such as Trojans and viruses. To identify this type of malicious tampering of the OS settings, we used the Microsoft's Hyper V virtualization platform to load a virtual hard disk image of a sample HDD. The following information was then checked in the OS settings:

1) Is the default firewall enabled?
2) Is the search engine for Internet Explorer set to the default?
3) Have the remote assistant default settings been changed?
4) Have the user account control (UAC) default settings been changed?
5) Has Windows defender been disabled?

E. ANALYSIS

The data collected in this paper was analyzed using statistical tools in Microsoft Excel as well as the open source software R. To compare the prevalence of various infections, we used the Mann-Whitney test [29] at a significance level of 95%.

TABLE 2. List of Symbols Used in the Analysis

Symbol	Description / Value
S_i	Source i, $i \in [1:\text{HDD}, 2:\text{DVD}, 3:\text{Downloads}]$
$\lvert S_i \rvert$	Number of software samples in all of S_i
M_j	Malware j, $j \in [1:\text{Adware}, 2:\text{Others}, 3:\text{Trojan}, 4:\text{Virus}]$
A_k	Anti-virus engine k, $k \in [1, 2, \ldots, 7]$
$\mathcal{M}_{ijk}$	Number of M_j detected by A_k in S_i
$\mathcal{P}_{ijk}$	Percentage of M_j detected by A_k in S_i, see (1)

Moreover, the average malware detection percentage and 95% confidence intervals were considered for evaluating various plots.

IV. RESULTS AND DISCUSSION

In this section, we analyze the data obtained by scanning software samples found in HDDs, DVDs, and downloads. Subsequently, we discuss the results and draw inferences related to the objectives of this paper. For a clearer discussion of the results, Table 2 provides a summary of key symbols and their descriptions. Instead of reporting $\mathcal{M}_{ijk}$, the number of M_j detections in each source S_i by the AV engine A_k, we chose to report the detection percentages $\mathcal{P}_{ijk}$ (as in (1)). This choice allows for comparative analysis across sources, considering variations in the number of software samples found in each source, i.e., 230 in HDDs, 220 in DVDs, and 300 in downloads.

$$\mathcal{P}_{ijk} = \frac{\mathcal{M}_{ijk}}{|S_i|} \times 100 \tag{1}$$

A. PREVALENCE OF INFECTIONS

To analyze the malware prevalence in the acquired dataset, we scanned files from all three sources using the 7 AV engines and recorded the number of detections as $\mathcal{M}_{ijk}$ as described in Section III-C. Using (2), we computed the average detection percentage of each malware (M_j) as reported by AV engines for software found in all three sources and present them in Fig. 2. The figure shows that out of all 750 analyzed samples, on average, 33.7% and 35% were infected by adware and Trojan malware respectively, whereas, the recorded infection rates for others and virus types were 7% and 3.4%, respectively. This sums to an overall 79% infection rate, on average, for the scanned pirated software, which is alarmingly high. The sheer high number of adware and Trojan infections are expected as these are the most common causes of further infections as they serve as gateways for other malicious parties to attack an infected system (see discussion in Section IV-C).

$$\mathcal{X}_j = \frac{1}{3} \times \frac{1}{7} \times \sum_{i=1}^{3} \sum_{k=1}^{7} \mathcal{P}_{ijk} \tag{2}$$

To compare the prevalence of the different malware types, we compute the likelihood that the observations from the four groups are generated from the same population. The classification of malware as adware, Trojan, virus, or others after running one of the AV engines and taking two categories of malware at one time forms one set of samples of observations. The medians of the number of detections for the two categories are then compared with the non-parametric Mann-Whitney test. The null hypothesis represents the case when the two samples are drawn from the same population, i.e., equal medians. The test is repeated for each pair of malware and the results are shown in Table 3.

