A Comparison of Systemic and Systematic Risks of Malware Encounters in Consumer and Enterprise Environments

Savino Dambra, Leyla Bilge, Davide Balzarotti

A Comparison of Systemic and Systematic Risks of Malware Encounters in Consumer and Enterprise Environments

SAVINO DAMBRA, Eurecom, France
LEYLA BILGE, Norton Research Group, France
DAVIDE BALZAROTTI, Eurecom, France

Malware is still a widespread problem and it is used by malicious actors to routinely compromise the security of computer systems. Consumers typically rely on a single AV product to detect and block possible malware infections, while corporations often install multiple security products, activate several layers of defenses, and establish security policies among employees. However, if a better security posture should lower the risk of malware infections, the actual extent to which this happens is still under debate by risk analysis experts. Moreover, the difference in risks encountered by consumers and enterprises has never been empirically studied by using real-world data.

In fact, the mere use of third-party software, network services, and the interconnected nature of our society necessarily exposes both classes of users to undiversifiable risks: independently from how careful users are and how well they manage their cyber hygiene, a portion of that risk would simply exist because of the fact of using a computer, sharing the same networks, and running the same software.

In this work, we shed light on both systemic (i.e., diversifiable and dependent on the security posture) and systematic (i.e., undiversifiable and independent of the cyber hygiene) risk classes. Leveraging the telemetry data of a popular security company, we compare, in the first part of our study, the effects that different security measures have on malware encounter risks in consumer and enterprise environments. In the second part, we conduct exploratory research on systematic risk, investigate the quality of nine different indicators we were able to extract from our telemetry, and provide, for the first time, quantitative indicators of their predictive power.

Our results show that even if consumers have a slightly lower encounter rate than enterprises (9.8% vs 12.0%), the latter do considerably better when selecting machines with an increasingly higher uptime (89% vs 53%). The two segments also diverge when we separately consider the presence of Adware and Potentially Unwanted Applications (PUA), and the generic samples detected through behavioral signatures: while consumers have an encounter rate for Adware and PUA that is 6 times higher than enterprise machines, those on average match behavioral signatures two times more frequently than the counterpart. We find, instead, similar trends when analyzing the age of encountered signatures, and the prevalence of different classes of traditional malware (such as Ransomware and Cryptominers). Finally, our findings show that the amount of time a host is active, the volume of files generated on the machine, the number and reputation of vendors of the installed applications, the host geographical location and its recurrent infected state carry useful information as indicators of systematic risk of malware encounters. Activity days and hours have a higher influence in the risk of consumers, increasing the odds of encountering malware of 4.51 and 2.65 times. In addition, we measure that the volume of files generated on the host represents a reliable indicator, especially when considering Adware. We further report that the likelihood of encountering Worms and Adware is much higher (on average 8 times in consumers and enterprises) for those machines that already reported this kind of signatures in the past.

CCS Concepts: • **Security and privacy** → *Malware and its mitigation*; • **Software and its engineering** → *Risk management*.

Authors' addresses: Savino Dambra, savino.dambra@eurecom.fr, Eurecom, Biot, France; Leyla Bilge, leyla.bilge@nortonlifelock.com, Norton Research Group, Biot, France; Davide Balzarotti, davide.balzarotti@eurecom.fr, Eurecom, Biot, France.

2471-2566/2022/10-ART $15.00
https://doi.org/10.1145/3565362

Additional Key Words and Phrases: Systemic risk, systematic risk, cyber-risk assessment, consumer malware, enterprise malware

1 INTRODUCTION

Recent statistics about cyber attacks [17, 18, 30, 38] show that malware is still a widespread phenomenon and malicious programs created to disrupt and damage computer systems are constantly on the rise, despite the improvements in anti-malware measures and the growth in cyber-security investments [37, 40]. As a result, malicious software remains one of the most common threats for both consumer and enterprise machines, although the hosts in the two groups show substantial differences in their purpose, installed software, configurations, and in the number and sophistication of their security products.

The way in which consumers and enterprises approach security is very different: while the former follow a reactive approach, installing defenses (typically in the form of AV software) to detect and remove possible malware infections, companies are expected to work more proactively, by relying on articulated risk assessment, mitigation, and risk transfer methodologies [9]. It is also well-known that consumers invest less in security, often preferring off-the-shelf, easy-to-use solutions that offer few customization options. On the contrary, organizations tend to protect their assets and data by deploying complex and multi-faced solutions that rely on several layers of defenses, such as firewalls and security proxies, intrusion detection and prevention systems, email protection and anti-exfiltration software, together with measures to prevent insider attacks and to limit the spread of infections. Consumers and enterprises also differ from a user point of view. In fact, in addition to educating employees on the best security practices, enterprises may adopt stricter security policies about what software can be installed—thus preventing users from running software of dubious origin that is often a vehicle for malware. For end-users, this choice is left to the sole user's security awareness and knowledge.

Each of these factors may affect the risk of experiencing cyber incidents and malware infections. This risk, known as *systemic*, is strictly related to the individual security posture and to the adopted counter-measures. However, there are also other factors to be accounted for, as both consumer and enterprise machines are not isolated entities. The interconnected nature of our society brings companies to rely on external services and to outsource computational tasks to third-party subjects, thus exposing the hosts of both parts to a potential *systematic* risk — which is a form of undiversifiable risk that is independent of how much a subject spends in security products and from its cyber hygiene.

While these two portions of risk might seem disconnected as they depend upon different factors, their joint analysis and comprehension are fundamental to carry out an exhaustive and holistic risk assessment. A methodology that tries to measure the risk by focusing only on one risk class might overlook fundamental indicators and provide as a result an inaccurate estimation that can be dangerous for both users and business decision makers. The financial sector is one of the best examples where risk estimations always rely on both kinds of risk. Indeed the two terms, systematic and systemic, are commonly used in the financial sector. In particular, systematic risks (also known as undiversifiable, volatility, or market risk) refer to the risk inherent to the entire market, that is not specific to a particular stock or industry and that therefore is impossible to completely avoid and cannot be mitigated through investment diversification. On the contrary, systemic risk (also known as nonsystematic, specific, or residual risk), is unique to a specific company, industry, or market segment. This second type of risk can be reduced by simply redirecting the investment towards multiple companies, stocks, and markets related to different sectors, thus reducing the likelihood that a failure in one of them could influence the others. The example related to the financial sector makes it clear that it is important to account for both risk classes when performing risk assessment. In the same way, the existence of different risk portions must be considered when evaluating the risks related to security events, in our case represented by malware encounters.

In this work, we want to study how these two forms of risks are applicable in the cyber domain and how they affect the consumer and enterprise environments. Cyber risk estimation is a very complex and challenging

problem, that to date was mainly approached from a *qualitative* perspective [9]. In this paper, we aim instead at exploring *quantitative* metrics, obtained by leveraging empirical data. In fact, while it might sound obvious to the reader that some factors are correlated to higher security risks, (e.g., the fact that machines with higher activity are more likely to encounter malware, or that a broader and more diverse set of software results in higher attack surface), the exact relationships that these variables have with the risk of encountering malicious software has never been measured before. In addition, our study provides numerous insights on the different impact these factors have on consumers and enterprises environments.

It is important to stress that the cyber security risks of consumers and enterprises have never been compared before. Although it is possible to infer some differences and similarities by looking at studies that analyzed either the first or the second segment in isolation, those studies relied on different data sources, focused on different aspects, and were performed over disjoint timeframes, thus making it difficult to compare results. On the contrary, the internal telemetry information we use for our experiments comes from a single AV sensor, and it has been collected in the same time period. In addition, the AV endpoint is dispatched with two different licence schemes that allow us to clearly distinguish between corporate and consumer machines. Moreover, while risk assessment is one of the cornerstones of computer security, the difference among consumer vs enterprise security has never been experimentally measured before: do enterprise machines encounter less malware because they are protected by more and more diverse cyber defenses? Are enterprise users more security conscious and, therefore, less likely to visit risky websites at work? Is there some relevant difference among the malicious files encountered by end-users and large companies employees?

To answer these questions, we articulate our work in two parts: 1) an extensive analysis of common aspects and differences in malware encounters between the two segments, and 2) an exploratory investigation of systematic risk indicators. To assess the implications of different choices in security investments and policies, the first part of the paper quantifies the malware encounter rate in consumers and enterprises and provides evidence of the most common classes and signatures observed by the two parties. We also look at the reporting frequency and different labels of popular malware families, the different incidence that PUA and Adware, and the impact that behavioral signatures have on corporate and consumer hosts.

In the second part of the paper, we conduct an exploratory study of undiversifiable risk indicators that we were able to extract from our real-world telemetry. For instance, we assess whether the days and hours of activity together with the volume of host-generated files can serve this purpose. We also look at the effect that the number of installed software vendors has on the malware encounter rate. We assess whether being in a recidivist infected state can be a good risk predictor, and finally, we verify whether the size of an enterprise or its industrial sector can provide useful insights on the systematic risk the company encounters.

We hope that our large-scale measurement and the quantitative insights provided in this work can help risk analysis experts to better understand the role of systemic and systematic risks and their impact on consumers and enterprise environments. We also believe that our study can provide valuable input for researchers working on malware detection and reputation systems, as well as for those interested in the cyber-insurance area.

2 RELATED WORKS

To the best of our knowledge, no scientific work exists that has specifically focused on the differences between consumers and enterprises when it comes to the cyber-threat landscape and the risk of malware infections. In the following sections, we discuss two different research areas that relate to ours: at first, we look at previous studies that have explored the threat landscape of either consumer or enterprise machines. In the second part, we cover those works that have correlated indicators extracted from telemetry data to the risk of cyber incidents and malicious software encounters.

2.1 Threat landscape

Many industrial reports published by security companies [5, 13, 14, 38] provide an annual summary of the malware families observed in the wild, the number of compromised machines, the losses due to data breaches, and the malicious campaigns that targeted different organizations. However, those documents only focus on statistics, do not carry a detailed analysis, and do not make any distinction between consumer and enterprise environments.

In the scientific community, scattered studies leveraged network telemetry or internal logs provided by ad-hoc software to delineate the status and the evolution of the malware landscape. Kotzias et al. [16] analyzed a 3-year-long collection of internal data from 28K enterprises to shed light on their vulnerability patching behaviors and existing threats. The investigation carried out by the authors shows a higher prevalence of malware with respect to potentially unwanted programs (PUP), the presence of more secure and affected industrial sectors, and the fact that the patching of server applications is much worse than the one on the client-side.

Two studies focused on the trends of malware that spreads through Pay-Per-Install (PPI) Services [2, 15]. Caballero et al. [2] built an infrastructure and deployed it in 15 countries to interact with 4 PPI providers. The authors found that 12 out of 20 of the most prevalent families of malware employ PPI services and that this distribution mechanism is more common in richer countries. The follow-up paper narrowed the analysis down to PUP families that spread through PPI services, performing a systematic study of their prevalence using AV telemetry. The results indicate that PUPs are installed on 54% of the considered machines and that up to 25% of them are distributed by a limited number of publishers.

2.2 Risk indicators

In recent years, an increasing number of studies have tried to identify risk indicators i.e., measurable features collected from external sources or internal telemetry, that can be correlated with the risk of suffering from cyber incidents. Some of them also applied the features they identified to train prediction algorithms and assess the prevalence of those risks in the future.

Yen et al. [42] used internal telemetry logs of a large organization to spot risk indicators that are correlated to malware encounters. The authors showed that user's demographic features, as age and job title, together with network-related features, such as the frequent use of untrusted internet connections and longer browsing sessions, are effective at predicting which users are more at risk of malware infections. RiskTeller [1] is a prediction tool that leveraged internal data of 18 enterprises to predict which of their machines will be at risk of being infected by a broad spectrum of malware classes. Its classification accuracy reaches 95%, showing that the identified features are strongly correlated with the likelihood of malware encounters. Liu et al. [22] studied the extent to which cyber security incidents can be predicted by using observed malicious activities associated with network entities, such as spamming, phishing, and scanning. The study shows that the resulting classifier is able to produce fairly accurate predictions over a forecasting window of 2-3 months. The same authors also attempted to predict the likelihood of organizations to suffer a cyber incident by using an algorithm that only uses externally observable features [21]. The authors trained a classifier by combining signs of network mismanagement, such as misconfigured DNS or BGP, with malicious activity time series, such as spam, phishing, and scanning activity sourced from these organizations. Despite 10% of false positives, the prediction reaches 90% accuracy, suggesting the possibility of forecasting an organization's breach without internal information. Thonnard et al. [39], discussed organization and individual-level features that are likely to reflect the risk of experiencing targeted attacks. The authors identify enterprise sizes and public profiles of individuals as potential risk factors and show that there exists a degree of correlation with receipt of targeted attacks. In a similar way, Sarabi et al. [33] built a predictor using a set of industry, business and web visibility/population information. The results demonstrate how, and to what

extent, these externally-observable features can help forecast an enterprise's relative risk of experiencing different types of cyber incidents.

