Understanding the Relationship between Human Behavior and Susceptibility to Cyber-Attacks: A Data-Driven Approach

Michael Ovelgönne, UMIACS, Univ. of Maryland, College Park.
Tudor Dumitras, Dept. of Elect. Eng. and UMIACS, Univ. of Maryland, College Park.
B. Aditya Prakash, Dept. of Computer Science, Virginia Tech., Blacksburg.
V. S. Subrahmanian, Dept. of Computer Science and UMIACS, Univ. of Maryland, College Park.
Benjamin Wang, Dept. of Computer Science, Virginia Tech., Blacksburg.

Despite growing speculation about the role of human behavior in cyber-security of machines, concrete data-driven analysis and evidence has been lacking. Using Symantec's WINE platform, we conduct a detailed study of 1.6 million machines over an 8-month period in order to learn the relationship between user behavior and cyber attacks against their personal computers. We classify users into 4 categories (gamers, professionals, software developers, others plus a fifth category comprising everyone) and identify a total of 7 features that act as proxies for human behavior. For each of the 35 possible combinations (5 categories times 7 features), we studied the relationship between each of these 7 features and one dependent variable, namely the number of attempted malware attacks detected by Symantec on the machine. Our results show that there is a strong relationship between several features and the number of attempted malware attacks. Had these hosts not been protected by Symantec's anti-virus product or a similar product, they would likely have been infected. Surprisingly, our results show that software developers are more at risk of engaging in risky cyber-behavior than other categories.

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems**]: Human factors; K.6.5 [**Security and Protection**]: Invasive software; H.2.8 [**Database Management**]: Database Applications–*Data mining*

Additional Key Words and Phrases: malware; computer virus; user behavior

1. INTRODUCTION

As is well known, cyber security systems based on rigorous theoretical proofs often fail in practice. In fact, many researchers consider human users to be the weakest link in the system [Anderson 1993; Whitten and Tygar 1999; Clark et al. 2011]. Though cyber-security has become an increasingly important problem, and despite much speculation about how human behavior leads to vulnerability, the role of human behavior in the vulnerability of machines has not been thoroughly studied in large-scale operational settings. In ordinary crime such as muggings, a victim is often selected by the criminal on the basis of his behavior and/or characteristics (e.g. walking late at night, being old and infirm, etc.). In the same way, the likelihood and intensity with which a machine is attacked by malware has been hypothesized to be related closely to the behavior exhibited on that machine [Stajano and Wilson 2011]. "Only amateurs attack machines; professionals target people" [Schneier].

Quantifying the behaviors that are likely to attract cyber attacks is important for coping with the growth and diversity of cyber attacks. Humans can fall prey to social engineering attacks which request them to visit Web sites or to download files that result in the

Author Email: {mov, tdumitra, vs}@umiacs.umd.edu; {badityap, benwang}@cs.vt.edu.

DOI:http://dx.doi.org/10.1145/0000000.0000000

installation of malware; they may also visit compromised Web sites that conduct drive-by-download attacks [Grier et al. 2012] which exploit vulnerabilities in browsers causing silent file downloads. In short, users often willingly download unknown applications and binaries or unknowingly perform actions that undermine security.

Nevertheless, the extent to which user behavior is related to the propensity of a host to be the target of cyber attacks is not well understood. There are excellent recent papers that conduct carefully designed experiments in order to understand the link between human behavior and specific kinds of susceptibility to attacks [Johnston and Warkentin 2010; Ifinedo 2012; Crossler et al. 2013; Sheng et al. 2010]—unfortunately though, these studies are typically in artificial lab settings with a small number of participants. For instance, [Johnston and Warkentin 2010] studies a cohort of 215 subjects in order to understand the reactions provoked by emails designed to induce fear on the part of the user. [Ifinedo 2012] looks at compliance with cyber-security procedures by integrating the theory of planned behavior (TPB) and protection motivation theory (PMT). They test hypotheses linking subjective norms with security protocols, attitudes towards security policies and compliance, self-efficacy and compliance, response cost and compliance, response efficacy and compliance, and severity of consequences and compliance. [Sheng et al. 2010] conducts a survey-based study of 1001 individuals to identify conditions under which the individuals fall for phishing attacks. However, as noted in [Crossler et al. 2013], the lack of hardcore operational data poses a threat to the validity of these studies, especially as they have very small "N's".

This paper focuses on the first-ever analysis of real-world *operationally gathered* cyber-security data about human behavior. More specifically we study the problem of *identifying human behaviors which increase the risk of malware attacks on a host based on real-world operational data.* Real-world cybersecurity companies such as Symantec, Kaspersky Labs, and others that deploy products have a strong interest in understanding and mitigating the risk of individual users around the world. Prior to this research, Symantec had created the Worldwide Information Network Environment (or WINE) dataset [Dumitraş and Shou 2011][1] and encouraged computer science researchers (such as the authors) to mine this data for interesting findings. In particular, using the WINE data, we systematically analyze the user behaviors and cyber attacks observed between January–August 2011 on 3.5 million end-hosts (which we downselected to 1.6M by a mechanism that we will describe later in Section 3).

Note that it was impossible for us to: (i) change the data collection method and/or (ii) run controlled trials/experiments. This is because the data is real-world operational data collected by Symantec's Norton Anti-Virus product which is a profit making operation where changes could potentially affect millions of users worldwide, and because opt-in permissions for data collection had already been gathered by Symantec from millions of users. We therefore did the best with the data that we were provided with, in answering our questions.

WINE's *binary reputation* data set includes information on binary executables downloaded by users who opt in for Symantec's reputation-based security program. The *anti-virus telemetry* data set includes reports about host-based threats (e.g., viruses, worms, trojans) detected by Symantec's anti-virus products. Because this data is collected on hosts targeted by cyber attacks—rather than honeypots or survey instruments or small-scale lab settings—it provides a unique window into the factors affect security of real computer users worldwide.

We identify several features that point to specific human behaviors, and we analyze how the risk of cyber attacks changes with different behaviors. As WINE contains no information that would allow us to identify users, we assume that each host in our data corresponds to one user and we assess the user's behavior anonymously, from the events recorded on that

[1] WINE data, as well as data sets created at Symantec by specific researchers such as us, is available for academic study through Symantec after an approval process controlled by Symantec.

host. We estimate the risk of attacks using the frequency of malware detections on each host. As all the hosts in our data set were protected by Symantec products, the observed attacks were actually blocked, but had the machine not had an anti-virus program installed, many attacks would have succeeded. Thus, our measurement of risk is actually directed at the population of machines *not* protected by any anti-virus program.