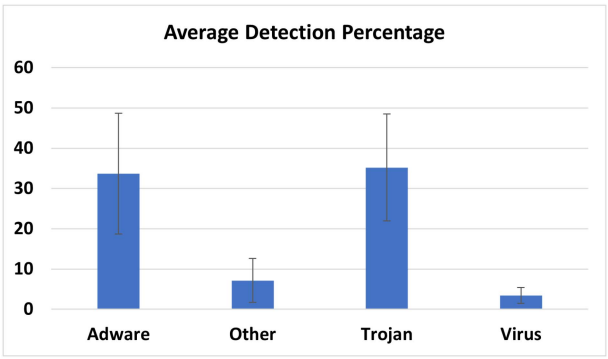

FIGURE 2. Average detection percentage for each malware type over 7 AV engines and 3 sources, along with the 95% confidence interval value.

TABLE 3. Mann–Whitney Significance Values for Consistency Among Malware Types

	Adware (p)	Other (p)	Trojan (p)	Virus (p)
Adware	–	0.0041	0.7104	0.0006
Other	0.0041	–	0.0069	0.1098
Trojan	0.7104	0.0069	–	0.0006
Virus	0.0006	0.1098	0.0006	–

Considering a significance level of $\alpha = 0.05$, we observe that the prevalence of adware as compared to virus and other type differs significantly, i.e., the p-value is less than α. However, the median number of infections for adware and Trojans are approximately the same. Similarly, the prevalence of other type against Trojans is significantly different while approximately the same to viruses. We also observe a significant difference in the prevalence of Trojans and viruses. Thus, we can infer that the most common type of malware in our samples is adware and Trojans. Similarly, the least number of infections in our samples was caused by others and viruses.

To study the prevalence of malware types in the three sources of pirated software, we apply the Mann-Whitney test to each source. The results for the software found in HDD samples are given in Table 4. We observe that the prevalence of adware, others, and Trojans is approximately the same in HDDs. However, the prevalence of viruses is significantly different. Similarly, the results for the Mann-Whitney test on DVDs and downloads are given in Tables 5 and 6, respectively. We observe that the prevalence of adware and Trojans

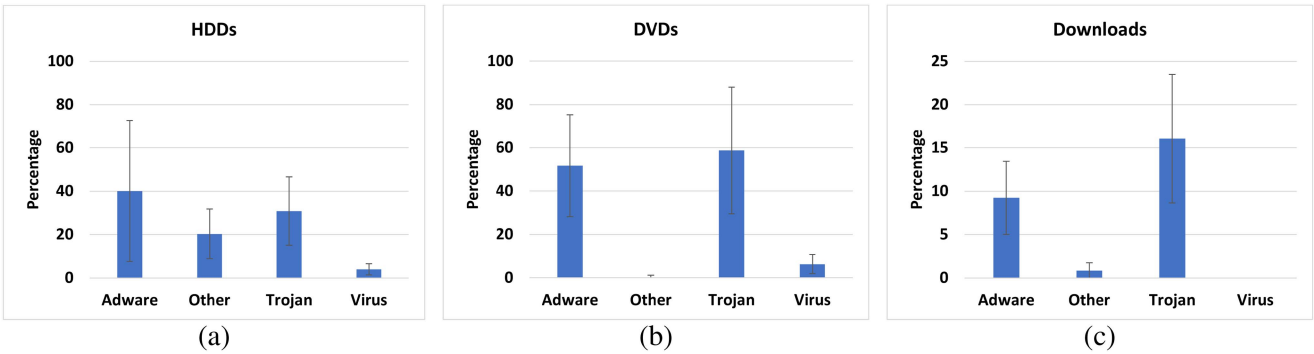

FIGURE 3. Average malware detection percentage found in software samples found in (a) HDDs, (b) DVDs, and (c) Downloads. 95% confidence intervals are also included.