Fewer prediction studies exist on the consumer side, probably due to the lack of telemetry data for this segment of users. Lévesque et al. [20], performed a 4-month study by collecting real-usage data of 50 subjects and monitoring both user behaviors and possible infections. Using neural networks, the authors developed a predictive model with 80% accuracy at predicting the users' likelihood of being infected. Canali et al., [4] assessed to what extent a user's web browsing behaviors can be used to predict her risk class. The results show how particular types of user actions, such as browsing the web late at night and during weekends, considerably affect the risk exposure. Finally, by leveraging mobile users' browsing patterns and self-reported data, Sharif et al. [35] tried to predict whether users will encounter malicious pages on a long and short term. With an overall accuracy of 87% TPR and 20% FPR, this work shows how useful on-the-fly predictions can be in protecting users from malware distributed on the web.

3 DATASETS

This section provides a detailed description of the different data sources we used in this study, as summarized in Table 1. Our main source of information is the telemetry data of a popular security company, collected on Windows machines throughout the year 2018 and made of different feeds. *Activity* reports provided a starting point to list all machines that had the antivirus sensor installed and opted in to share their data, allowing us to compute the number of hours each machine was active every day. *File appearance logs* helped us to identify vendors of installed programs. Using *malware encounters logs* we identify where, how many times, and which signatures were triggered for each malware encounter. Finally, we scraped the company website to retrieve a list of all existing *signatures* along with their class and description.

The datasets were encoded in pickle files by using Pandas, a Python open-source data-analysis and manipulation tool [29]. Statistics and results were computed by using NumPy, which offers a comprehensive set of mathematical functions [28]. When training and testing the models of Section 5, we used Scikit-learn, a library that offers efficient tools for predictive data analysis [34]. All the figures have been rendered by using Matplotlib [23].

Table 1. Overview of datasets used

		Unique instances	
Dataset	Info About	Consumers	Enterprises
Activity	Hosts	144.9 M	226.4 M
	Enterprises		45.6 - 640 K
	Countries	239	235
File appearance	Vendors	59.9 K	40.9 K
Encounters	Hosts	14.2 M	27.1 M
	Enterprises		26.6 - 244.2 K
	Records	62.4 M	76.5 M
	Signatures	24.0 K	23.3 K
	Countries	239	235
Signatures	Labels	32.0 K	
	Subclasses	41	
Industrial sectors	Sectors		10 - 1215

3.1 Consumers vs Enterprises

Our data contains 640 K unique enterprise identifiers. However, since big corporations can span multiple countries and comprise several subsidiaries—each of which may possess a different identifier— we use a second mapping to further group those cases to a single organization. In total, we were able to identify 45.6 K (2nd record in Table 1) unique organizations. We distinguish 6.5 K micro ($\leq$ 10 hosts), 12.3 K small ($\leq$ 50 hosts), 11.9 K medium ($\leq$ 250 hosts) and 14.8 K large enterprises ($>$ 250 hosts), with the biggest of them having 3.4 M machines.

In the period of our experiments, we observed a total of 144.9 M distinct consumer machines and 226.4 M enterprise machines. Our dataset covers 239 (for consumers) and 235 (for enterprises) of the 249 countries, territories or areas of geographical interest with an assigned ISO 3166-1 code [12]. The two tables below (grouped under Table 2) report the geographical breakdown of the machines in our dataset: North America is the most represented region (38% of the machines), followed by Europe (27%) and Asia (22%). In South America, Africa and Oceania we measure the lowest concentrations ($<$ 10 % overall).

Table 2. Host distribution per countries (left) and continents (right)

Consumers		Enterprises	
Country	% hosts	Country	% hosts
United States	33.87	United States	35.52
Japan	7.46	India	6.60
Germany	5.41	China	4.51
United Kingdom	4.60	Brazil	3.39
China	3.74	Japan	3.12
Brazil	3.52	United Kingdom	3.02
Canada	3.45	Germany	2.22
France	3.25	France	2.10
Australia	3.07	Canada	1.90
India	2.61	Australia	1.55
Italy	1.98	Mexico	1.52
Others	27.04	Others	34.55

Consumers		Enterprises	
Continent	% hosts	Continent	% hosts
North America	38.89	North America	42.55
Europe	27.57	Asia	27.18
Asia	22.32	Europe	19.76
South America	5.49	South America	5.80
Oceania	3.49	Africa	2.57
Africa	2.24	Oceania	2.13

3.2 Host activity and file appearance

All the 371 M machines in our dataset have an anonymized identifier linked to the AV software licence and thus stable throughout the period under analysis. Each of them routinely queries a centralized system to assess the reputation of files that appear on the host. These requests are made possible thanks to the explicit consent of both consumer and enterprise users, who opted-in to share their data in an anonymized and privacy-preserving form. We leverage this process for two different purposes. First, for each machine and for each day in the time frame of this study, we computed the number of active hours. We then computed the number of active days per month by counting the days in which the machine submitted at least one request. On average, consumer and enterprise hosts are active 6.4 and 7.6 days per month, respectively for 2.9 and 3.7 hours per day. Second, for all executed applications we extract the vendor name (if the file is signed), thus identifying more than 40 K distinct vendor names for enterprises and around 60 K for consumers.

3.3 Malware Encounters

When a file is flagged as malicious by the host AV sensor, the event (including the hash and the signature identifier) is reported to the central server. We use these logs to create a register that, for each machine, records the day, the number of encounters (as the same object can be reported multiple times), and the matching signature name. Our data do not allow us to perform a retroactive analysis of files to catch newly identified threats, but only consider those reported by existing signatures at the time of detection. In addition, we filter out all the signatures that were

matched but that do not appear anymore in the vendor's list of signatures at the time of the study: the rationale behind this choice is that we want to limit the impact of wrong signatures and remove those generating false positives. Over the 140 M collected events, we identified 14.2 M distinct consumers and 27.1 M distinct enterprise hosts that encounter at least one malicious file within the year. Overall, malware was encountered by 58.3% of the enterprise.

We scraped the website of the AV vendor to obtain the list of available signatures—together with their descriptions, years of creation, and subclasses. In this way, we were able to gather information about 18143 labels classified in 41 subclasses (out of roughly 24 K signatures observed in the dataset). For a more concise classification, we decided to merge similar and smaller subclasses into seven broader groups: *Adware*, *PUA*, *Trojan*, *Ransomware*, *Worms*, *Viruses*, and *Others*. The full mapping among the different classes is reported in Table 3.

Table 3. Malware classes grouping

Class	Subclass
Adware	Adware
	Adware-trojan
Potentiallyunwantedapp (PUA)	Misleadingapplication
	Misleadingapplication-trojan
	Potentiallyunwantedapp
Ransom	Ransom
Trojan	Trojanhorse
	Trojanhorse-macro
	Trojanhorse-virus
	Trojanhorse-worm
	Trojanhorse-worm-macro
	Trojanhorse-worm-virus
	Trojan-virus
	Trojan-worm
Virus	Virus
	Virus-macro
Worm	Worm
	Worm-macro
	Worm-virus

Class	Subclass
Others	Dialer
	Dialer-adware
	Dialer-hacktool
	Dialer-trojan
	Hacktool
	Hoax
	Joke
	Joke-trojan
	Macro
	Other
	Other-trojan
	Other-worm
	Parentalcontrol
	Remoteaccess
	Removalinformation
	Securityassessmenttool
	Securityassessmenttool-trojan
	Spyware
	Spyware-trojan
	Trackware
	Trackware-trojan

3.4 Enterprise industry sectors

For a subset of the anonymized enterprise identifiers, we were provided with a number of additional information; including their industry sectors and the countries in which their registered offices are based. This industry classification is available in different granularities, ranging from a fine-grained classification of up to 1215 distinct sectors to a coarse version of only 10 macro-sectors. Table 4 shows the number of machines and enterprises per sector, according to the most concise classification: information technology is the prevalent industry with more than 3 M hosts and 4732 enterprises. Globally, our dataset shows good industry coverage, with all sectors having at least 200 K active machines, and half of the sectors having more than 1 M hosts.

Table 4. General sector statistics

Sector	Enterprises	Hosts
Consumer Discretionary	5030	1.99 M
Consumer Staples	1495	912.22 K
Energy	654	210.71 K
Financials	5052	2.96 M
Healthcare	2349	1.96 M
Industrials	7715	2.79 M
Information Technology	4732	3.63 M
Materials	2159	427.00 K
Telecommunication Services	314	307.08 K
Utilities	496	245.59 K

3.5 Ethical considerations

The datasets we analyzed in this work derive from logs and data collected only from consumer and enterprise users who voluntarily opted-in to share their data. This choice is left to users at installation time, when they are presented with information about the data collection mechanism and a checkbox to tick if they wish to opt-in. Specifically, each piece of information is anonymized on the client-side and sent in this form to a central system, to preserve the customers' privacy and identity. In our study, we observe enterprises and hosts only through alphanumeric anonymized identifiers that do not contain any detail or endpoint attribute able to trace back to their origin. The data analyzed in this study, although it might come from different sources, is similar in nature to what has already been observed in other studies, such as the one of Yen et al. [42], Kotzias et al. [16] and Dambra et al. [10].

3.6 Selection Bias and Limitations

The dataset we used for our study is the largest ever adopted for risk-based experiments: while the telemetry of previous works included at most 20 K consumer devices [35], and 82M machines of 28k enterprises [16], the one used in this work has been collected on more than 226M organization hosts and 144M home-user computers located in almost 250 countries. However, it is not completely unbiased. For instance, we only analyze consumers and enterprises that invest in security products: it is reasonable to believe that those without any protection should have a worse security posture, thus making our results conservative. Moreover, our datasets are obtained from a single vendor and only from those users who opted-in to share data: although this allows us to better compare the two classes of machines, software from other vendors may provide different security, and users who opted-out due to privacy concerns could be more security conscious. In addition, our telemetry is only collected on Windows hosts. Although Windows is still by far the most adopted operating system with 75% of the market share [36], it is possible that users running other OSes (e.g., macOS, Unix-like) may have a different security posture. Finally, the prevalence of Windows machines can be different between the two segments of machines. Nevertheless, Windows is still the predominant OS in both consumers and enterprises when it comes to common activities carried by users, such as gaming, document editing and other office tasks. On the other hand, the remaining OSes are often used for very specific purposes and installed on hosts that carry no human-interactive task, such as servers and machines hosting public-facing services [41].

Table 5. Most common malware signatures and classes for consumers and enterprises. For each malware class, percentages represent a normalization to the total number of hosts and organizations that encounter malware. Malware classes are sorted by the percentage of hosts on which they are detected.