For each host h, WINE's binary reputation data included information on: the number of binaries on host h, the number of unsigned binaries on host h, the number of unique binaries on host h, the number of downloaded binaries on h, the number of low-frequency and high-frequency binaries on h. These quantities were captured by Symantec's daily operational experts (that included Tudor Dumitras, a co-author of this paper) who recorded this data based on seeing thousands of cyber-attacks on a day in - day out basis and based on their interaction with other cybersecurity experts. The number of binaries on a host has always been viewed as significant in the cyber-security industry – for instance, a Georgia Tech report [Ahamad et al. 2008] with extensive industry input reports numbers of binaries in specific attacks such as RAT-SZ-1, Sality-1, Poebot-1, Kraken, amongst others. The report additionally quotes Georgia Tech Professor Wenke Lee stating that "several recent bot variants have exhibited more than 100 distinct binary payloads used to hide the communications path and to vary the command and control IP address." This suggests that the number of binaries may have some link to cyber-security and cyber-vulnerability of a host. Similarly, [Niki 2009] examines the importance of drive-by-downloads in disseminating malware and studies methods to detect such drive-by download activity – likewise, [Provos et al. 2007] present ample evidence that drive by downloads are a common malware distribution mechanism. In a similar vein, unsigned binaries have long been thought to pose a cyber-security risk by researchers in the cyber-security industry – [Niemelä 2010] discusses such a perspective from industry leader F-Secure. Likewise, in a CTO round-table, cyber-security experts from different companies stated that "Ad hoc experiments have shown that unsigned applications from well-known server sites carry a risk of infection" [Creeger et al. 2010], a hypothesis that needs more careful validation. Likewise, users all over the world who travel tend to log in to the Internet from airports, hotels, and cafes. The idea that connecting from such "public" or semi-public networks is linked to increased cyber-risk has long been felt. For instance, [Hu et al. 2009] studies how malware can spread through such networks.

These quantities measured in the WINE data capture various behaviors. The number of binaries captures the tendency of a user to install binaries, either intentionally or otherwise, on his machine. The number of unique and low-frequency binaries on a host indicates the willingness of the user to install less popular software. The number of unsigned binaries says something about the user's risk taking behavior - savvy users prefer to install software from respected vendors who sign their software. The number of downloaded binaries is a proxy for the types of websites the user is visiting. In addition to WINE's binary reputation data, we also looked at the number of ISPs that a host machine logged in from. Connecting to wifi networks poses a bigger risk [Henry and Luo 2002]—therefore, the number of ISPs that a user logs in from is a proxy for the user's travel habits.

Table I summarizes our key findings showing that for all categories of users, the number of low-prevalence binaries downloaded by the users, number of unique binaries on users' machines, number of unsigned binaries on the users' machines, and number of binaries downloaded by users all increase the number of malware attacks. In the case of software developers, the number of binaries they installed on their machine is also related to the number of attacks. Finally, for the number of ISPs that they access to connect to the network, we saw a statistically significant influence on the number of attacks. However, we think the magnitude of that influence is too low to claim a relationship given the potential sources of error we discuss later.

Table I. Overview of analyzed independent variables.

Ind. Variable	gamer	pro	SW-dev	Other	All
# Binaries	no	no	yes	no	no
% Low-Freq Bin.	yes	yes	yes	yes	yes
% Hi-Freq Bin.	yes	yes	yes	yes	yes
% Unique Bin.	yes	yes	yes	yes	yes
% Unsigned Bin.	yes	yes	yes	yes	yes
% Downloaded Bin.	yes	yes	yes	yes	yes
# of ISPs	no	no	no	no	no

Implications for the security industry. It is increasingly difficult to protect users against malware because of the growth in volume and diversity of cyber attacks; characterizing the user behaviors that are more likely to attract cyber attacks opens up new opportunities for identifying and defending the hosts that are at risk. For example, security analysts estimate that 403 million new malware samples were created in 2011 [Symantec Corporation 2012]. This growth results in a large number of low-prevalence files, which are present on few hosts and are likely to be malicious, as attackers employ polymorphism techniques in order to evade detection. This observation represents the basis of recent *reputation-based security* techniques [Chau et al. 2010; Rajab et al. 2013], which compute a reputation score for each unknown file based on features such as the file's prevalence in the wild, before analyzing the content of the file. Today, reputation-based security systems are included in several anti-virus products, as well as in the Windows 8 operating system [Cowan 2013]; however the association between the users' propensity to download low-prevalence files and cyber attacks has not been validated at a large scale.

Similarly, security best practices recommend reducing the *attack surfaces* of end-hosts [Manadhata and Wing 2011]. Attack surface reduction works by decreasing the number and severity of potential attack vectors that each host exposes (e.g., open sockets, RPC endpoints, running services). Even if the software contains vulnerabilities—perhaps not yet discovered—the attacks will succeed only if a corresponding attack vector is available. However, users can alter the attack surfaces of their computers by downloading and installing new software, which may enable additional attack vectors. By helping analysts understand how the number of binary executables present on a host affects the volume of cyber attacks, our work allows them to assess the impact of attack-surface reduction techniques on security in the field.

Roadmap. The paper is organized as follows: we review related work and then describe our dataset, and then the statistical features signifying human behaviors of interest. We then present our approach and hypotheses. Finally we conclude with implications of our findings.

2. RELATED WORK

We considered related work in cyber-security and data mining. However, to the best of our knowledge, we are the first to do an in-depth analysis of a complete (though cleaned) data set spanning an 8 month period.

2.1. Human Factors in Security

A growing body of research points to the importance of human behavior in creating security products. Anderson [Anderson 1993] first observed that strong cryptographic protocols do not usually fail because of errors in their mathematical underpinnings, but are overcome in practice owing to the errors of human users and operators. Whitten and Tygar [Whitten and Tygar 1999] and Clark et al. [Clark et al. 2011] revisited this question 6 and 18 years later, respectively, and found that human errors continue to be an important source of security failures. Leach [Leach 2003] found that "as many as 80% of major security failures could be the result of not poor security solutions but of security behavior...". Abraham et. al. [Abraham and Chengalur-Smith 2010] specifically studied "social-engineering" malware, which adopt a combination of psychological and technical ploys, with the eventual goal of luring a computer user to execute the malware. Visualization tools for security analysts is also an active research area [Nataraj et al. 2011].

Carlinet et al. [Carlinet et al. 2008] analyzed the network traffic of ADSL users to identify risk factors. They identified the usage of web and streaming increases the infection risk while for peer-to- peer and chat applications usage no such link could be established. Very recently, Lévesque et al. [Lalonde Lévesque et al. 2013] conducted a field study where they installed monitoring software on the computers of 50 subjects to identify risk factors for malware attacks. They analyzed mainly the types of websites their subjects visited (e.g. mp3/streaming, sport, gambling, illegal). Our analysis overlaps with theirs in two points: number of applications/binaries and computer expertise. Like them we identified a significant relationship between attacks and application/binaries. However, in our *large-scale analysis* we also see that although the relationship is statistically significant the influence is low. Interestingly, using completely different methods, with respect to computer expertise, our results for SW-developers validate the observation of [Lalonde Lévesque et al. 2013] that higher computer expertise is a risk factor.

The Fear Appeal Manipulation model[Johnston and Warkentin 2010] tested whether fear based manipulation of users (e.g. by posting phishing messages where a user is threatened with dire consequences such as imprisonment if he does not follow instructions) is linked to perceptions of threat severity, threat susceptibility, self-efficacy, and response efficacy. Their study looks at 215 users in all. [Ifinedo 2012] looks at compliance with cybersecurity procedures by building on and integrating two psychological theories - the theory of planned behavior (TPB) and protection motivation theory (PMT). The author tests hypotheses linkiing subjective norms and compliance with security protocols, attiitudes towards security policies and compliance, self-efficacy and compliance, response cost and compliance, response efficacy and compliance, severity of consequences and compliance. A specific phishing-based study [Sheng et al. 2010] conducts a survey-based study of 1001 individuals to identify conditions under which the individuals fall for phishing attacks. [Crossler et al. 2013] discusses possible future work in behavior-based information security - they state "Gaining access to individuals' actual behavior is one consistent challenge for Behavioral InfoSec research", and suggest a number of methods to achieve this. They later state that corporate data can be very valuable, but that "this access can prove to be elusive as gaining access to corporate data, especially security data, can be a difficult or virtually impossible." We believe this paper follows upon their suggestion, at least to the extent that we have gained access to a huge trove of corporate data.