TABLE 4. Mann–Whitney Significance Values for Consistency Among Malware Types in HDDs

	Adware (p)	Other (p)	Trojan (p)	Virus (p)
Adware	–	0.3638	0.7721	0.0150
Other	0.3638	–	0.4134	0.0213
Trojan	0.7721	0.4134	–	0.0071
Virus	0.0150	0.0213	0.0071	–

TABLE 5. Mann–Whitney Significance Values for Consistency Among Malware Types in DVDs

	Adware (p)	Other (p)	Trojan (p)	Virus (p)
Adware	–	0.0017	0.7104	0.0048
Other	0.0017	–	0.0017	0.0196
Trojan	0.7104	0.0017	–	0.0071
Virus	0.0048	0.0196	0.0071	–

TABLE 6. Mann–Whitney Significance Values for Consistency Among Malware Types in Downloads

	Adware (p)	Other (p)	Trojan (p)	Virus (p)
Adware	–	0.0093	0.2183	0.0010
Other	0.0093	–	0.0023	0.0293
Trojan	0.2183	0.0023	–	0.0010
Virus	0.0010	0.0293	0.0010	–

is approximately the same while the prevalence of others and viruses is significantly different in software found in DVDs and downloads.

To study the effect of the source of a pirated software on the prevalence of these infections, we computed the average detection percentage for each M_j found in software samples found in HDD, DVD, and download sources separately using (3) and show them in Fig. 3.

$$\mathcal{Y}_{ij} = \frac{1}{7} \times \sum_{k=1}^{7} \mathcal{P}_{ijk} \tag{3}$$

From Fig. 3 we observe that the DVDs have the highest cumulative infection score of 117% while downloads have the lowest at 26%. For HDDs, this number is at 96%. This statistic is alarming as it highlights that, on average, one should expect more than one malware per pirated software if the source is a DVD. Additionally, the 96% average infection rate for HDDs emphasises the users to perform a fresh OS install or a thorough AV scan even for a newly purchased computer. In contrast, the downloaded software had the lowest average infection rate, which shows that the websites offering such pirated software are taking positive steps towards ensuring malware free distribution.

In terms of infection type, adware was prevalent in all sources, with DVDs having the highest adware prevalence recorded at 52%, whereas, adware prevalence in HDDs and downloads was found at 40% and 9% respectively. The prevalence of Trojans was highest in DVDs as well, recorded at 59%, which is higher than the 28% for HDD and 16% found in downloaded software samples. HDD samples had the most others malware infections at 22%, whereas, the DVD and download sources have $< 1\%$ infections categorized as others. On the other hand, the virus malware had consistently low detection rates for all sources, especially in downloaded sources, where we found no viruses.

The results above provide the answer to the first question in the objectives of this paper, i.e., "which type of malware is more common in pirated software?". We observe that the prevalence of adware and Trojans is the highest. We can also answer the second question, i.e., "which source of pirated software has the highest prevalence of malware?". Adware malware was more prevalent in HDDs while Trojans were more prevalent in DVDs and downloaded samples. Moreover, we observed the lowest number of malware in the downloaded samples. However, we do note that a user may be exposed to malware at various steps of the process of downloading

pirated software from Internet based sources [20], [21], [22]. These include drive-by-downloads and pop-ups from the websites hosting these pirated software or the torrents for the software. Furthermore, based on our findings, the Trojans are more prevalent than viruses in all our sources, which is inline with [30]. Additionally, [31] reported that Trojans accounted for approx. 64% of all malware based attacks on windows system, followed by viruses at 15%. This high prevalence of Trojans can be attributed to their ease of delivery as compared to viruses as they can be hidden inside the legitimate programs or files, making them more likely to be installed by a user. Additionally, Trojans are often used as a "delivery mechanism" for other malware, which can explain their prevalence as well.