Consumers			
Class/Label	Hosts	Reported Events	Labels
Trojan	11.3M (79.5%)	186.7 M	3.4 K
W97M.Downloader	627.3 K	3.1 M	
Mdropper	305.8 K	1.1 M	
Dromedan	303.6 K	916.8 K	
PUA	6.3M (44.4%)	32.6 M	747
InstallCore	699.2 K	1.0 M	
DownloadSponsor	509.3 K	1.6 M	
OpenCandy	335.4 K	438.6 K	
Others	4.6M (32.4%)	12.6 M	820
Jswebcoin	148.5 K	669.8 K	
Remacc.Ammyy	101.2 K	155.5 K	
Remacc.Radmin	10.9 K	18.9 K	
Adware	770.9K (5.4%)	2.4 M	491
Browext	154.0 K	623.3 K	
DealPly	54.3 K	87.6 K	
DriverUpdater	48.2 K	56.0 K	
Worm	559.1K (3.9%)	4.1 M	1.1 K
Silly	125.3 K	353.8 K	
Ippedo	64.7 K	206.8 K	
Dunihi	53.4 K	1.5 M	
Virus	279.5K (2.0%)	15.1 M	589
Sality	56.1 K	2.1 M	
Virut	47.8 K	493.1 K	
Bursted	34.5 K	154.8 K	
Ransom	112.0K (0.8%)	416.1 K	326
Wannacry	51.4 K	299.8 K	
Crysis	15.1 K	26.7 K	
Cerber	7.5 K	10.3 K	

Enterprises				
Class/Label	Enterprises	Hosts	Reported Events	Labels
Trojan	16.1K (60.5%)	22.7M (83.8%)	217.1M	3.2 K
Dromedan	2.6 K	481.7 K	1.4 M	
W97M.Downloader	4.0 K	179.3 K	603.5 K	
JS.Downloader	1.5 K	98.9 K	187.7 K	
Others	10.9K (41.0%)	2.3M (8.5%)	7.1 M	616
Remacc.Ammyy	985	74.6 K	115.2 K	
Jswebcoin	1.6 K	70.2 K	286.0 K	
Remacc.Radmin	172	26.5 K	42.2 K	
PUA	10.5K (39.5%)	1.9M (7.0%)	3.8 M	548
InstallCore	3.6 K	245.6 K	307.9 K	
OpenCandy	3.0 K	186.7 K	231.2 K	
DriverPack	1.1 K	105.1 K	149.5 K	
Worm	4.1K (15.4%)	692.1K (2.6%)	5.1 M	884
Silly	1.8 K	164.5 K	438.5 K	
Ippedo	1.0 K	83.5 K	325.2 K	
Dunihi	1.1 K	68.5 K	2.0 M	
Virus	2.6K (9.8%)	320.9K (1.2%)	17.6 M	396
Sality	1.1 K	74.8 K	3.2 M	
Virut	933	59.2 K	739.5 K	
Bursted	639	52.8 K	232.4 K	
Ransom	1.2K (4.5%)	160.6K (0.6%)	665.8 K	307
Wannacry	550	109.2 K	546.3 K	
Crysis	210	21.5 K	37.2 K	
Locky	31	4.0 K	7.8 K	
Adware	2.8K (10.5%)	149.5K (0.5%)	444.6 K	429
Browext	1.0 K	30.5 K	121.0 K	
Lop	339	16.4 K	20.5 K	
Funshion	153	6.9 K	15.5 K	

4 MALWARE SPECIFICITY

In this section, we describe similarities and differences in malware encounters among consumer and enterprise hosts. We start by analyzing the overall picture of encountered malware signatures and classes in section 4.1. Considerations about the number of malware classes on each host and the average age of signatures follow in sections 4.2 and 4.3. In section 4.4, we finally discuss how behavioral signatures, Potentially Unwanted Applications, and Adware impact consumers and enterprises in a different way.

4.1 Overall picture

We start our analysis by measuring malware encounter prevalence in consumers and enterprises. Over the twelve months observation period at our disposal, we found that the percentage of hosts that encounter malware slightly differs between the two groups: for consumers, 14.2 M of the 144.9 M active hosts have suffered at least one encounter (9.80%), while in enterprises 27.1 M out of 226.0 M machines (12.0%) detected malicious software. We verified that this difference is statistically significant ($p<.001$) by running a Chi-squared test on a 2-by-2 contingency table obtained by considering infected and clean devices in consumers and enterprises.

It is worth noting that the malware encounter rate we measured in enterprise environments is consistent with prior works. In fact, in their conservative estimation along three years (from 2015 to 2017), Kotzias et al. [16] report a prevalence rate of 13%; the same ratio increases to 15% in the study of Yen et al. [42], who consider hosts of a large enterprise in a four-month time frame in 2013. This shows that, once averaged over a sufficient number

of computers, the malware encounter rate in enterprises remained relatively constant across different studies, AV vendors, and even across multiple years.

No prior study exists instead that specifically focuses on consumer hosts encompassing every class of malware. Some measured a combined encounter rate —therefore also including enterprise machines— on a global scale [14], others restricted their analysis to only few malware classes to investigate their distribution vectors [2, 15, 27]. Although in the report published by Microsoft [14] there is no clear distinction between consumer and enterprise machines, our study reveals an encounter rate that is higher than the 6% assessed by their researchers in the security bulletin over the same period.

While the overall encounter rates are similar, a closer look at the malware families shows that there are some relevant differences between consumer and enterprise encounters. Table 5 summarizes the most common malware signatures and their corresponding classes in our telemetry data, together with percentages that represent a normalization to the fraction of devices and organizations that encounter malware. Labels are sorted by the number of distinct hosts in which they appeared, after removing generic records and those for which we could not assign a class (as explained in Section 3.3). As a single signature could be triggered multiple times in the same machine, we also measure and report these occurrences. We complete the picture by counting the total number of distinct labels for each class and the number of enterprises in which each signature has been observed.

Results show that Trojan is by far the most popular class: these signatures alone represent 47% of total number of signatures matched for consumers and nearly 80% for enterprises. Although this malware class is also prevalent in organization environments, home users show higher infection frequency and a more diverse set of labels: on average, consumer hosts report Trojan detection events 16.46 times during the year and encounter 2.02 distinct families in the same period. Enterprise frequency and distinct labels are lower (respectively 9.56 and 1.33). Again, the differences between the two means are statistically significant (Reporting frequency: Welch's ANOVA F-test = 5104, $p<.001$; Families: Welch's ANOVA F-test = 1709257, $p<.001$). The most common families are respectively *W97M.Downloader*, a well-known set of malicious macros embedded in Microsoft Word document files, and *Dromedan*, a label associated with a Trojan family spread via email attachments.

Table 5 also highlights the completely different incidence of PUA and Adware between the two groups. Although *InstallCore* —a large family of bundlers that install Adware and Potentially Unwanted Programs (PUPs)— and *Browext* — malicious software that shows advertisement and slows down the system to frustrate the user— are the most observed labels on both sides, PUA and Adware account upwards 29% for consumers, but not more than 7.1% for enterprises. In addition, home users report Adware and PUA detections on average 5.18 times per year, while enterprise machines only 2.05 times (Welch's ANOVA F-test = 649, $p<.001$). Since this is an important difference between the two groups we decided to dedicate Section 4.4 to investigate it in more detail.

On the contrary, Viruses and Worms (respectively 1.1% and 1.2% of all the signatures matched) appear with similar frequency in both groups. Although we register a statistically significant difference in the mere detection rate between the two segments of machines (Virus: X-squared = 12447, $p<.001$; Worm: X-squared = 14164, $p<.001$), we find no such difference when considering the reporting-event frequency and distinct-label encounters: Viruses are respectively detected on average 53.92 and 54.99 times during the year on home-user and organization machines (Welch's ANOVA F-test = 0.12, $p=.12$), showing the same average presence of 1.21 different signatures per host (Welch's ANOVA F-test = 2.88, $p=.09$). Similarly, Worms are reported 7.40 (consumers) and 7.38 (enterprises) times on average (Welch's ANOVA F-test = 0.02, $p=.90$), in the form of 1.24 and 1.21 distinct labels (Welch's ANOVA F-test = 1.06, $p=.08$). Our data reveals that the family of *Silly* Worms, that replicates through email attachments and local copies to steal sensitive information and disable other software, is the most common in its corresponding class. *Sality*, a popular malware that infects executable files acting as backdoor or botnet, dominates instead the scene when it comes to Viruses.

4.2 Distribution of malware subclasses

Figure 1 shows the cumulative distribution of the number of distinct malware subclasses observed in enterprise and consumer hosts. For each machine, a subclass is counted if at least one of its signatures is matched by the AV product. The maximum number of distinct classes (22 for consumers and 21 for enterprises) has been reported by two machines per group. While at a first sight the graph might suggest similar behaviors in the two categories, the Chi-squared tests separately considering up to 20 distinct encountered categories reported significant differences with $p < .001$. In particular, substantial differences are present in the leftmost part of the plot: while nearly 82% of enterprise hosts have encountered only a single subclass of malware, this percentage drops below 57% for consumers. This, in turn, reveals that on average consumer machines are more likely to encounter a more diverse set of malicious files than enterprise computers. As already discussed in the introduction, a possible explanation for these differences can be the adoption of stricter security policies and multiple layers of defenses present in enterprises but not in consumer environments.

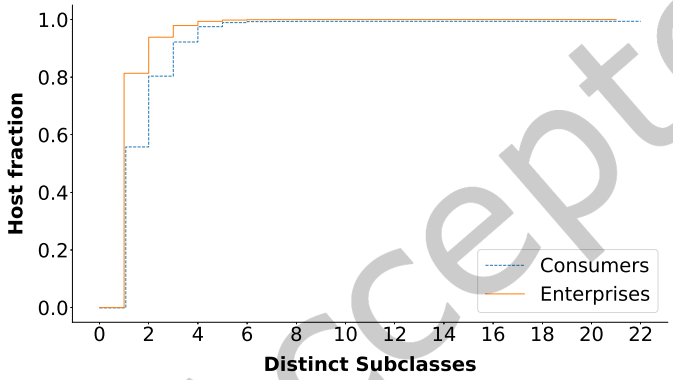

Fig. 1. Cumulative distribution of the number of distinct subclasses per host

Our measurements show that for most of the malware categories there was no relevant change over the year in terms of the fraction of hosts that detect them. This supports the hypothesis that different malware classes reach a plateau that they maintain over time despite the effort of security companies to mitigate them. There were only two exceptions to this rule, which we present in Figure 2. The first was a slight but steady decrease of *Ransomware* families, both in consumer and enterprise data. The second was a rapid increase of *Cryptominer* families, followed by a general downward trend. Ransomware and Cryptominers are the last two malware classes that emerged over the last few years and their curves show that in fact they did not yet reach a stable trajectory.

4.3 Age of encountered malware

We continue our analysis by estimating how *old* the malware encountered by the hosts in our dataset is, by looking at the date in which each signature was first introduced by the vendor. Figure 3 depicts the average age of matched signatures in our one-year observation period. For each of the 12 months, we group all the labels based on the year in which they were created. Then, for each of the 29 years (from 1990 to 2018) we average the number of distinct records over the months and compute the 95% confidence interval. Despite a common peak of over 300 signatures written in 2014 and a drop for those developed in 2018, the number of matching signatures present in our dataset is almost constant since 2003. This corroborates what has already been observed in other studies about the fact that it is still common to encounter today samples belonging to very old malware families [19]. In fact, about 174 K consumer hosts and 151 K corporate machines (respectively 1.0% and 0.5% of

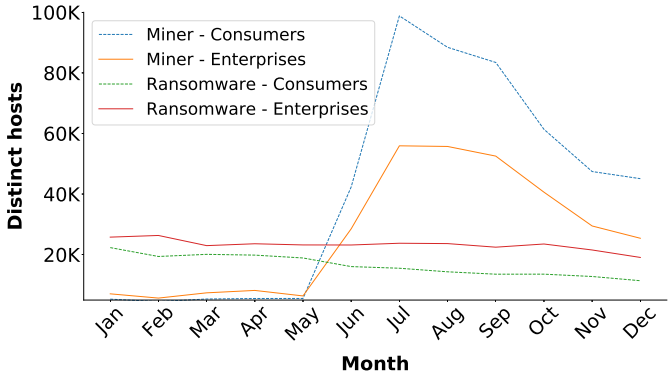

Fig. 2. Ransomware and Miner trends for consumers and enterprises

those that suffered at least a malware encounter) report encounters for signatures whose creation even predates the year 2000. Among those, the most common for consumers (4858 hosts) and enterprises (1990 hosts) is *CIH*, a 22-year-old signature to identify a computer Virus that targets Microsoft Windows 9x systems.

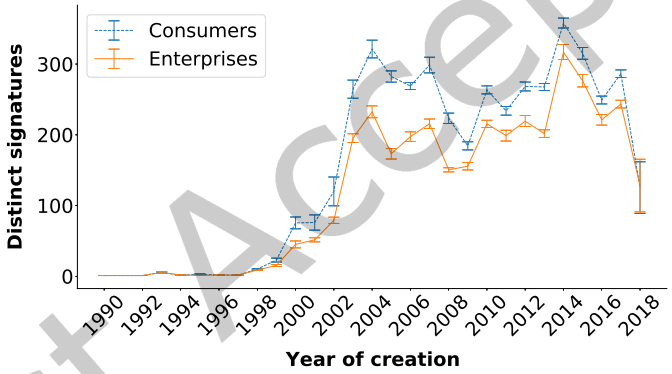

Fig. 3. Average number of different signatures per year of creation. Error bars provide a 95% confidence interval

4.4 Behavioral signatures, Adware, and PUA prevalence

So far, in this manuscript we have used the word *signature* to indicate without distinction the set of unique data that allows an AV software to detect, quarantine, and remove specific malware. However, two main approaches exist to create a signature: the older pattern-based methodology in which a model was built to match a particular family of malware, and the more recent behavioral-based approach in which generic heuristics are used to capture different aspects of malicious behavior. While the first leverages object attributes to create a unique fingerprint, the latter typically evaluates an object based on its runtime actions [6].