2.2. Data Mining for Security

Much research has tried to model malware propagation. Staniford et. al. [Staniford et al. 2002; Staniford et al. 2004] analyzed the Code Red worm traces and proposed an analytical model for its propagation. They also argue that optimizations like hit-list scanning, permutation scanning can allow a worm to saturate 95% of vulnerable hosts on the Internet in less than 2 seconds. Papalexakis et. al. [Papalexakis et al. 2013] propose the SharkFin and GeoSplit models of spatio-temporal propagation of malware based on an analysis of the WINE data. Their system models only the total volume of malware attacks as a whole over time without considering the human behavioral aspect. In contrast, in this paper we model the *magnitude* of malware attacks *per machine* in context of the *machine usage* by humans. As such, our work can be also thought of as providing a fine-grained picture of human behavior characteristics that seem to be related to increased vulnerability to malware.

3. DATASET AND SET-UP

To characterize the link between human behavior and cyber attacks, we integrated information in several data sets collected from different observation perspectives. We describe our problem statement, and the datasets we used (including associated caveats) for our research in this section.

3.1. Problem Statement

More specifically, our research problem can be defined as follows:

GIVEN: Security telemetry from Symantec's WINE datasets (details in Section 3.2).

FIND: The statistical proxies of human behaviors which are related to increased malware reports on a machine.

Clearly, addressing this problem involves (a) extracting carefully constructed data-based features from the WINE datasets, and then (b) performing sound statistical tests to relate the *independent* behavioral variables (the features) to the *dependent* variable (the number of malware detections).

3.2. The WINE Datasets

Symantec's WINE data is collected from real-world hosts running their consumer anti-virus software. Users of Symantec's consumer product line have a choice of opting-in to report telemetry about the security events (e.g. executable file downloads, virus detections) that occur on their hosts. The events included in WINE are representative of events that Symantec observes around the world [Papalexakis et al. 2013]. WINE enables reproducible experimental results by archiving the reference data sets that researchers use and by recording information on the data collection process and on the experimental procedures employed.

We analyze the complete set of events recorded in the *binary reputation* and *anti-virus telemetry* data sets from WINE during the 8 month period between January–August 2011. For this 8-month period, the dataset contains 13.7 billion reports collected on 3.5 million hosts. WINE does not include user identifiable information.

Anti-virus telemetry. Anti-virus (AV) telemetry records detections of known malware for which Symantec generated a signature that was deployed in an anti-virus product. As commercial security products generally aim for low false-positive rates, we have a high degree of confidence that the files detected in this manner are indeed malicious. From each record, we use the detection time, the associated threat label, the hash (MD5 and SHA2) of the malicious file detected and the manner of the detection (signature scanning or behavioral features extracted from an execution of the file on the end host). Each record indicates that the anti-virus has blocked an attack that may have resulted in an infection.

Binary reputation. The binary reputation data records all binary executables – benign or malicious – that were downloaded/copied on end-hosts worldwide. From each record we extract the time stamp of the file creation event, the country in which the host is located, the hash (MD5 and SHA2) of the binary, and the URL from which it was downloaded (if available).

Data cleaning. First, we restrict our analysis to the 20 countries with the most hosts in the dataset. That left 2.9 million machines. We also removed any machine that was active for less than 200 days in the 8-month period. We take the range of activity as a proxy for potential virus threats. After applying the 200 day filter rule, 1.7M machines remained. The percentage of machines that did send reports for the full period studied seems to be very high. However, it seems very likely that either many hosts revoked the data sharing opt-in at some point or replaced the Norton software with something else (or nothing) after their license expired. Some hardware vendors sell computers with pre-installed trial versions of Symantec's software. Conversion rates from trial users to paying customers are usually low

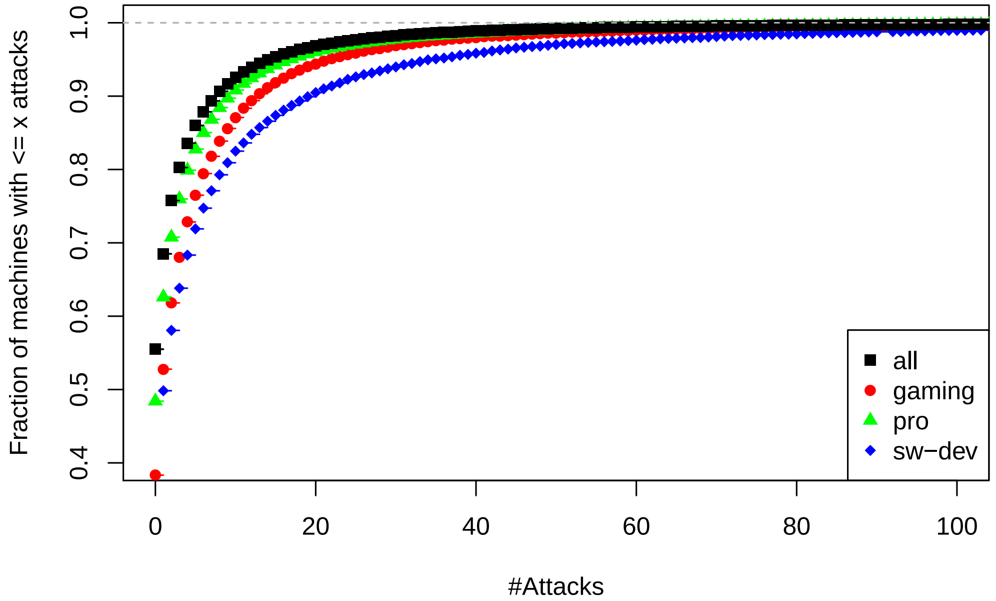


Fig. 1. Empirical cumulative density of the number of malware attacks per machine shown by type of user machine. We see that the worst affected category of user machine is software developers, followed by gamers, followed by professionals.

for any type of product or service. We removed outliers (hosts that have more binaries or attacks that are more than 2 standard deviations away from the mean) we ended up with a *cleaned dataset* of 1.6M hosts — all our results were derived from this cleaned data.

Figure 1 shows the distribution of number of malware found per host. Encouragingly, more than 50% of hosts in the cleaned data encountered no malware during our observation period. Most machines had 50 attacks or less.

Shortcomings of our study. As WINE does not include telemetry from hosts without Symantec's anti-virus products, our results may not be representative of the general population of platforms in the world. In particular, users who install anti-virus software might be more careful with the security of their computers and, therefore, might be less exposed to attacks. Additionally, our data was collected on hosts running various versions of Windows; the trends we observe may not apply to other operating systems. Although we cannot rule out the possibility of such selection bias, the large size of the population in our study (3.5 million hosts) and the fact that Windows has been the primary target for cyber attacks for the past decade suggest that our results have a broad applicability. The anti-virus applications that gather WINE data operate on end-hosts. Hence, we do not know how many attacks are deflected by security measures in the environment (e.g., a hardware firewall, or intrusion-prevention services provided by an Internet Service Provider) or by operating system defenses that sit in front of Symantec's software. Therefore, the number of attacks observed in the anti-virus telemetry data should be interpreted as a lower bound.