B. EFFECTIVENESS OF AV ENGINES

Given the vast size of our dataset, comprising millions of files within HDDs and DVDs, establishing a definitive ground truth for malware infections across the entire dataset proves to be impractical. Consequently, evaluating the true detection accuracy of AV engines becomes unfeasible. In this section, instead of attempting a comprehensive evaluation, we focus on comparing the performance of the seven AV engines concerning their sensitivity in detecting various types and number of malware. This sensitivity factor stands as a crucial feature of an AV as in standard situations, a false positive carries a significantly lower cost compared to a false negative. The latter, if overlooked, could result in severe security ramifications, impacting individual users on a smaller scale as well as entire corporate infrastructures. For analysis, we compute the average percentage of the individual M_j detected by each AV engine A_k in samples acquired from all three sources (S_i) using (4).

$$\mathcal{Z}_k = \frac{1}{3} \times \frac{1}{4} \times \sum_{i=1}^{3} \sum_{j=1}^{4} \mathcal{P}_{ijk} \tag{4}$$

Fig. 4 presents these computed metrics highlighting that the average malware detection percentage by different AV engines is not the same, i.e, some AV engines may have a higher detection rate than others. For example, A_7's average detection rate was the lowest at 30% for software samples coming from all sources, whereas, A_3 had the highest score of approx. 132%. This shows that some AV engines are very conservative in flagging a detection as infection, whereas, some (like A_3) are more trigger happy. Moreover, based on our analysis we have found that some AV engines would flag rather benign programs as malware, even though they aren't. An example of this is the `AutoKMS`, which is a generic hacktool used for illegal activation of Microsoft Office and Windows OS applications. Although if acquired from trusted sources, the `AutoKMS` is benign, but it may as well contain other malware if the source was untrustworthy. To further analyze this, we present the results of the Mann-Whitney test to compare the median number of malware detected by different AV engines in Table 7. Here we observe that the malware detection rate on the average does not differ significantly among the different AV engines.

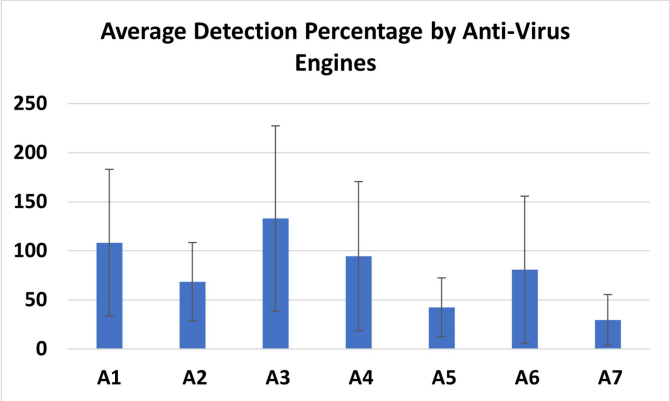

FIGURE 4. Average detection percentage for different AV engines over all malware types and 3 sources, along with the 95% confidence interval value.

TABLE 7. Mann–Whitney Significance Values for Consistency Among Malware Detection in AV Engines

	$\mathbf{A}_1$ (p)	$\mathbf{A}_2$ (p)	$\mathbf{A}_3$ (p)	$\mathbf{A}_4$ (p)	$\mathbf{A}_5$ (p)	$\mathbf{A}_6$ (p)	$\mathbf{A}_7$ (p)
$\mathbf{A}_1$	–	1.000	0.6857	0.6857	1.0000	1.0000	0.3429
$\mathbf{A}_2$	1.000	–	0.8857	0.6857	0.6857	0.8857	0.2000
$\mathbf{A}_3$	0.6857	0.8857	–	1.0000	0.3429	0.4857	0.2000
$\mathbf{A}_4$	0.6857	0.6857	1.0000	–	0.3429	0.8857	0.3429
$\mathbf{A}_5$	1.0000	0.6857	0.3429	0.3429	–	1.0000	0.4857
$\mathbf{A}_6$	1.0000	0.8857	0.4857	0.8857	1.0000	–	0.8857
$\mathbf{A}_7$	0.3429	0.2000	0.2000	0.3429	0.4857	0.8857	–

To evaluate the performance of an AV engine according to the malware type, Fig. 5 presents $\mathcal{P}_{ijk}$ (see (1)), the percentage of M_j malware detected by A_k AV engine in software samples present in a particular source S_i, which shows that the performance of an AV varies and depends on the malware type. For example, in HDD samples as shown in Fig. 5(a), AV engine A_1 detected approx. 51% less adware as compared to A_3. However, the same AV engine A_1 detected 200% more Trojans as compared to A_3. We observe a similar behavior in DVDs and downloads as well. Note that we made sure that the same malware was not classified as two different types by two AV engines by carefully analyzing the location, names and types returned by each engine and if such a sample was found, we placed it in the 'Other' category.