In our dataset, we identified 6.7 K behavioral signatures by using their label and report their prevalence for consumers and enterprises in Figure 4. The reported percentages are obtained by dividing the number of distinct hosts with at least one behavioral-based detection and the number of distinct hosts that have suffered one or more encounters of any kind. We verify that all the monthly differences are statistically significant ($p < .001$)

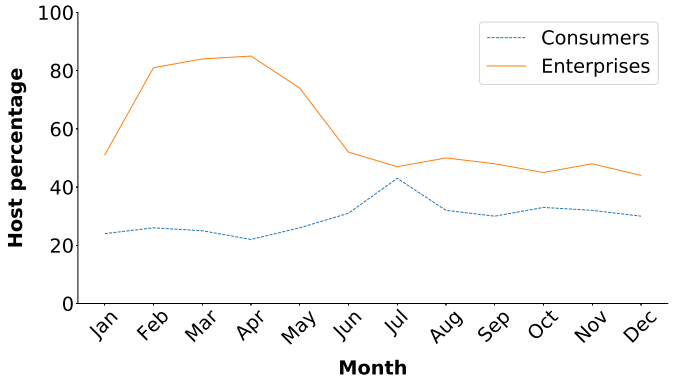

Fig. 4. Prevalence of behavioral signatures in consumer and enterprise machines

by running a Chi-squared test on the contingency table obtained by considering devices that trigger behavioral signatures and those that do not, in consumers and enterprises.

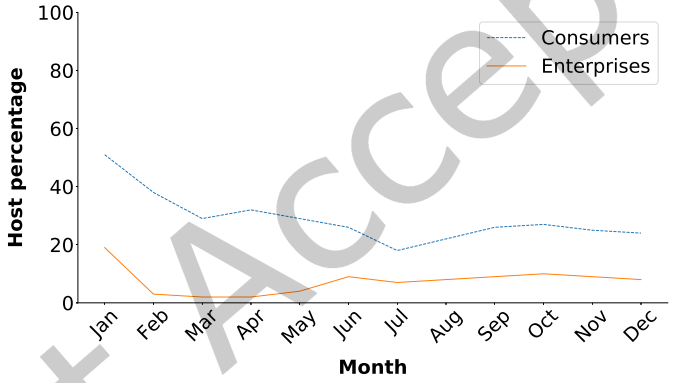

Fig. 5. Prevalence of PUA and Adware signatures in consumer and enterprise machines

The curve for enterprise hosts lies considerably above the one of consumers, a sign that behavioral signatures match much more in the former environment (an average of 59% of hosts in enterprise vs 30% in consumer hosts). This could be due to the presence of less popular software and of custom applications built and compiled on corporate machines, for which the AV has not been tested against to whitelist or tune its behavioral signatures. On the contrary, consumer machines mostly run well-known applications that are therefore accounted for by AV vendors. However, since the totality of behavioral signatures is categorized as Trojan by the AV vendor, we speculate that this difference could also be due to sophisticated malware that targets specifically certain enterprises, which could not be easily detected with a traditional pattern-based signature.

In Figure 5, the trends are inverted when considering Adware and PUA. In fact, their prevalence in consumer hosts is constantly higher (6.06 times on average with a statistically significant difference for each month — $p < .001$) than in enterprises. As already discussed in section 4.2, a very likely explanation can be found in the freedom that consumer users have to install any kind of software, whereas more rigorous rules are enforced in enterprises.

5 UNDIVERSIFIABLE RISK ANALYSIS

In the previous section, we extensively analysed the systemic risks that affect consumers and enterprises when it comes to malware encounters. As discussed at the beginning of the manuscript, systemic risk takes into account the different use of machines, security posture and policies that characterize the two environments. The analysis of the telemetry at our disposal revealed that the different adoption of security measures and policies implies a diverse threat landscape in the two segments, with the prevalence of specific classes of malware in consumers (e.g., Adware, PUA) and enterprises (e.g., behavioral signatures).

Differently from the previous one, the portion of risk that we define as systematic refers to the risk introduced by objective factors that do not reflect any aspect of the security posture, policies or nature of machines. In the next sections, our goal is to investigate whether this kind of risk exists in the cyber domain, and identify correlated indicators for consumers and enterprises that can help us to measure its significance together with the differences between the two classes.

To this end, we employ regression analysis by constructing several models that simultaneously use a combination of host attributes as regressors, thus controlling for conflicting explanatory variables when modeling the risk of encountering malware. We detail the model generation in section 5.1 and deeply discuss each risk factor in the subsequent sections.

5.1 Model generation

We postulate that the monthly risk of encountering malware for one host is influenced by a combination of the following seven independent variables: active days and hours, file-request volume, reputation and number of installed vendors, geographical location and whether or not malware has already been detected on the machine the month before.

Our objective is to obtain a Log-Odds distribution for the dependent variable Y, that expresses the odds —the ratio of successes (host encounters malware) and failures (host is clean)— as a linear combination of the regression variables. Since Y is monthly given in our telemetry as a boolean value (i.e., host encounters malware or is clean), we transform it as to obtain a count by bucketing numerical variables (days, hours, files created, vendor number and enterprise size) into bins to reduce granularity, grouping all the machines that share the same combination of values, and counting how many of them are infected or clean.

We then make use of Generalized Linear Models (GLMs) [3], test them in different configurations, and analyze the outcome of several goodness-of-fit quantities (Pseudo R-Squared, Log-Likelihood, Dispersion, and the estimation provided by the Akaike Information Criterion (AIC)). We achieve the best results when modeling the risk of malware encounters Y as a Binomial distribution using a Logit link function. The analysis of the pseudo-R-Squared values obtained when modeling the different malware classes along the year revealed that, on average, between 68.4% and 89.9% of variance in the encounter rate is explainable by the chosen control variables.

Once the model has been fitted to the data, the extent to which the independent variables influence the dependent variable is captured by their regression coefficients. In particular, for each regressor, we select a bin (e.g., 0-4 days) or categorical value (e.g., North America) as a reference baseline, and express the odds ratios of other bins or values to derive the attribute's importance.

We separately model consumer and enterprise machines. We are aware that comparing the *magnitude* of odds ratios from models that use different samples from different populations may introduce an error [26]. However, our ultimate goal is to analyze the *trends* within each segment —odds ratios increasing, decreasing or fluctuating— and hereinafter we never directly compare the magnitude of the coefficients between consumers and enterprises. As further evidence of the correctness of our results, we also examined the case of a GLM that combines the two set of machines (consumers and enterprises) by adding a regressor (machine_type: 0 = consumer, 1 = enterprise). For the sake of completeness we report the odds ratios obtained in Table 11 of Appendix A.

Table 6. Odds ratios of encountering malware according to our regression models

Host Attribute	Bin Category	Malware family	Consumers Monthly Odds μ	Consumers Monthly Odds σ	Enterprises Monthly Odds μ	Enterprises Monthly Odds σ
Activity Days	4-8	Any	2.10	0.19	1.44	0.19
	8-12	Any	2.78	0.45	1.59	0.30
	12-16	Any	3.26	0.67	1.82	0.57
	16-20	Any	3.58	0.85	1.91	0.71
Ref: [0-4]	20-24	Any	4.01	1.11	1.97	0.75
	24-28	Any	4.15	1.25	1.79	0.73
	28+	Any	4.51	1.33	1.85	0.48
Activity Hours	3-6	Any	1.34	0.09	1.02	0.20
	6-9	Any	1.57	0.32	0.95	0.14
	12-15	Any	1.25	0.47	0.88	0.21
	15-18	Any	1.35	0.38	0.98	0.23
	18-21	Any	1.59	0.49	0.99	0.39
Ref: [0-3]	21+	Any	2.65	1.67	1.32	0.56
	18-21	Adware	1.68	1.32	0.63	1.46
	21+	Adware	3.30	2.23	0.08	0.25
File-volume Activity	1K-2K	Any	1.05	0.07	1.19	0.26
	3K-4K	Any	1.64	0.33	1.33	0.54
	5K-10K	Any	2.21	0.55	1.59	0.90
	10K-50K	Any	3.19	1.05	1.85	0.79
Ref: [0-1K]	50K+	Any	4.77	1.23	2.34	1.38
	10K-50K	Adware	9.67	3.87	2.62	1.77
	50K+	Adware	13.52	4.71	9.79	3.76
Vendors	20-40	Any	1.11	0.04	1.09	0.12
	40-60	Any	1.22	0.06	1.30	0.28
	60+	Any	1.39	0.09	1.54	0.55
Ref: [0-20]	60+	Adware	1.46	0.31	4.86	3.74
	60+	PUP	1.56	0.09	3.37	1.06
Reputable vendors only	Yes	Any	1.00	0.05	0.99	0.25
	Yes	PUP	0.98	0.05	0.82	0.06
Ref: No	Yes	Virus	0.64	0.04	0.70	0.09
Repeat player	Yes	Any	1.77	0.77	1.33	0.49
	Yes	Adware	8.33	3.15	5.86	1.14
	Yes	Virus	2.21	1.03	5.50	2.56
Ref: No	Yes	Worm	10.56	2.45	8.44	2.82
Geographical location	AF	Virus	6.35	2.10	12.14	2.49
	AS	Virus	4.19	0.51	9.72	1.21
	AF	Worm	20.77	2.61	18.59	4.61
	AS	Worm	5.39	0.23	9.49	2.31
Ref: NA	OC	PUP	0.86	0.18	1.25	0.45
	OC	Trojan	1.04	0.10	0.82	0.22

In our experiments, we consider each month separately, as data are monthly aggregated due to anonymity constraints. We run a separate model for each month starting from February, by only considering hosts that have been active all the 12 months (11.7M consumer machines and 2.8M hosts of 33.7K enterprises), as to have information of the previous-month clean/infected state always available. At first, we define one host being targeted by malware if it encounters any kind of malware in that specific month. In addition, we separately consider and model five different malware classes —Adware, Trojan, PUP, Virus, Worm— to explore any variations in host-attribute importance or differences between consumers and enterprises when narrowing down the analysis to a specific class. In Table 6, we report the average μ and the standard deviation σ of the odds ratios along the 11-month period for the most explanatory cases that we discuss in the following sections. We note that all reported values are statistically significant with $p < .001$ for all the months of the considered period.

We do not include the enterprise size (i.e., number of hosts) and its industrial sector in the previous experiment, as these regressor variables are not available for consumers. In fact, the odds analysis of models that have been constructed with different variables is statistically unsound [26]. We instead repeat the experiment by isolating enterprise machines and simultaneously modeling all the 9 attributes at our disposal for this segment of hosts. In Table 10, we only report the odds ratios of the two features that were added at this step. Also in this case, all reported values —including those that are not reported in Table 10— are statistically significant with $p < .001$ for all the months of the considered period.

5.2 Time-based activity

It is reasonable to expect that the longer a machine is active, the more likely it is to encounter malware. Indeed, the odds of detecting malware for consumers linearly increase with the number of active days, reaching a 4.51 factor with respect to the reference class for those active on average more than 28 days per month. A similar relationship also exists for enterprises, where the odds reach a peak of 1.97 when considering those hosts active between 20 and 24 days per month. Activity days represents a stable indicator along the months, as detailed by the low standard deviation in relative odds. A similar trend exists also with respect to the number of hours of activity per day but, in this case, both enterprise and consumers show a similar random behavior for those machines active on average more than 9 hours per day.