Finally, the WINE data sets do not provide sufficient information for determining the rate of successful infection on the targeted hosts. However, by correlating the attacks blocked and the files present on each host in our data set, we can derive unique insights on the link between human behavior and susceptibility to cyber attacks.

4. FEATURE CONSTRUCTION

The goal of this paper is to analyze human behaviors w.r.t. cyber security. Hence an important step is constructing statistical features based on the WINE datasets, which can act as proxies for human behaviors. To this end, we analyzed the following features (i.e. the 'independent variables'):

(1) **Number of binaries present on a host.**
DEFINITION: The number of executable files on a machine.
MOTIVATION: The total number of executables represents a measure of the host's attack surface because each executable file may include known or unpatched vulnerabilities and may provide distinct attack vectors.
ASSOCIATED BEHAVIOR: installation of software. The human user can increase this attack surface by downloading/installing more executables.

(2) **Percentage of low-prevalence binaries on a machine.**
DEFINITION: A low-prevalence binary is one that is present in less than 1000 hosts in our cleaned WINE dataset. The percentage of low-prevalence binaries is the ratio of number of low-prevalence binaries to number of binaries in the cleaned WINE dataset.
MOTIVATION: Hackers often create numerous minor polymorphic variants of malware in order to evade detection—these are low-prevalence files. This is one of the key observations behind modern reputation-based security technologies [Chau et al. 2010; Rajab et al. 2013].
ASSOCIATED BEHAVIOR: Some users have a tendency to accumulate low prevalence binaries (e.g. by downloading "cheats" in online games which are often infected with malware [Bono et al. 2009] or by downloading free software).

(3) **Percentage of high-prevalence binaries on a machine.**
DEFINITION: A high-prevalence file is one that is present in over 1M hosts in our cleaned WINE dataset.
MOTIVATION: We studied high-prevalence files solely in order to complement our study of low-prevalence files.
ASSOCIATED BEHAVIOR: Some users have a tendency to download popular binaries, perhaps binaries that lots of their friends are downloading such as new social network apps.

(4) **Percentage of unique binaries**
DEFINITION: A binary that appears on only one host represents an extreme case of low prevalence.
MOTIVATION: As discussed above, a unique binary may be a polymorphic variant of a piece of malware.
ASSOCIATED BEHAVIOR: A user who downloads unique binaries is one who is downloading low-prevalence files as mentioned above. We note, however, that software vendors sometimes create unique binaries by embedding digital watermarks or customer-specific information, for licensing purposes and some unique binaries may fall into this category.

(5) **Percentage of unsigned binaries on a machine.**
DEFINITION: Commercial software vendors usually sign their binaries digitally to verify the integrity of software and to establish the identity of the vendor. This feature measures the percentage of binaries on a host that is unsigned. Lack of a digital signature does not necessarily mean that the binary is malicious; for example, open-source software is typically distributed without being signed.
MOTIVATION: As unsigned binaries do not have a reputed entity affirming the integrity of the binary, one may hypothesize that these binaries are more likely to be malware.
ASSOCIATED BEHAVIOR: A user with a high percentage of unsigned binaries exhibits a tendency to "go with" fewer well known software vendors. This may indicate that

he cares less about the reputations of the vendors whose software he installs on his machines.

(6) **Percentage of downloaded binaries**
DEFINITION: Percentage of binaries that the user downloads from the web, intentionally or otherwise.
MOTIVATION: Malware is often distributed via the Web [Grier et al. 2012].
ASSOCIATED BEHAVIOR: Users who download a high percentage of binaries from the web may be visiting more questionable sites, especially if these downloaded binaries are not signed. Note that the WINE dataset does not allow us to track binaries downloaded through means such as email attachments or copying them from a physical medium, such as a CD-ROM or USB drive.

(7) **Travel history of a user**
DEFINITION: This is the number of ISPs from which a host has connected to the network.
MOTIVATION: In a time when many users use laptops, tablets, and smartphones for their computing needs, there is a high probability that these machines "travel". People carry laptops from home to work to conference sites, airports, and hotels. At each such site, their machine may connect to a local, less secure, ISP.
ASSOCIATED BEHAVIOR: Individuals who feel an absolute need to be connected through free Wifi networks may value connectivity more than security. By using the number of ISPs that a machine connects to as a proxy for a user's travel habits, we wondered whether amount of travel by a user can be linked to the risk of malware attacks on the machine.

5. USER CLASSIFICATION

We classify (anonymized) users into four categories, based on the application programs present on their computers. This is because cyber-attackers may target different categories of users by exploiting vulnerabilities in software primarily used for professional tasks (e.g. SAP) or by taking advantage of the fact that some users download more binaries than others (e.g. for gaming). As our curated WINE data set includes 352.8 million binaries, we use the manufacturer information provided by the certificates used to digitally sign binaries. We identify a total of 902 software development companies that had signed at least 100K reported binaries. We then classify all users via four categories: `gamer, pro, SW-dev, other` describing gamers, professionals (other than software developers), software developers, and all others, as follows:

Pros. A host with more than 50 binaries from companies that only manufacture professional software is considered a professional host.

Gamers. A host with more than 50 binaries from companies that only manufacture games is considered a gamer.

Software Developers. We created a list of popular tools (Table II) and classified any machine having any of these tools as a software development host.

Other. All hosts that have not be classified in one or more of the aforementioned categories.

Overall, 14.8% of users were classified as Gamers, 27.1% as Pros, 0.6% as Software Developers and the rest were classified as Other. Note that these criteria may classify some hosts into more than one category. We chose this approach because, in real life, users may also utilize one computer for multiple purposes, e.g. a computer science student who completes programming assignments and plays games on the same laptop. The overlaps though, for each pair of categories, were miniscule: only 0.1% of users were both Gamers and Software Developers, 0.3% of users were both Software Developers and Pros, and 5.9% of users were both Gamers and Pros.

Table II. Definition and examples for classification of software vendors and software

Category	Description	Examples
Professional	Software vendors whose products are only used in professional contexts and have no dual consumer/business use like office packages, examples are vendors for ERP, CAD, data center or call center software	SAP, EMC, Sage Software, Autodesk, Dassault Systemes, Citrix, TiFiC AB
Gaming	Every software vendor that publishes only gaming software including more traditional video games and/or social entertainment and virtual worlds, no multi-product companies like Microsoft	Valve, Electronic Arts, Blizzard, Duowan, Epic Games, WildTangent, Jorudan, IMVU
Software Development	Software used in the software development process including compilers, IDEs, version control systems	VisualStudio, Eclipse, NetBeans, Java SDK, Subversion, Git, Mercurial