Thus, we can now answer our third question, i.e., "Can we determine the most sensitive antivirus engine based on its malware detection rate?". The answer to this is clear, from Fig. 4 we see that the A_3 is the most sensitive reporting the highest detection rate of 132% whereas A_7 is the least sensitive with a lowly 30% detection rate. We also observe that some AV engines performed better in one source (i.e., HDD, DVD, and downloaded software) as compared to other sources. For example, the detection percentage of Trojans by AV engine A_4 in HDDs is 94% lower than AV engine A_1.

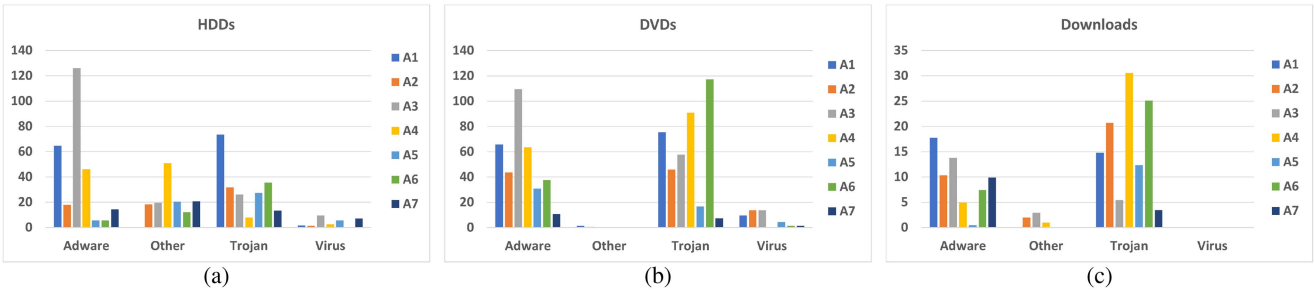

FIGURE 5. Percentage of malware detected by anti-virus engines in software samples found in the three sources, categorized by different types of malware.

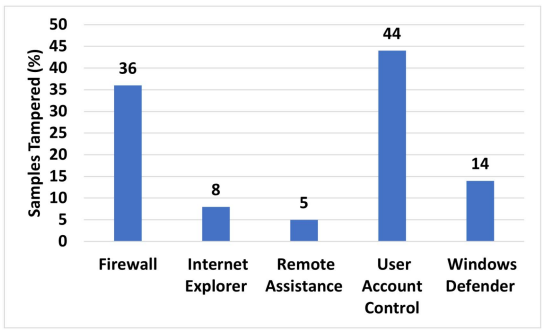

FIGURE 6. Proportion of HDDs with tampered OS settings.

However, the same AV engine A_4 detected 20% more Trojans than AV engine A_1 in DVDs. Similar trend is observed when comparing the Trojan detection percentage of A_2 and A_3 who had similar Trojan detection rates in HDD, but have different rates in case of DVDs and downloads.

C. MALICIOUS TAMPERING OF OS SETTINGS

In this section, we present and discuss our findings in terms of critical OS settings that are maliciously modified after installing pirated software. Fig. 6 shows the proportion of HDD samples whose OS settings were found to have been modified from the default settings. We observe that the most affected settings are firewall and UAC settings. This suggests that installing pirated software opens doors for coordinated attacks where an attacker may use one type of malware to compromise a computer and tamper with its critical OS settings. After that the attacker may exploit those settings to transfer more harmful malware to the victim computer. For example, this may be done by enabling the installation of malicious software without the need for permission from the user, i.e., by tampering with the UAC settings.