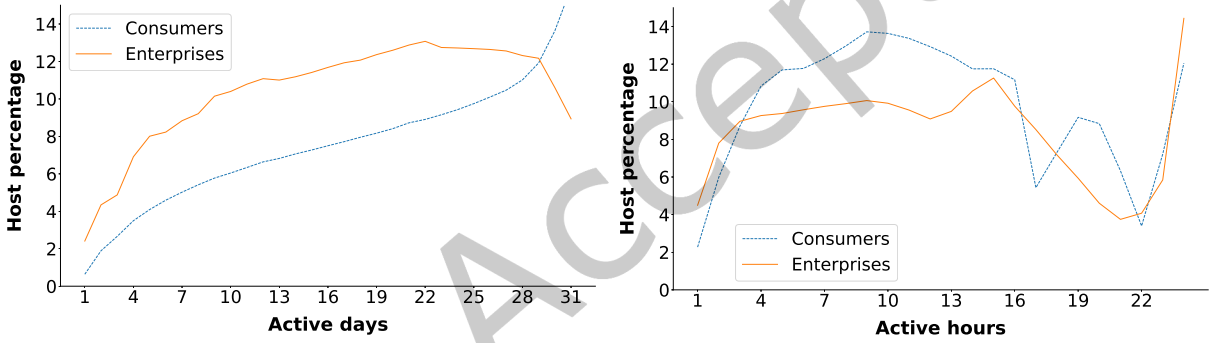

Fig. 6. Disjoint influence of activity days (left) and hours (right) on malware encounters

To better understand this phenomenon, we separately assess the influence of activity days and hours in Figure 6. We split the machines based on their average uptime days and, for each of the 31 days, we compute the percentage of hosts that detect malware. Regardless of the number of days, we repeated the same task for the number of uptime hours. While for consumers the plot suggests that malware detection rates keep increasing alongside the number of active days, for enterprises this growth stops at around 20 days (roughly the number of working days in a month), but then the curve considerably drops for machines that are always running. The same trend is exhibited by looking at the daily hours of activity. In this case, the growth of the encounter rate stops at around eight hours for both groups (which again seems to align with the number of working hours in a day). As we clarify later in the section, these values seem to suggest that the active time changes with the *role* of the machine, and different roles may have very different encounter rates.

With these results in mind, we identified a set of machines for which the time-to-risk relationship was more regular. These include machines with up to eight hours of activity per day and, for enterprises, hosts that are active no more than 20 days per month. This group accounts respectively for 96% of the consumer hosts and for 73% of the enterprise machines. Figure 7 shows the joint influence that activity time has on the Regular Group:

for each day X and each hour Y, the point on the 3D surface is given by selecting the machines active for X days and Y hours on average, and computing the percentage of those that detect malware. Interestingly, both plots follow a smooth behavior according to the one of the two control variables, confirming the goodness of time activity as a risk indicator for this type of machine.

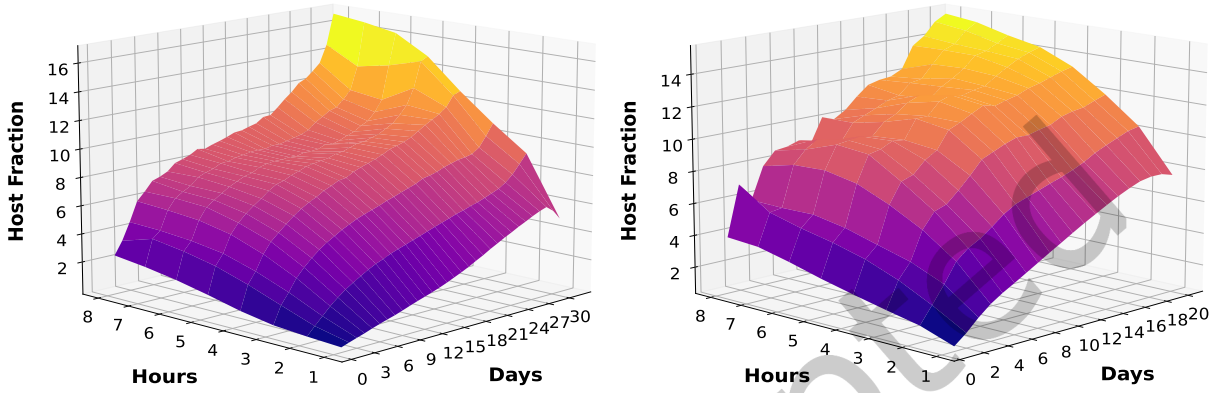

Fig. 7. Joint influence of activity days and hours on malware encounters for consumers (left) and enterprises (right)

For machines in the regular group, we also computed to what extent each additional day or hour of activity increases the odds of encountering malware by fitting a model that considers days and hours as integer variables, while keeping unchanged the other regressors. We measure that for any additional day of activity the odds of encountering malware increases by 4% for consumers and 3% for enterprise machines. An additional hour of daily activity results instead in an additional 17% and 6% extra risk. At first, both results suggest that adding more daily uptime has a stronger impact than adding more days of activity (consumers: 17 > 4; enterprises: 6 > 3), but we need to keep in mind that machines in the regular group have a maximum of 8 hours of daily activity vs 20 (for enterprise) and 31 (for consumers) days per month. If we repeat the experiment by considering only specific classes of malware, in the case of number of active days we find a consistent behavior with the general case in both odds magnitude and increasing trends.

The odds related to number of active hours per day deserve instead special attention. In fact, we observe that for enterprise machines running more than 8 hours per day (i.e., the threshold we identified for the regular groups), the odds across all malware classes are lower than for hosts active fewer hours per day. We speculate that the reason is that those machines are likely dedicated to performing not-interactive tasks (e.g., servers). This hypothesis is confirmed by looking at the almost-zero odds of encountering Adware in enterprise machines that are always running: since this particular malware is usually shipped during software installations or web-browsing activity, very low odds of suffering this kind of infection can be explained by the lack of this kind of tasks. On the contrary, we observe a decrease-increase behavior for consumers, an indicator that those machines are probably used in both automated and interactive fashion.

Summary: Time activity can clearly reflect the risk of encountering malware only when a subset of "regular machines" is considered: both uptime days and hours can reliably act as control variables when evaluating detection rates of machines active less than 8 hours and 20 days for enterprises. However, time activity for general enterprises is not a good predictor.

5.3 File-based activity

As we already mentioned, the machines in our dataset routinely query a centralized system to assess the reputation of new objects: by monthly counting the number of these requests, we build a second metric for host activity and correlate its magnitude to the odds of malicious program detection. We find that the odds of detecting malware steadily increase with the level of activity in terms of files generated for both consumer and enterprise hosts and across malware families. This relationships does not vary month by month, as confirmed by the very low standard deviation reported with the mean. While we observe a similar magnitude in the odds of machines that generate less than 5K files per month, the effect of a greater file-volume activity (5K+) more consistently impacts consumer hosts. At its extreme, we observe that the odds of infection reach twice those of enterprises when selecting machines that generate a very high file-volume activity (50K+).

In Figure 8, we provide the reader with a visual representation of the relationship between files generated and malware encounters: for a given number X of file requests, we group the machines that queried the centralized system exactly X times in a month, and compute the percentage Y of those that encounter malware. The orange curve in the graphs provides an indication of the underlying trend, and it has been obtained by sampling the percentage every 100 values. The two figures reveal a similar logarithmic trend for both corporate and consumer machines. While for a low number of queries (up to roughly 5 K for consumers and 2.5 K for enterprises) a rise in the file-based activity entails a severe increase in the malware encounter rate, this effect gets weaker as we move in the right part of the plot.

Summary: The odds of detecting malware steadily increase with the level of activity in terms of files generated for both consumer and enterprise hosts and across malware families. The effect of file-based activity is more prominent when considering a low number of requests, up to roughly 5 K for consumers and 2.5K for enterprises.

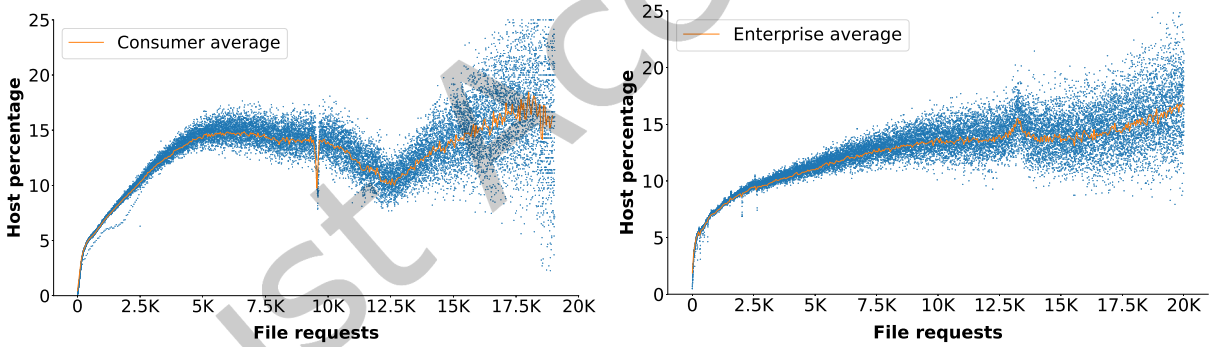

Fig. 8. File volume influence on malware encounters for consumers (left) and enterprises (right)

5.4 Software vendors

We now measure to what extent various machine profiles might have an impact on the overall risk. We achieve this by looking at the set of software installed on the computers, extracting the vendor name from the publisher subject that can be obtained from signed binaries. On the vast majority of computers (around 80% for both groups), we identify software that is signed by between 10 and 15 different publishers. The maximum numbers of publishers identified on a single machine were 2312 for consumers and 349 for corporations. We first test whether an increasing number of software vendors implies a higher risk of detecting malicious programs. The rationale

behind including the vendor number as a regressor in our model is that the odds of encountering malware –and in turn suffering from security issues– may raise according to the number and diversity of software installed in a system.

Our modeling reveals that a relationship exists between the two variables, and that enlarging the set of software installed on a machine results in higher odds of encountering malware. For instance, consumer and enterprise machines with a number of vendors between 20 and 40 are 1.11 and 1.09 times more likely to be targeted by malware than those with less than 20 signers. Odds increase to 1.22 and 1.30 for hosts with a number of vendors between 40 and 60, and reach 1.39 and 1.54 for those with more than 60 vendors. Once again we measure a very low standard deviation, which suggests that results persist for all the considered months. When restricting to Adware and PUA, we find that the presence of a very high number of vendors entails higher odds ratios in enterprises (4.86 and 3.37). This trend is not reflected for home users, for which the odds follow the general case.

We further dive into the relationship between a diverse set of programs and malware encounters by dividing both consumer and enterprise machines into groups based on the number of different software vendors installed. For each group in which we have at least 100 elements, we compute the fraction of hosts that encountered malware at least once.

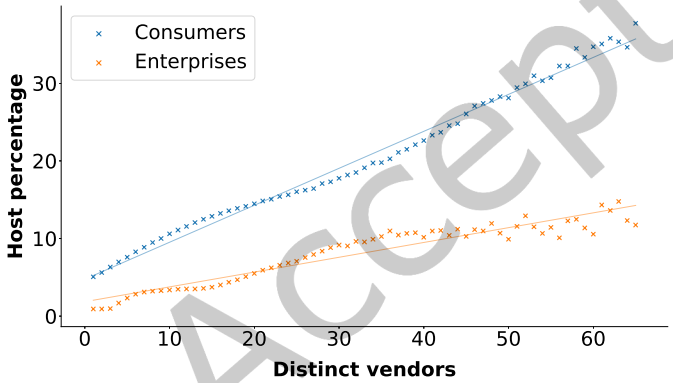

Fig. 9. Relationship between the number of distinct vendors installed and hosts that encounter malware.

In Figure 9 we report these percentages together with straight lines, that represent linear regressions obtained using the least-squares method with a mean squared error of 1.86 for consumers and 4.75 for enterprises. Again, the diversity of software installed on the computers positively and *linearly* correlates with the rate of malware encounters. This is true both for consumers and enterprises, with the difference that the slope associated with the consumer trend is steeper than the one of enterprises. This discrepancy is also reflected by the higher odds ratios in the former group. We can justify this behavior with the fact that in enterprise contexts, even if a user were to install a diverse set of applications, each of them would probably serve the purpose to carry some tasks related to her job: indeed, with the existence of security policies, users are less likely to install software from dubious origin on the machines provided by their employers.

As a further insight, we also consider whether the nature of the installed software influences the odds of malware encounters and whether that could be used to profile the role of the machines that installed them. We rank the top 20 vendors in our dataset based on the number of hosts on which they appear and report their list in Table 7.

Our hypothesis here is that the machines that installed only those could be used as a control group, as they might belong to regular user profiles who only use common software, such as browsers, document editing

tools and such. We therefore create two different profiles, isolating machines with only top-20 vendors installed from the rest: while for enterprises this set is composed of around 12% of the active hosts, this percentage rises above 42% for consumers. In our tests, we found that a higher vendor reputation has a negligible contribution to lowering the odds of encountering malware. Indeed, we register no changes in odds for consumers ($\mu = 1.00$ and $\sigma = 0.05$) and a small decrease for enterprises ($\mu = 0.99$ and $\sigma = 0.25$). However, we register a more significant impact when modeling malware classes as PUA (0.98 consumers - 0.82 enterprises) and Adware (0.64 consumers - 0.70 enterprises): in this case, the presence of only reputable vendors is an important factor that contributes in lowering the odds of encountering particular families that are usually shipped with dubious software.