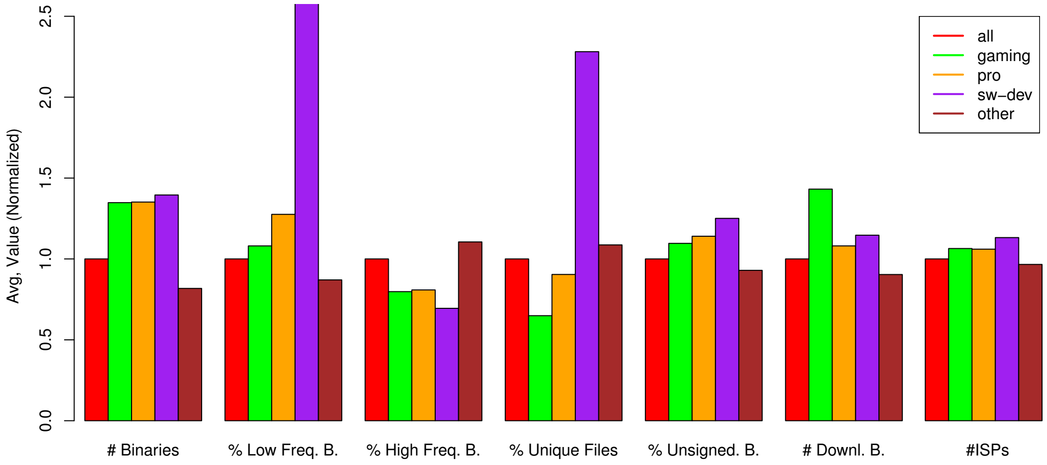


Fig. 2. Independent variable averages by user category (normalized, values for `all` set to 1).

Figure 2 shows the average values of the features described in Section 4 by user category. We normalize these values by the averages for all users in each category, to highlight the behaviors that are typically associated with each user category. The most striking deviation from the average values of `all` machines are the high fraction of low-prevalence and unique files of software developers. This is not surprising. When a software developer compiles code, a new and most likely unique binary is created. Additionally, Figure 2 shows that `gamers` have a higher than normal number of downloaded binaries on their computers and `SW-devs` spend an above average fraction of time working at night.

This user classification allows us to make the first observation about human behavior and cyber attacks.

Observation 1. On average, gamers encountered 83% more malware attacks than non-gamers, while professional users encountered have 33% more malware attacks than non-professional users.

As gamers have the habit of downloading more binaries from the Web than other users, this puts them at risk. Malware that targets popular gaming platforms (e.g. with the aim of stealing World of Warcraft credentials [Group 2013]), further raises the risk profile of gamers. The higher risk of cyber attacks against professional users reflects the recent increase in targeted attacks aimed at stealing sensitive/proprietary information from corporations [O'Gorman and McDonald 2012; Mandiant 2013]. As the ability to remain stealthy (e.g. through the use of zero-day exploits) is key in these attacks, they typically target only a few selected employees and do not result in a large volume of malware aimed at professional users.

OBSERVATION 2. The amount of malware present on software development hosts is significantly higher than on non-software development hosts. On average, software development hosts (`SW-dev`) have 8.1 pieces of malware on them, compared with just 3.3 on non-`SW-dev` hosts (3.3). Thus `SW-dev` hosts have approximately 2.5 times the amount of malware when compared to non-`SW-dev` hosts (3.3).

We do not know the reason for this. Perhaps some software developers build tools to analyze malware, or perhaps they learn how to develop exploits, or perhaps they participate in vulnerability rewards programs. However, as the group of users most intimately familiar with the inner workings of computer systems, it is also possible that developers find ways around the restrictions imposed by firewalls and anti-virus software (e.g. in order to deliver a project on time) and they may also disregard security best practices that are aimed at regular users, by downloading risky binaries of unknown provenance. Some software developers may mistakenly believe that their knowledge of software arms them with better protection against cyber attacks than automated tools or simple rules of thumb could. Because the WINE data reflects attacks rather then infections (see Section 3.2), we do not know whether this is indeed true, or, conversely, whether their knowledge lulls software developers into a false sense of security. This could be the topic of future work. However, the WINE data shows that software developers attract considerably more cyber attacks than other users.

6. USER BEHAVIOR AND CYBER ATTACKS

In order to analyze the relationship between each independent variable (IV; the features of Section 4) and the number of observed attacks, we first learned a quasi-Possion regression model with a sqrt-link function to measure the validity of the connections between attacks and IVs. [2] When we fit models separately for all 7 IVs, we get p-values of less than 2e-16 for all independent variables. This indicates that each feature is statistically correlated to the number of attacks a host receives. In our multivariate model with all 7 features considered together, we get the following p-values, residuals, errors shown in Figure 3.

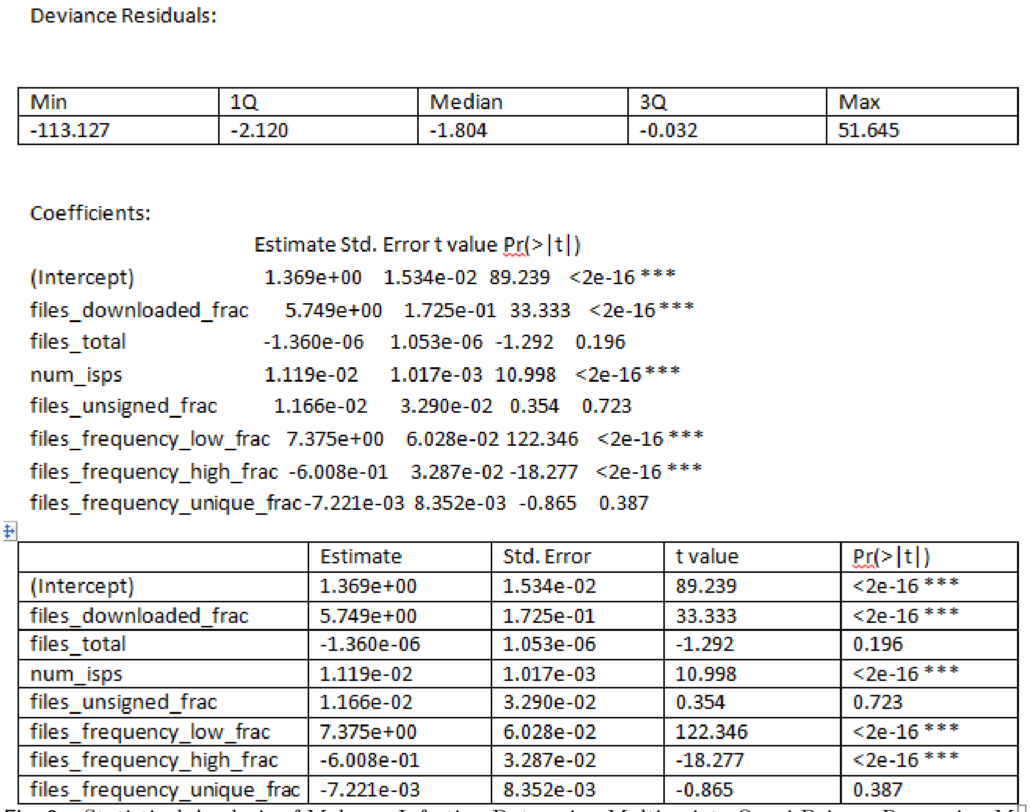
This indicates that the features for unique and unsigned binaries do not provide additional information when used in combination with the other features, even though on a univariate basis, those features are significant.

This result shows that when we aggregate the data and group hosts by their feature values, we see clear trends as we will discuss in the remainder of this section. Moreover, our subsequent analysis (reported below) shows that with data from more then 1.6 million hosts, it is highly unlikely that any observed pattern is random.