D. DISCUSSION

This study involves a sample size of 750 pirated software samples obtained from Southeast Asian countries, with each country's computer purchases limited to fewer than 10 units and an average of 30 DVD samples per country. It's essential to recognize that the results should be interpreted within a specific context and cannot be generalized to any particular nation due to these limitations. Additionally, our dataset doesn't provide a fully comprehensive sample from the population, introducing a limitation in the form of sampling bias towards certain countries. It's important to acknowledge that our analysis does not consider alternative procurement channels for personal computers, such as second-hand sales, or untapped sources of pirated software on the internet. Furthermore, we recognize that there might be unidentified factors at play in this type of research, as highlighted by previous work [32]. Nevertheless, research of this nature offers valuable insights into the various factors that influence malware prevalence and the intricate relationship between software piracy and the dissemination of malware. While our findings should be understood within the context of the study's limitations, they contribute to a broader understanding of these critical aspects.

To evaluate the impact on individual affected users, our methodology can be extended to perceive each source sample of pirated media (HDD, DVD, or download) as representative of a single user. This simplification is applicable to HDDs extracted from computers, as they typically belong to a single user. However, the situation becomes more complex for DVDs and downloads, as a user might acquire multiple DVDs or downloads as per their requirements. Within this context, as demonstrated in Fig. 3, our analysis suggests the probability of encountering malware upon purchasing a computer to be as high as 96%. Therefore, it is advisable to perform a fresh OS installation or conduct a comprehensive antivirus scan on newly acquired computers. Additionally, given the 117% infection rate in DVDs, it is recommended to scan the contents of a DVD package using antivirus software before installing any software to ensure its integrity. Here we emphasise that, once malware infiltrates a system, it poses a severe threat to the system's security, potentially compromising data integrity, confidentiality, and availability. Consequences of malware infections include data breaches, unauthorized access, and data theft, endangering sensitive information. Furthermore, malware can lead to system crashes, performance degradation, and network congestion, adversely affecting overall system functionality. Additionally, it can serve as a conduit for more advanced cyberattacks, making it a gateway to further threats.

Moreover, selecting appropriate AV solutions for system security is not a straightforward task. Our analysis indicates that different AV engines may exhibit varying levels of sensitivity. While an AV with low sensitivity might suffice for individual users, it could prove detrimental in corporate environments if it fails to detect malware. Therefore, the choice of AVs demands careful consideration, as it should align with the specific security needs and settings of the users or organizations. Furthermore, an intriguing avenue for future research could involve delving into a comparative analysis of what individual AVs successfully detect and what they may potentially overlook in comparison to their counterparts.

V. CONCLUSION

This paper presented an empirical study of prevalence of malware in pirated software from three different sources, i.e., HDDs of newly purchased computers, DVDs, and the Internet. 750 samples of pirated software found in the sources acquired from eight different countries in Southeast Asia were analyzed using seven different AV engines. The results show that adware and Trojans are the most prevalent types of malware in pirated software. Moreover, adware malware was more prevalent in HDDs while Trojans were more prevalent in DVDs. The prevalence of viruses in HDDs and DVDs was approximately the same. However, downloaded samples had the lowest prevalence of malware, even-though users may be exposed to malware during the download process. Our findings further underscore the substantial variation in detection sensitivity across various AV solutions. For some AVs, the average detection rates exceeded 100%, while one AV detected as few as 30% of malware infections. This stark contrast emphasizes the critical importance of users, be they individuals or corporations, meticulously evaluating the sensitivity of the chosen AV for system security. Opting for an AV with overly aggressive detection could lead to minor productivity loss due to false positives. Conversely, selecting an AV with low sensitivity risks overlooking malware threats, potentially resulting in severe security breaches and data damage.