Summary: The larger and the more diverse the set of applications installed on a machine is, the higher is its odds to encounter malware. On average, an increase in installed programs entails more risk of malicious software detections for consumers than enterprises. In addition, software reputation is important in lowering the likelihood of encountering particular malware classes, as Adware and PUA.

Table 7. Top-20 vendors for consumers and enterprises

Consumers	Enterprises
Microsoft Corporation	Microsoft Corporation
Symantec Corporation	Symantec Corporation
Google Inc	Google Inc
Apple Inc	Adobe
Adobe	Intel
Dell Inc	Oracle America Inc
Mozilla Corporation	Citrix Systems Inc
Intel Corporation	VMware Inc
NVIDIA Corporation	ESET
HP Inc	Mozilla Corporation
McAfee Inc	Cisco Systems Inc
Dropbox Inc	Hewlett Packard Company
Hewlett Packard Company	Lenovo
OracleAmerica Inc	Pulse Secure LLC
ESET	Dell
Garmin International Inc	Sun Microsystems Inc
Wild Tangent Inc	Apple Inc
Valve	NVIDIA Corporation
CyberLink	LogMeIn Inc
Lenovo	CrowdStrike Inc

5.5 Repeat players

We now assess whether being a repeat player has an impact on the odds of encountering malware. When fitting the model for a specific month, we consider a machine being a repeat player if malicious software was detected on it the month before. Our hypothesis is that repeated encounters with malware can be a sign of users' hazardous behaviors or of their poor security practices during the year under analysis.

In fact, we found a difference ($\mu = 1.77$ for consumers and $\mu = 1.33$ for enterprises) between the odds that a recidivist host will encounter malware versus a clean machine. The importance of this risk factor and the

differences between home and corporate users increase when considering malware classes as Adware, Worm and Virus. When looking at consumers and at the first two cases, repeat players are 8.33 and 10.56 times more likely to encounter malicious software than machines that were clean the previous month. Odds increase also for enterprises, where we register factors of 5.86 and 8.44.

Summary: Recidivists hosts have higher odds of encountering malware with respect to clean machines. This finding is more pronounced when individually considering malware classes as Adware, Worm and Virus.

5.6 Geographical location

Previous works show that the number and types of malware that computers encounter vary greatly across countries [2, 24, 25]. To verify these findings, we consider the continent in which one host is located as a regressor variable, and model how the odds of encountering malware vary with the geographical location.

When considering all malware categories, we register the same order of odds magnitude both across countries and types of machines. On the contrary, geographical location constitutes a considerable risk factor when restricting to Worms and Viruses. For those classes, we measure comparable odds in North America, South America, Europe and Oceania, but register a massive increase in continents like Africa (> 18 for Worms, > 6 for Viruses) and Asia (> 5 for Worms, > 4 for Viruses) for both consumers and enterprises.

In addition to the analysis of the odds ratios, we separately assess the incidence that different malware classes have across continents. We report in Tables 8 and 9 a complete geographical breakdown of the percentage of hosts that encounter a specific family.

Table 8. Geographical breakdown of malware classes for consumers

	Consumers					
Country	Trojan	PUA	OT	Worm	Adware	Virus
Africa	64.01	15.36	7.15	**9.3**	0.8	**3.38**
Asia	**66.71**	14.47	10.94	3.81	1.47	2.61
South America	60.17	23.62	12.49	1.5	1.65	0.57
Europe	56.57	**25.51**	14.4	0.99	2.22	0.31
North America	58.04	24.8	13.68	0.61	2.59	0.27
Oceania	58.17	22.19	**15.67**	0.93	**2.8**	0.25

Table 9. Geographical breakdown of malware classes for enterprises

	Enterprises					
Country	Trojan	PUA	OT	Worm	Adware	Virus
Africa	77.05	**5.80**	4.87	**9.57**	0.34	**2.36**
Asia	81.76	5.59	**6.63**	3.28	**0.45**	2.29
Europe	87.42	4.88	6.21	0.94	0.3	0.26
North America	**93.39**	2.34	3.53	0.44	0.17	0.12
Oceania	90.88	3.17	4.84	0.72	0.2	0.2
South America	86.64	5.18	5.72	1.42	0.4	0.65

The results are in line with what reported in a previous study by Mezzour et al. [24], who found a predominant prevalence of Worms and Viruses in Sub-Saharan Africa and South Asia. In the opposite direction, we find that machines in Oceania have lower odds of encountering Trojans and PUA. Here, we find that the odds home-users facing PUA are reduced by a factor of 0.86 with respect to those in North America. A similar result holds for corporate machines whose odds ratio of encountering Trojan is 0.82.

We also tested whether there exist geographical regions where many machines encounter some malware families that appear very rarely elsewhere. To analyze this aspect, we first ranked all the signatures in our dataset based on the number of distinct hosts on which they have been detected. We then isolated the top-100 labels among behavioral signatures, PUA and Adware, and the remaining set of malware and, for each label, we broke down the machines that have encountered it across continents (Figure 10).

Although we identified some differences, machines located in North America, Europe and Asia encountered the top-100 signatures with a similar frequency, while Africa and South America follow different behaviors.

After discarding generic cases, we observe that the family of the trackware *TransitGuide* (218 K hosts), developed to monitor browser activity of the targets, and of the Trojan *Kotver* (122 K hosts), that performs click-fraud operations in order to generate revenue for its authors, are almost exclusively detected in consumers located in North America (97% and 92%). At the same time, the Adware families of *KpZip* (22 K machines for consumers and enterprises) and *Funshion* (19 K consumer and 11 K enterprise machines), both created with the aim of displaying ads to profit from user clicks, are mostly encountered by computers located in Asia (92%).

Summary: Some categories of malware (e.g., Trojan, PUA, and Adware) affect equally companies and home-users in all continents. In contrast, the odds of encountering Worms and Viruses are on average from 4.19 to 20.77 times higher in Asia and Africa.

5.7 Enterprise size and industrial sector

We finally focus our analysis on the risk profiles of enterprises with different sizes and industrial sectors. As reported in Table 10, the odds ratios related to small, medium and large organizations slightly differ from the baseline of micro firms. In addition, we do not observe any trend that relates an increasing number of hosts to higher or lower odds of malware detection, but instead register a fluctuating value when considering any malware class as well as when narrowing to specific categories. This suggests that the enterprise size is not correlated with the likelihood of malicious software encounters.

To get a clearer picture of this relationship, we decide to separately consider the enterprise size as a risk factor. Figure 11 shows two scatterplots in which each blue dot represents a separate enterprise, and on the axis we report its size (i.e., number of computers) and the fraction of its machines that encountered malware at least once in the one-year period of our experiments. Green crosses indicate clean enterprises, i.e., companies whose hosts do not encounter malware in the considered timeframe. The orange line shows the average among companies of the same size, considering both clean entities and those that encounter malware. We also plot a dotted line showing the average consumer rate—i.e., the ratio of consumer machines that had at least one encounter (9.8%)—with the aim of detecting whether the consumer encounters distribution is more similar to that observed in enterprises with a particular size.

In line with the insights gathered analyzing the odds ratios, the left figure depicts an almost constant trend, slightly above the consumer line, with a flexion of the curve for those companies with sizes lower than 50 machines or higher than 100 K hosts. This may seem to suggest that small (<50) and large enterprises (>100 K) tend, in proportion, to have a smaller number of computers that encounter malware. However, the difference is very small and the Pearson correlation coefficient for size and the fraction of hosts that encounter malware is 0.01, indicating a negligible relationship between the two. Once again, this is a sign the number of machines in enterprises is not correlated to how much malware is detected.

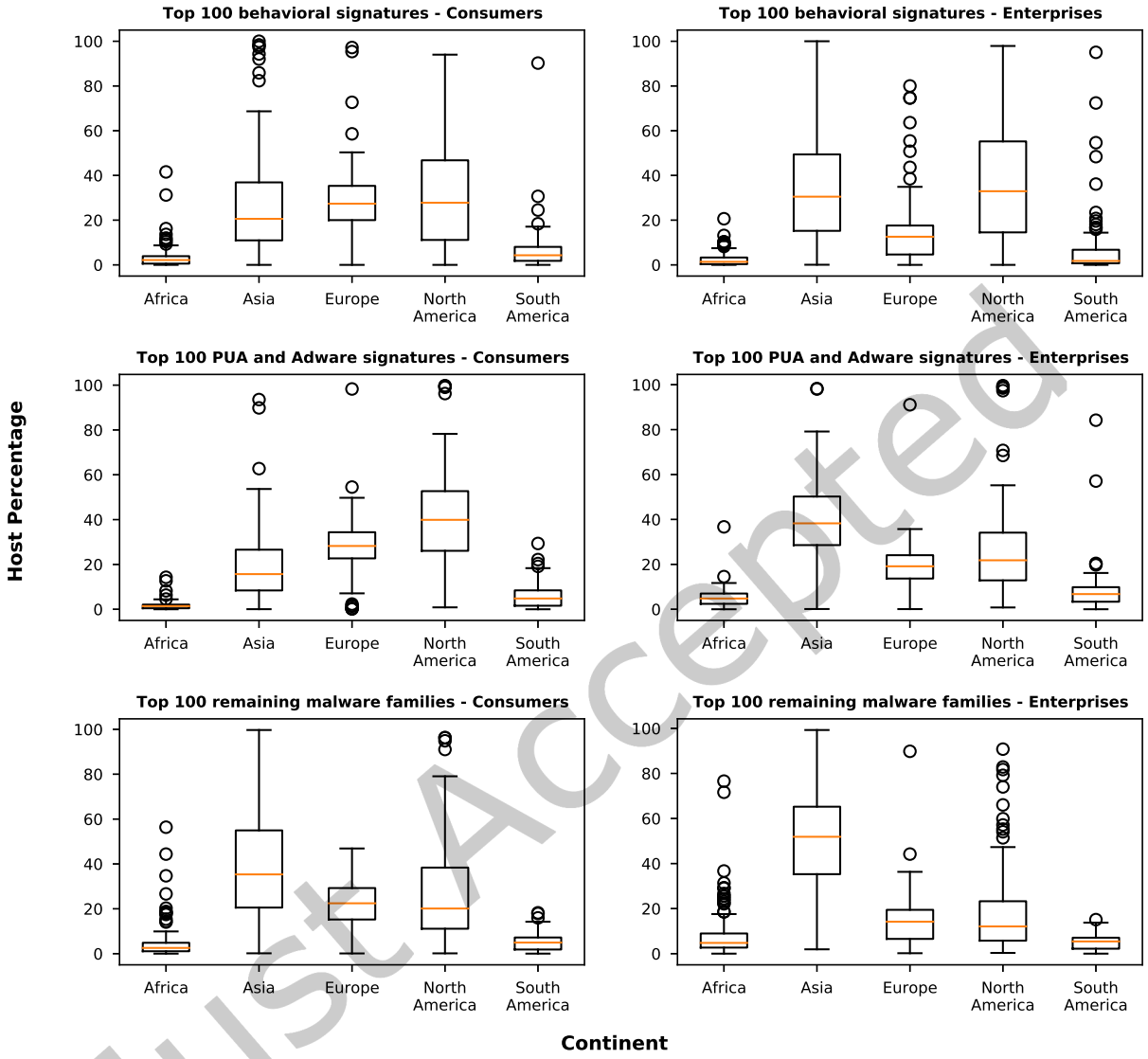

Fig. 10. Breakdown of top 100 behavioral signatures, PUA and Adware, and remaining malware families. Percentages are sorted according to the number of distinct hosts on which the signatures have been detected.

To further investigate this aspect, we decided to focus our analysis only on those machines that were active for each of the 12 months of our experiment (2.8 M hosts of 33.7 K distinct enterprises). The rationale behind this choice is that hosts active only few months have less likelihood of reporting detections, thus lowering the average encounters rate we are interested in measuring. In the right part of Figure 11, companies are still represented by blue dots. However, while the x-coordinate indicates the enterprise size (as in the previous case, obtained considering all machines), the y-coordinate is computed by considering only hosts active 12 months, and thus dividing those that encounter malware by their total number.