6.1. Analysis methods

We analyze the behavioral features in three ways which are better suited to test our hypotheses. For a given IV X, we provide *Median Tables* and *Decile Plots*, which sort all machines in ascending order by the value of X. Think of this as generating a line, $sort(X)$ that lines up all hosts from left to right with the left-most host having the lowest value of X and the right most having the highest value of X.

[2] We tried several other regression models and chose the best one, via. the quasi-Possion regression model with a sqrt-link function.



Deviance Residuals:

Min	1Q	Median	3Q	Max
-113.127	-2.120	-1.804	-0.032	51.645

Coefficients:

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.369e+00   1.534e-02  89.239   <2e-16 ***
files_downloaded_frac      5.749e+00   1.725e-01  33.333   <2e-16 ***
files_total               -1.360e-06   1.053e-06  -1.292   0.196
num_isps                  1.119e-02   1.017e-03  10.998   <2e-16 ***
files_unsigned_frac        1.166e-02   3.290e-02   0.354   0.723
files_frequency_low_frac  7.375e+00   6.028e-02 122.346   <2e-16 ***
files_frequency_high_frac -6.008e-01   3.287e-02 -18.277   <2e-16 ***
files_frequency_unique_frac-7.221e-03 8.352e-03  -0.865   0.387
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.369e+00	1.534e-02	89.239	<2e-16 ***
files_downloaded_frac	5.749e+00	1.725e-01	33.333	<2e-16 ***
files_total	-1.360e-06	1.053e-06	-1.292	0.196
num_isps	1.119e-02	1.017e-03	10.998	<2e-16 ***
files_unsigned_frac	1.166e-02	3.290e-02	0.354	0.723
files_frequency_low_frac	7.375e+00	6.028e-02	122.346	<2e-16 ***
files_frequency_high_frac	-6.008e-01	3.287e-02	-18.277	<2e-16 ***
files_frequency_unique_frac	-7.221e-03	8.352e-03	-0.865	0.387

Fig. 3. Statistical Analysis of Malware Infection Data using Multivariate Quasi-Poisson Regression Model with a sqrt-link function.

Median tables. In the median tables, we divide the $sort(X)$ line at the median value for X, and we show the average number of attacks per host for the hosts to the left of the median (lower 50% of hosts) and the ones to the right of the median (upper 50% of hosts). For instance, Median Table III shows that hosts with fewer than the median number of binaries receive 2.9 attacks on average, while hosts with an above-median binary count receive 3.7 attacks on average.

Decile plots. These plots split $sort(X)$ into 10 equal-sized deciles from left to right. The first decile has the 10% of machines with the lowest values of the IV, while the tenth decile has those with the top 10%.

KD-plots. Decile plots line up hosts by the value of an independent variable and look at the attacks counts. In contrast, kernel-density (KD) plot (e.g. Figure 5) sort hosts by how *often they were attacked*: (1) hosts without attacks, (2) hosts with a few (1 or 2) attacks, and (3) hosts with a higher (≥ 3) number of attacks. For each group, we plot the density function of their score of some independent variable.

Statistical tests. We want to test whether the differences the upper and lower 50%, reported in the median tables, are statistically significant or whether they are due to pure chance. To check whether the means for the two groups of hosts ($> median$ vs. $< median$) are statistically significant, we conduct the well-known Mann-Whitney U-test (using a Bonferroni correction to account for the number of significance tests conducted on the same data set).

In all the experiments reported in this section, the p-value is under* 0.001*, suggesting that our results are correct with probability 99.9% or more. This is true even when e.g. the values in the median tables are identical up to the first decimal place. This high statistical significance is not surprising given our massive dataset.

6.2. Number of Binaries and Risk

We compare the number of attacks against machines that have less than and more than the median number of binaries via Median Table III.

Table III. Average number of attacks for hosts with greater or less than the median of the number of binaries. (diff. significant at $p < 0.001$)

Binary count	Gamer	Pro	SW-dev	other	All
< Median	5.4	3.5	6.6	2.6	2.9
> Median	5.4	4.2	8.6	2.9	3.7

Our main finding is listed below.

Finding 1.

(1) *In the case of* `SW-dev` *hosts, there is strong evidence to support the hypothesis that the number of binaries on a host is linked to the number of attacks: developers with above-median binary counts receive 8.6 attacks on average, a 30% increase over the value for developers with fewer than the median number of binaries on their hosts.*

(2) *For the other categories of users, the link between the number of binaries and the risk of cyber attacks is weaker, while still statistically significant.*

The Decile Plot in Figure 4 also shows that `SW-dev` exhibit a clear upward trend in the number of infections as the number of binaries goes up, suggesting that a high number of binaries is associated with more attacks against hosts used for software development purposes. Moreover, we see that `SW-dev` machines are more heavily attacked irrespective of the decile considered, followed by `gaming` and `pro` categories.

Additionally, the slopes of the corresponding curves in Figure 4 are closer to the horizontal. This suggests that there is something uniquely risky about the way software developers acquire binaries. Moreover, the more binaries they acquire, the more likely it is that they will be attacked, which suggests that the total number of binaries present on a host is an important risk factor for software developers.

Figure 5 shows kernel-density (KD) plots of the number of binaries per host. We see the same trend. The higher the attack count, the more the density curve shifts right, i.e. has more binaries. However, the shift is minor and there is a large overlap in the density curves. This means, that the number of binaries explains only some of the difference between the low, medium, and highly attacked machines.

6.3. Percentage of Low/High Prevalence and Unique Binaries and Risk

We checked if there was a difference in risk associated with low prevalence binaries (present on under 1K hosts), medium prevalence (present on 1K to 1M hosts) binaries, and high prevalence binaries (present on over 1M hosts). We also consider unique binaries, i.e. binaries we found on only one host. Median Table IV shows how the number of attacks changed when we considered Low/High prevalence and unique files which were below/above the median number of files in each category.

Table IV. Average number of attacks for hosts with greater or less than the median of the fraction of low/high/unique binaries. (diff. significant at $p < 0.001$)

Prevalence	Gamer	Pro	SW-dev	other	All
< Median Low	2.9	2.5	4.0	2.0	2.2
> Median Low	7.5	4.9	8.4	3.6	4.4
< Median High	6.1	4.6	9.0	3.5	4.2
> Median High	3.7	2.8	3.9	2.2	2.4
< Median Unique	4.5	3.5	7.2	2.4	2.9
> Median Unique	6.8	4.5	8.6	3.1	3.7

Our main finding is described below.