REFERENCES

[1] A. Greenberg, "Crash override: The malware that took down a power grid," Jun. 12, 2017. Accessed: Jan. 2023. [Online]. Available: https://www.wired.com/story/crash-override-malware/

[2] "Every minute, 2,900,000 is lost to cybercrime" The evil internet minute, 2019. [Online]. Available: https://www.riskiq.com/resources/infographic/evil-internet-minute-2019/

[3] "Mcafee labs threats report," McAfee, San Jose, CA, USA, Tech. Rep., May 2021. [Online]. Available: https://www.mcafee.com/enterprise/en-us/assets/reports/rp-threats-jun-2021.pdf

[4] B. S. Alliance, "Software management: Security imperative, business opportunity," *BSA Glob. Softw. Surv.*, 2018. [Online]. Available: https://gss.bsa.org/wp-content/uploads/2018/05/2018_BSA_GSS_Report_en.pdf

[5] V. DeMarines, "Revulytics software piracy statistics," Sep. 2019. [Online]. Available: https://www.revulytics.com/blog/infographic-2019-revulytics-software-piracy-statistics.html

[6] "BSA calls out 'ghost piracy' in southeast asia," 2022. [Online]. Available: https://www.digitalnewsasia.com/business/bsa-calls-out-ghost-piracy-southeast-asia

[7] "Cybercrime costs global economy 2.9 m per minute" *InfoSecurity Mag.*, Jul. 2019. [Online]. Available: https://www.infosecurity-magazine.com/news/cybercrime-costs-global-economy/

[8] E. G., "Microsoft office flaws exploited in nearly 80% of malware attacks," Jul. 2022. [Online]. Available: https://atlasvpn.com/blog/microsoft-office-flaws-exploited-in-nearly-80-of-malware-attacks

[9] M. N. Aman and B. Sikdar, "ATT-Auth: A hybrid protocol for industrial IoT attestation with authentication," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5119–5131, Dec. 2018.

[10] M. N. Aman et al., "HAtt: Hybrid remote attestation for the Internet of Things with high availability," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7220–7233, Aug. 2020.

[11] L. Zeltser, "Analyzing malicious software in cyberforensics," J. Bayuk, Ed., 2010.

[12] J. Bayuk et al., "Malware risks and mitigation report," BITS Financial Serv. Roundtable, Washington, DC, USA, Tech. Rep., 2011. [Online]. Available: https://www.nist.gov/system/files/documents/itl/BITS-Malware-Report-Jun2011.pdf

[13] R. A. Grimes, "9 types of malware and how to recognize them," Nov. 2020. [Online]. Available: https://www.csoonline.com/article/2615925/security-your-quick-guide-to-malware-types.html

[14] M. Kammerstetter, C. Platzer, and G. Wondracek, "Vanity, cracks and malware: Insights into the anti-copy protection ecosystem," in *Proc. ACM Conf. Comput. Commun. Secur.*, 2012, pp. 809–820.

[15] F. Lalonde Levesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji, "A clinical study of risk factors related to malware infections," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2013, pp. 97–108.

[16] G. Maier, A. Feldmann, V. Paxson, R. Sommer, and M. Vallentin, "An assessment of overt malicious activity manifest in residential networks," in *Proc. Int. Conf. Detection Intrusions Malware Vulnerability Assessment*, 2011, pp. 144–163.

[17] D. Canali, L. Bilge, and D. Balzarotti, "On the effectiveness of risk prediction based on users browsing behavior," in *Proc. 9th ACM Symp. Inf., Comput. Commun. Secur.*, 2014, pp. 171–182.

[18] M. Botacin, G. Bertão, P. de Geus, A. Grégio, C. Kruegel, and G. Vigna, "On the security of application installers and online software repositories," in *Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment*, 2020, pp. 192–214.

[19] K. Onarlioglu, U. O. Yilmaz, E. Kirda, and D. Balzarotti, "Insights into user behavior in dealing with internet attacks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2012. [Online]. Available: https://www.ndss-symposium.org/ndss2012/ndss-2012-programme/insights-user-behavior-dealing-internet-attacks/

[20] P. A. Watters, M. Watters, and J. Ziegler, "Malicious advertising and music piracy: A New Zealand case study," in *Proc. IEEE 5th Cybercrime Trustworthy Comput. Conf.*, 2014, pp. 22–29.