Table 10. Odds ratios of encountering malware according to our regression models for enterprise size and industrial sector

Host Attribute	Bin Category	Malware family	Enterprises Monthly Odds μ	σ
Enterprise Size Ref: [0-10]	10-50	Any	1.09	0.20
	50-250	Any	1.04	0.19
	250+	Any	0.98	0.49
Enterprise Sector Ref: Information Technology	Consumer Discretionary	Any	1.55	0.47
	Consumer Staples	Any	1.02	0.25
	Energy	Any	1.82	0.51
	Financials	Any	1.01	0.36
	Healthcare	Any	0.93	0.30
	Industrials	Any	1.32	0.28
	Materials	Any	1.48	0.38
	Telecommunication	Any	1.37	0.66
	Utilities	Any	1.56	0.26

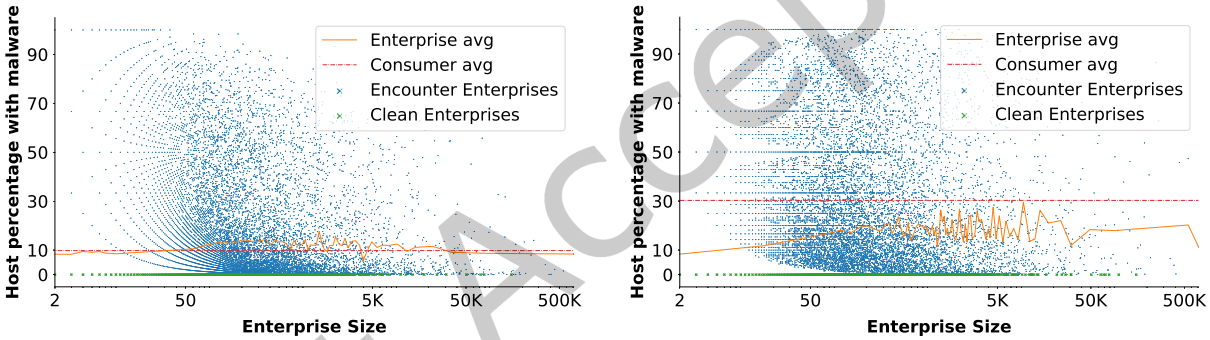

Fig. 11. Relationship between enterprise size and the fraction of hosts that encounter malware, computed for any host (left) and for those active each of the 12 months of our experiment (right)

Interestingly, the effect of this filtering is more pronounced for consumers, where the percentage of machines that encounter malware raises to 30.3% (+ 20.4%), while in enterprises we register an average of 21.5% (+9.5%). We also observe a discrete gap between small organizations (<50) and those with a number of hosts comprised within 50 and 500 K: while for the former the mean stays around 16%, in the other case it reaches 23%. While this may indicate the existence of a relationship between enterprise size and malware detection rate, overall we still observe a very low Pearson correlation coefficient (0.02). In fact, excluding the companies with less than 50 machines, the remaining set of organizations (> 50 and $< 500K$, i.e., 92% of the total) exhibit an almost constant trend regardless of their size.

To gather further insights, we verify whether the industrial sector affects the relationship between the size of enterprises and the malware encounter rate. For this, we compute the Pearson correlation coefficient to measure the extent to which an increase in enterprise size leads to a higher number of hosts that detect malware. We also report in Figure 12 a plot for each sector. Again, we do not observe any general correlation, similar to the one

obtained by looking at all enterprises (0.01), a sign that the number of hosts alone does not play a very important role in explaining the encounter rate.

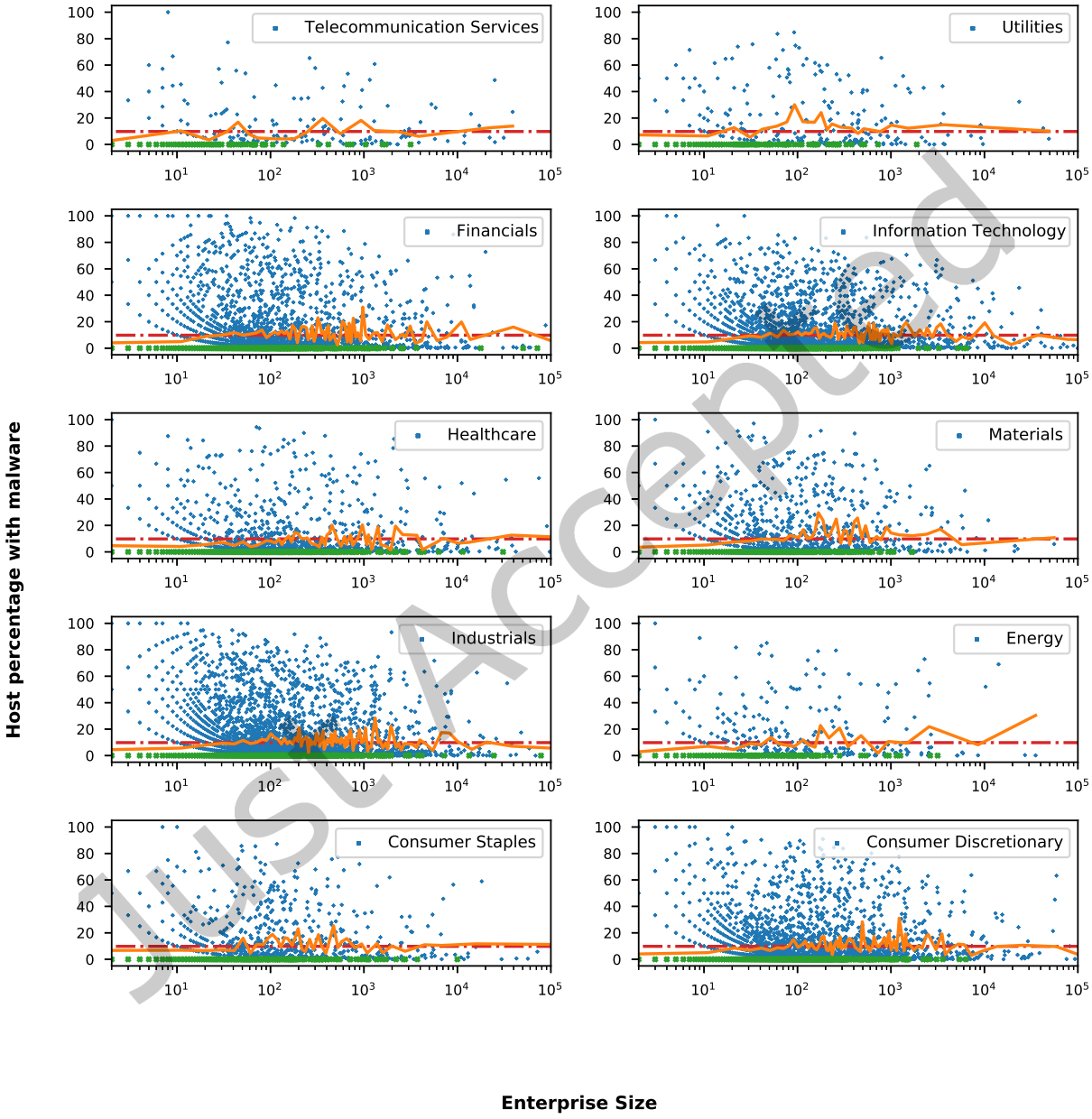

Fig. 12. Relationship between the enterprise size and the fraction of hosts that encounter malware. Each plot represents a distinct sector.

In addition, we conduct a test to verify whether a statistically significant difference exists among the distributions in Figure 12 across industrial sectors. We opt for a non-parametric Kruskal-Wallis one-way ANOVA rather than a parametric one-way ANOVA, as this allows us to relax the one-way ANOVA assumption of data normality, which is not met in our case. The Kruskal-Wallis test assumes that a) the independent variable (enterprise size) has two or more independent groups; b) the measurement scale of the dependent variable (ratio between hosts that encounter malware and enterprise size) is ordinal, ratio or interval; c) the observations within a group and among groups must be independent. d) no data distribution assumptions if the test is used as a test of dominance, i.e., to verify whether at least one group stochastically dominates another one. With those assumptions verified, we run the test, our null hypothesis being that the samples come from populations with the same distribution. We obtain a test statistic H = 13.75 (p=.13), values that do not allow us to reject the null hypothesis: we conclude, once again, that the malware encounter rate based on enterprise size is not influenced by its industrial sector.

To conclude the study of enterprise environments, we evaluate how the risk of encountering malware varies across organizations in the different fields. In this case, we consider Information Technology (IT) as a baseline for comparisons when evaluating odds ratios. We measure that machines of firms in the fields of Consumer Staples and Financials show negligible differences with those in the IT segment (1.5% higher likelihood of infection). Overall, we also find that the Healthcare industry is the best sector with the odds ratio with respect to the reference segment being 0.93. On the other hand, firms dealing with Energy, Consumer Discretionary, Utilities, Industrials, Materials and Telecommunications reveal a higher likelihood of encountering malware, We end up to similar conclusions when narrowing the analysis down to specific malware classes.

Summary: An increasing number of hosts does not translate into higher odds of encountering malware. Enterprise size results to be uncorrelated even when narrowing to specific industrial sectors or malware classes. In contrast, organizations operating in Consumer Discretionary, Energy, Materials, and Utilities have higher risk factors than firms related to Information Technology, Healthcare, Financials, Consumer Staples, Industrials, and Telecommunications.

6 DISCUSSION AND CONCLUSION

Home-computer users and enterprises tend to face malware in two different ways: while consumers approach the problem in a reactive fashion, often relying on a single AV product to detect and block possible malware infections, corporations act in a proactive manner, installing multiple security products, activating several layers of defenses, and establishing policies among employees.

In the first part of our work, we investigate whether the different measures in the two environments have an impact on their risks. In other words, we want to answer the question whether more security products, tools, policies and restrictions in the enterprise segment are effective to lower the risk of malware encounters. Globally, we measure for 144.9 M consumer machines and 226.4 M corporate hosts an encounter rate of 9.8% and 12.0% respectively. According to these results, home-machine users encounter slightly less malware than the counterpart, suggesting, at first glance, that all the choices that enterprises adopt are not effective in practice. However, we believe this first impression to be misleading: when considering all the available hosts in our dataset, a lot of them have been found to be active for only a few months, or even a few days, and these low-activity hosts are more prevalent among end-users than corporate machines. When we restrict the two sets of machines to only those active every month of the year, we find an opposite result: around 30% of consumer hosts report malware encounters vs 21% of enterprise machines. If we go one step further and select only those machines that are active more than 20 days and 15 hours per day, the gap widens as 89% of consumers encounter malware against 53% of corporate machines. Moreover, we also found that the average consumer machine encounters a more diverse set of malicious files compared with its corporate counterpart, and this finding holds for all the malware classes considered in our study.

Security policies and restrictions also seem to have a relevant impact on reducing risks. Indeed, when analyzing the presence of Adware and PUAs, we report a concentration of such malware families 6 times higher in consumers, due to the freedom in installing any kind of software that this group of users has. Since the presence of less reputable programs is often a vehicle for malware, we believe the same findings apply also when considering other families. On the opposite, generic behavioral signatures (who might match unknown threats or suspicious files) are twice as likely to trigger in enterprise environments than in consumers hosts.

If on the one hand a good security posture and a better cyber hygiene are important to reduce the risk of malware encounters, on the other hand it is not the only factor to take into account. Indeed, the interconnected nature of our society, the use of third-party software and the sharing of the same networks expose all the classes of machines to undiversifiable and systematic risk, regardless of the number and type of security measures and policies in place.

In the second part of this work, we leverage the data at our disposal to investigate whether this portion of risk exists and provide quantitative indicators that can be used to measure its significance: for this purpose, we extract seven indicators for each consumer machine and nine for each enterprise host that carry no information about its security level, and test their correlation with malware encounter risks. Interestingly, we find that height of them serve this purpose: host uptime days and hours can act as control variables for the encounter rate of a subset of regular machines; with a logarithmic relationship, the same holds for file-based activity; encountering malware over and over and being recidivist along time represents an important risk factor, which is even more pronounced when considering malicious categories as Adware, Virus and Worm; for the same classes, host geographical location can explain the risk of suffering from higher encounter rate; finally, we also verify the effectiveness of vendor number and reputation; For organization environments, we compare industrial sectors and spot those that have higher odds of reporting malicious software; we fail, instead, in proving any correlation between enterprise size and malware encounter rate, even when separately considering each industrial sector.