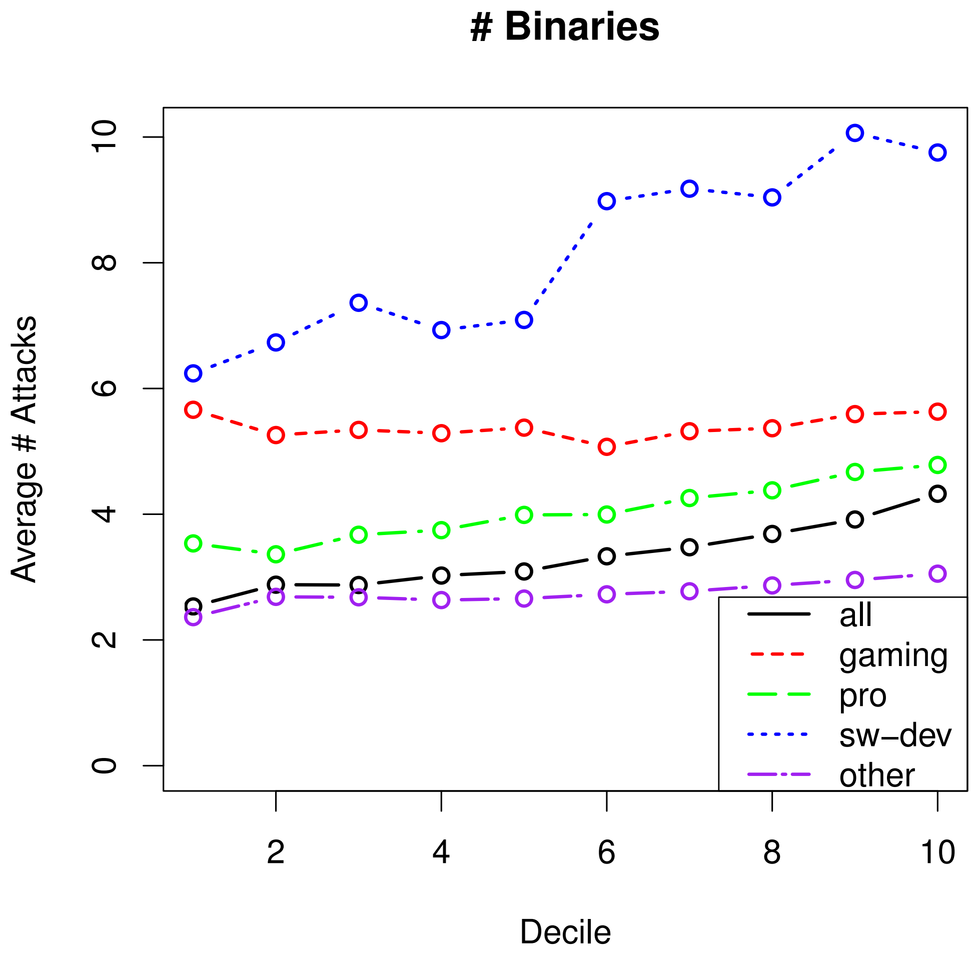


Fig. 4. Graph showing average number of attacks against machines in the i'th decile by number of binaries for $i = 1, \dots, 10$

FINDING 2. *Table IV shows a clear trend – `SW-dev` hosts are most at risk, followed by `gamer` followed by `pro`, and this risk goes up as the percentage of low prevalence binaries increases, irrespective of which category is considered. When we look at the high prevalence binaries, the trend is reversed as we would expect. The Decile Plot of Figure 6 shows the number of attacks for machines in each category by decile.*

Software developers create unique executable binaries by writing and compiling programs. These executables are included in our low-prevalence binary counts but they do not pose a threat to the developers' hosts. Our analysis of the impact of unique files highlight this trend: software development hosts with above-median unique files receive only 20% more attacks, in contrast to the >100% increase when considering low-prevalence files, and the corresponding line in the decile plot is not monotonically increasing. We observe a similar trend for professional users.

The huge increase in attack numbers for `gamer` hosts with a high ratio of low-prevalence binaries is remarkable. From the 8th to the 10th decile, the average attack count more than

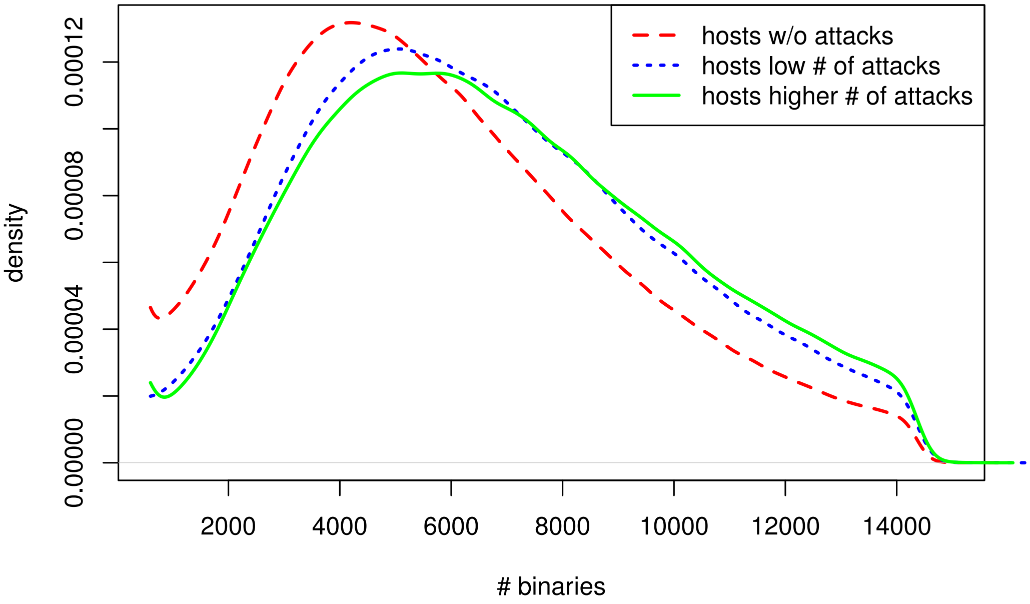


Fig. 5. Kernel-density plots of number of binaries

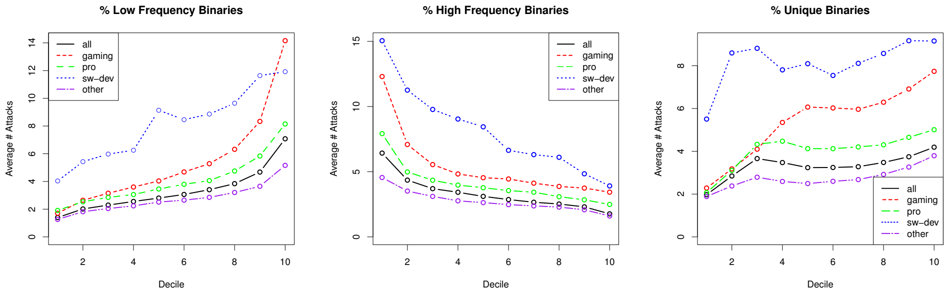


Fig. 6. Plots showing number of low/high density files by decile and average number of attacks by decile for deciles $1, \ldots, 10$

doubles. The results suggest that there is a subpopulation of gamers with especially risky behavioral patterns.

These plots provide further evidence supporting the hypothesis that an increased fraction of low prevalence files on a host increases the risk of malware attack—and, as expected, as the fraction of high-prevalence files increases, the risk level decreases.

Figure 7 shows separate KD-plots for binaries with low, medium and high prevalence. We observe the same pattern. In the first graph, the KD-plot for low-prevalence binaries shows that every type of host (with a low, medium or high number of malware attacks) has larger numbers of attacks than in other cases.