[21] P. Watters, "Consumer risks from piracy sites in the Philippines," Aug. 1, 2023, doi: 10.2139/ssrn.4536945.

[22] P. Watters, "Time to compromise: How cyber criminals use ads to compromise devices through piracy websites and apps," Dec. 2021, doi: 10.2139/ssrn.4536943.

[23] J. Haffejee and B. Irwin, "Testing antivirus engines to determine their effectiveness as a security layer," in *Proc. IEEE Inf. Secur. South Afr.*, 2014, pp. 1–6.

[24] Sonicwall, "SonicWall cyber threat report," 2020. [Online]. Available: https://www.sonicwall.com/medialibrary/en/white-paper/2020-sonicwallcyber-threat-report.pdf

[25] S. Kumar, L. Madhavan, M. Nagappan, and B. Sikdar, "Malware in pirated software: Case study of malware encounters in personal computers," in *Proc. IEEE 11th Int. Conf. Availability, Rel. Secur., Conf. Proc.*, pp. 423–427.

[26] G. Mezzour, K. M. Carley, and L. R. Carley, "An empirical study of global malware encounters," in *Proc. Symp. Bootcamp Sci. Secur.*, 2015, pp. 1–11.

[27] C. Colby, R. Hodge, A. D. Rayome, and T. Attila, "Best antivirus software for 2023," Sep. 2023. [Online]. Available: https://www.cnet.com/tech/services-and-software/best-antivirus/

[28] A. Spadafora, "The best antivirus software 2023: Free and paid options," Oct. 2023. [Online]. Available: https://www.tomsguide.com/us/best-antivirus,review-2588.html

[29] H. Mann and D. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 3 1947.

[30] "Bitdefender malware and spam survey finds e threats adapting to online behavioral trends," Aug. 2009. [Online]. Available: https://www.bitdefender.com/news/bitdefender-malware-and-spam-survey-finds-e-threats-adapting-to-online-behavioral-trends-1094.html
[31] R. Roul, "49 malware statistics businesses should take seriously," Sep. 2021. [Online]. Available: https://www.g2.com/articles/49-malware-statistics-businesses-should-take-seriously
[32] T.-F. Yen, V. Heorhiadi, A. Oprea, M. K. Reiter, and A. Juels, "An epidemiological study of malware encounters in a large enterprise," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1117–1130.

ASIF IQBAL (Member, IEEE) received the B.S. degree in telecommunication engineering from NUCES-FAST, Peshawar, Pakistan, in 2008, the M.S. degree in wireless communications from LTH, Lunds University, Lund, Sweden, in 2011, and the Ph.D. degree in electrical and electronics engineering from The University of Melbourne, Melbourne, VIC, Australia, in 2019. He is currently a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He was an Assistant Professor with the Faculty of National University of Computer and Emerging Sciences, Pakistan. His research interests include signal processing, deep learning, sparse signal representations, and privacy preserving machine learning.

MUHAMMAD NAVEED AMAN (Senior Member, IEEE) received the B.Sc. degree in computer systems engineering from KPK UET, Peshawar, Pakistan, in 2006, the M.Sc. degree in computer engineering from the Center for Advanced Studies in Engineering, Islamabad, Pakistan, in 2008, and the M.Eng. degree in industrial and management engineering and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2012, respectively. He is currently an Assistant Professor with the University of Nebraska-Lincoln, Lincoln, NE, USA. His research interests include IoT and network security, hardware systems security and privacy, wireless and mobile networks, and stochastic modeling.

RAMKUMAR REJENDRAN received the M.Sc. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2018. His research focuses on malware analysis.

BIPLAB SIKDAR (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from North Eastern Hill University, Shillong, India, in 1996, the M.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1998, and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2001. From 2001 to 2013, he was with the Faculty of Rensselaer Polytechnic Institute, first as an Assistant and then as an Associate Professor. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include wireless network, and security for IoT and cyber physical systems.