To our knowledge, no scientific or empirical work has looked at the systematic nature of cyber risks, although the topic is largely discussed in other domains. The existence and quantification of systematic cyber risks is an emerging problem among risk management experts and cyber-insurance underwriters, as the number of events that simultaneously affect a large number of hosts across different enterprises and countries is increasing every year. Hypotheses to explain it have also been advanced considering global-scale incidents and the subsequent market reactions: experts agree that factors and events such as common widespread vulnerabilities, infrastructure failure cascade, loss of integrity of trusted systems, concentrated dependencies and indirect attacks to central actors characterize its nature [7, 8, 11, 31].

Despite these conclusions, systematic risks need a deeper understanding for what concerns their underlying factors and likelihood. An objective analysis of the extent to which these indicators can explain cyber risks would definitely be beneficial for particular tasks, such as premium establishment for cyber insurance policies [32]. Indeed, in order to compute premiums, insurance carriers scale a base rate by factors depending on the enterprise size, industrial sector, and by considering whether or not the company had already suffered cyber security events (i.e., it is a repeat player). In this respect, our study shows that an assessment done considering the enterprise size as a factor may not be appropriate - we find no correlation with malware encounter- and that different indicators are needed to come up with a correct assessment.

Security companies can also benefit from the insights presented in this work. Security vendors can use our analysis for pre-selecting a scanning aggression level of their tools (low, medium, high) based on the factors identified in our work. Currently, end users or companies administrators are asked to make this decision. For example, our results suggest that a more aggressive scanning of devices that are more active or generate a higher file volume is justifiable due to the higher risk. Furthermore, security vendors can adjust the notification level/wording and warn the users who undertake riskier behaviors defined by the indicators of our model. While

currently relying on a reactive approach that tries to match a malicious signature and report it to the user, AVs could proactively notify users about these risky behaviors.

In this work, we try, for the first time, to shed light on systematic risk indicators, by shifting the analysis at the host level and by using real-world data telemetry. With the findings previously discussed in mind, we support the hypothesis that this portion of risk exists in the cyber scenario —in both consumer and enterprise context— and that the factors we identified can be used as good indicators to quantify it. We believe these insights can help both companies and academic researchers to better understand the global picture of malware encounters in the wild, and that our study can be used as a foundation for future works in the area of systematic risk.

ACKNOWLEDGMENT

We would like to thank all the anonymous reviewers for their constructive feedback. This project was supported by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme under grant agreement No 771844 (BitCrumbs).

REFERENCES

[1] Leyla Bilge, Yufei Han, and Matteo Dell'Amico. 2017. Riskteller: Predicting the risk of cyber incidents. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, 1299–1311.

[2] Juan Caballero, Chris Grier, Christian Kreibich, and Vern Paxson. 2011. Measuring pay-per-install: the commoditization of malware distribution.. In *USENIX security symposium*, Vol. 13. The Advanced Computing Systems Association.

[3] A Colin Cameron and Pravin K Trivedi. 2013. *Regression analysis of count data*. Vol. 53. Cambridge University Press.

[4] Davide Canali, Leyla Bilge, and Davide Balzarotti. 2014. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*. 171–182.

[5] Cisco. 2019. Cisco Annual Cybersecurity Report. https://www.cisco.com/c/dam/m/hu_hu/campaigns/security-hub/pdf/acr-2018.pdf. Accessed: 2021-08-09.

[6] John Cloonan. 2017. Advanced Malware Detection - Signatures vs. Behavior Analysis. https://www.infosecurity-magazine.com/opinions/malware-detection-signatures/. Accessed: 2021-08-09.

[7] Shaen Corbet and Constantin Gurdgiev. 2019. What the hack: Systematic risk contagion from cyber events. *International Review of Financial Analysis* 65 (2019), 101386.

[8] Cyber Insurance and Systemic Market Risk 2018. Cyber Insurance and Systemic Market Risk. https://www.eastwest.ngo/sites/default/files/ideas-files/cyber-insurance-and-systemic-market-risk.pdf. Accessed: 2021-08-09.

[9] Savino Dambra, Leyla Bilge, and Davide Balzarotti. 2020. SoK: Cyber Insurance–Technical Challenges and a System Security Roadmap. In *2020 IEEE Symposium on Security and Privacy (SP)*. 293–309.

[10] Savino Dambra, Iskander Sanchez-Rola, Leyla Bilge, and Davide Balzarotti. 2022. When Sally Met Trackers: Web Tracking From the Users' Perspective. In *31st USENIX Security Symposium (USENIX Security 22)*. 2189–2206.

[11] Is Cyber Risk Systemic? 2017. Is Cyber Risk Systemic? https://www.aig.ie/latest-insights/is-cyber-risk-systemic. Accessed: 2021-08-09.

[12] ISO 3166-1 1997. ISO 3166-1. https://en.wikipedia.org/wiki/ISO_3166-1. Accessed: 2021-08-09.

[13] Kaspersky. 2018. Kaspersky Security Bulletin 2018. Threat Predictions for 2019. https://bit.ly/2Wq5eIw. Accessed: 2021-08-09.

[14] Diana Kelley. 2019. Microsoft Security Intelligence Report. https://www.microsoft.com/security/blog/2019/02/28/microsoft-security-intelligence-report-volume-24-is-now-available.

[15] Platon Kotzias, Leyla Bilge, and Juan Caballero. 2016. Measuring PUP Prevalence and PUP Distribution through Pay-Per-Install Services. In *25th USENIX Security Symposium*. 739–756.

[16] Platon Kotzias, Leyla Bilge, Pierre-Antoine Vervier, and Juan Caballero. 2019. Mind Your Own Business: A Longitudinal Study of Threats and Vulnerabilities in Enterprises.. In *The Network And Distributed System Security Symposium (NDSS)*. 739–756.

[17] McAfee Labs. 2018. McAfee Labs Threats Report. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-dec-2018.pdf. Accessed: 2021-08-09.

[18] MalwareBytes labs. 2019. 2019 State of Malware. https://resources.malwarebytes.com/files/2019/01/Malwarebytes-Labs-2019-State-of-Malware-Report-2.pdf. Accessed: 2021-08-09.

[19] Chaz Lever, Platon Kotzias, Davide Balzarotti, Juan Caballero, and Manos Antonakakis. 2017. A Lustrum of Malware Network Communication: Evolution and Insights. In *Proceedings of the IEEE Symposium on Security and Privacy* (San Jose, CA). IEEE Computer Society.

[20] Fanny Lalonde Lévesque, José M Fernandez, and Anil Somayaji. 2014. Risk prediction of malware victimization based on user behavior. In *2014 9th International Conference on Malicious and Unwanted Software: The Americas (MALWARE)*. IEEE, 128–134.

[21] Yang Liu, Armin Sarabi, Jing Zhang, Parinaz Naghizadeh, Manish Karir, Michael Bailey, and Mingyan Liu. 2015. Cloudy with a chance of breach: Forecasting cyber security incidents. In *24th USENIX Security Symposium*. 1009–1024.
[22] Yang Liu, Jing Zhang, Armin Sarabi, Mingyan Liu, Manish Karir, and Michael Bailey. 2015. Predicting cyber security incidents using feature-based characterization of network-level malicious activities. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*. 3–9.
[23] Matplotlib. 2022. Visualization with Python. https://matplotlib.org/.
[24] Ghita Mezzour, Kathleen M Carley, and L Richard Carley. 2015. An empirical study of global malware encounters. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*. 1–11.
[25] Ghita Mezzour, L Carley, and Kathleen M Carley. 2014. Global mapping of cyber attacks. *Available at SSRN 2729302* (2014).
[26] Carina Mood. 2010. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review* 26, 1 (2010), 67–82.
[27] Alexander Moshchuk, Tanya Bragin, Steven D Gribble, and Henry M Levy. 2006. A Crawler-based Study of Spyware in the Web.. In *The Network And Distributed System Security Symposium (NDSS)*, Vol. 1. 2.
[28] Numpy. 2022. The fundamental package for scientific computing with Python. https://numpy.org/.
[29] Pandas. 2022. Python data analysis library. https://pandas.pydata.org/.
[30] PurpleSec. 2019. The Ultimate List Of Cyber Security Statistics For 2019. https://purplesec.us/resources/cyber-security-statistics/. Accessed: 2021-08-09.
[31] Quantifying Systemic Cyber Risk 2018. Quantifying Systemic Cyber Risk. http://web.stanford.edu/~csimoiu/doc/Global_CRQ_Network_Report.pdf. Accessed: 2021-08-09.
[32] Sasha Romanosky, Lilian Ablon, Andreas Kuehn, and Therese Jones. 2017. Content analysis of cyber insurance policies: How do carriers write policies and price cyber risk? *Available at SSRN 2929137* (2017).
[33] Armin Sarabi, Parinaz Naghizadeh, Yang Liu, and Mingyan Liu. 2015. Prioritizing Security Spending: A Quantitative Analysis of Risk Distributions for Different Business Profiles.. In *WEIS*.
[34] Scikit-learn. 2022. Machine Learning in Python. https://scikit-learn.org/stable/.
[35] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. 2018. Predicting impending exposure to malicious content from user behavior. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 1487–1501.
[36] StatCounter. 2022. Desktop Operating System Market Share Worldwide. https://gs.statcounter.com/os-market-share/desktop/worldwide.
[37] Moore Susan and Keen Emma. 2018. Gartner Forecasts Worldwide Information Security Spending to Exceed $124 Billion in 2019. https://gtnr.it/2zQUueM. Accessed: 2021-08-09.
[38] Symantec. 2019. Internet Security Threat Report. https://docs.broadcom.com/doc/istr-24-executive-summary-en. Accessed: 2021-08-09.
[39] Olivier Thonnard, Leyla Bilge, Anand Kashyap, and Martin Lee. 2015. Are you at risk? Profiling organizations and individuals subject to targeted attacks. In *International Conference on Financial Cryptography and Data Security*. Springer, 13–31.
[40] OMERS Ventures. 2019. Cybersecurity: Industry Overview, Market Map, Global Investments. https://bit.ly/2L52hbn. Accessed: 2021-08-09.
[41] W3techs. 2022. Usage statistics of operating systems for websites. https://w3techs.com/technologies/overview/operating_system.
[42] Ting-Fang Yen, Victor Heorhiadi, Alina Oprea, Michael K Reiter, and Ari Juels. 2014. An epidemiological study of malware encounters in a large enterprise. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 1117–1130.

A COMBINING CONSUMER AND ENTERPRISE MACHINES INTO A SINGLE GLM

Table 11 reports the odds ratios obtained by combining the two set of machines (consumers and enterprises) with an added regressor (machine_type: 0 = consumer, 1 = enterprise) in addition to the 7 already considered (active days and hours, file-request volume, reputation and number of installed vendors, geographical location and whether or not malware has already been detected on the machine the month before).

Table 11. Odds ratios of encountering malware according to our regression models

Host Attribute	Bin Category	Malware family	Monthly Odds μ	Monthly Odds σ
Activity Days	4-8	Any	2.04	0.17
	8-12	Any	2.68	0.40
	12-16	Any	3.14	0.59
	16-20	Any	3.63	0.84
	20-24	Any	3.82	0.98
Ref: [0-4]	24-28	Any	3.96	1.10
	28+	Any	3.97	1.18
Activity Hours	3-6	Any	1.31	0.09
	6-9	Any	1.50	0.30
	12-15	Any	1.09	0.40
	15-18	Any	1.14	0.31
	18-21	Any	1.29	0.39
Ref: [0-3]	21+	Any	1.89	0.79
	18-21	Adware	1.58	1.22
	21+	Adware	3.06	2.06
File-volume Activity	1K-2K	Any	1.00	0.07
	3K-4K	Any	1.63	0.32
	5K-10K	Any	2.13	0.50
	10K-50K	Any	2.98	0.93
	50K+	Any	4.05	0.80
Ref: [0-1K]	10K-50K	Adware	9.52	3.72
	50K+	Adware	13.41	4.24
Vendors	20-40	Any	1.02	0.04
	40-60	Any	1.13	0.07
	60+	Any	1.41	0.08
Ref: [0-20]	60+	Adware	1.40	0.31
	60+	PUP	1.60	0.09
Reputable vendors only	Yes	Any	1.00	0.04
	Yes	PUP	0.97	0.05
Ref: No	Yes	Virus	0.63	0.03
Repeat player	Yes	Any	1.88	0.84
	Yes	Adware	8.80	3.24
	Yes	Virus	2.46	0.94
Ref: No	Yes	Worm	10.13	2.13
Geographical location	AF	Virus	13.42	2.07
	AS	Virus	4.81	0.55
	AF	Worm	20.52	2.35
	AS	Worm	5.61	0.20
Ref: NA	OC	PUP	0.86	0.17
	OC	Trojan	1.02	0.09