Software Developers are at Higher Risk. Figure 2 shows that software developers have a higher rate of low-prevalence binaries and unique binaries, possibly because they compile binaries. They also have much higher attack rates than other groups. We wondered if the high attack rate is the result of misclassification of custom-build binaries as malware. Symantec anti-virus product detects malware using both static signatures (e.g. checking the

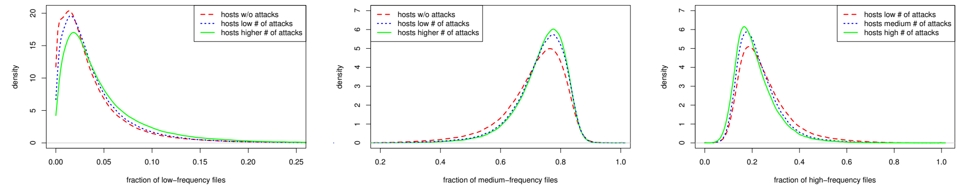


Fig. 7. Kernel-density plots of fraction of files with low/medium/high overall prevalence

file hash against a blacklist, scanning for strings or regular expressions in the binary content) and the behavior of a binary, observed at runtime (e.g. downloading files, modifying Windows registry entries). Perhaps newly compiled binaries, which are unique files and look naturally suspicious to an anti-virus, further trigger behavior-detection heuristics that cause them to be reported as malicious. We therefore break down the attack numbers separately for signature-based and behavioral virus detection. `sw-dev` hosts have on average 7.2 pieces of malware detected based on signatures and 0.9 pieces of malware detected by behavior. For non-`sw-dev` these numbers are 3.0 and 0.2, respectively. While behavioral detections do result in 4.5 more attack reports for software developers—as opposed to 2.4 more in the case of signature-based detections—their contribution to the overall attack averages is small. This suggests that our results are not distorted by misclassification of benign binaries as malware.

6.4. Percentage of Unsigned Binaries & Risk

Median Table V shows the link between the number of unsigned binaries on a host machine and the total risk.

Table V. Average number of attacks for hosts with greater or less than the median of the fraction of unsigned/unsigned&unique binaries. (diff. significant at $p < 0.001$)

Binaries	`Gamer`	`Pro`	`SW-dev`	`other`	`All`
$<$ Median Unsigned	3.9	2.9	5.3	2.4	2.6
$>$ Median Unsigned	6.3	4.6	8.9	3.2	4.0
$<$ Median Unsigned & Unique	3.5	3.5	7.0	2.4	2.9
$>$ Median Unsigned & Unique	6.7	4.5	8.7	3.1	3.7

Our main finding is listed below.

FINDING 3.

(*1*) *Table V shows that hosts with a larger than median percentage of unsigned binaries are more at risk than those with a less than median percentage of unsigned binaries. It also suggests that `gamers` are more vulnerable than `pros` who in turn are less vulnerable than `SW-dev`.*

(*2*) *However, the higher risk associated with this feature seems comparable across the three user categories: we observe 59%–68% more attacks against hosts with above-median numbers of unsigned binaries.*

The Decile Plot in Figure 8 further substantiates this observation.

We also checked for binaries that are both unsigned and unique. As mentioned before, binaries might be unique because they are infected by a polymorphic virus that changes its own code to avoid signature-based detection. Additionally, legitimate binaries might be unique because they result from just-in-time (JIT) compilation during installation or on the first launch or because they include customer-specific licensing information; however, such legitimate binaries are typically signed. We therefore checked whether unsigned and unique binaries are better indicators for malware attacks. For developers and professional users, this does not seem to be the case (the unique binaries that result from program compilation

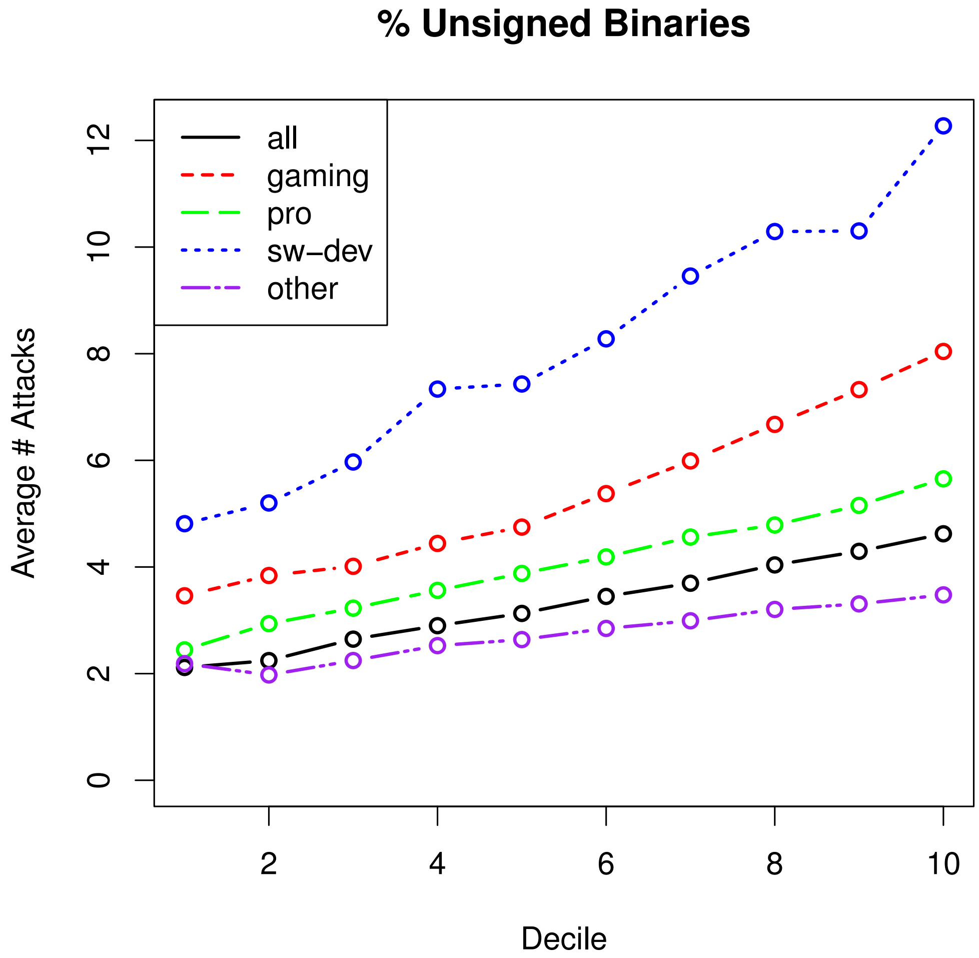


Fig. 8. Plot showing number of attacks for host machines in the i'th decile by number of unsigned binaries for $i = 1, \ldots, 10$

are not typically signed). However, for gamers the addition of the uniqueness criterion leads to a higher risk profile: 91% more attacks for hosts with above-median numbers of unique unsigned files, as opposed to 62% more attacks when considering all unsigned files.

Finally, Figure 9 shows via a kernel density plot showing that hosts with a higher fraction of unsigned files experience a higher occurrence of malware.

6.5. Percentage of Downloaded Binaries & Risk

Median Table VI compares the average attack numbers of hosts on which we found more or less than the median fraction of downloaded binaries.

Table VI. Average number of attacks for hosts with greater or less than the median of the fraction of downloaded binaries. (diff. significant at $p < 0.001$)

Downloaded binaries	Gamer	Pro	SW-dev	other	All
< Median	3.7	2.8	5.5	2.2	2.4
> Median	6.4	5.0	9.8	3.4	4.2

We are able to statistically validate the following clear result.

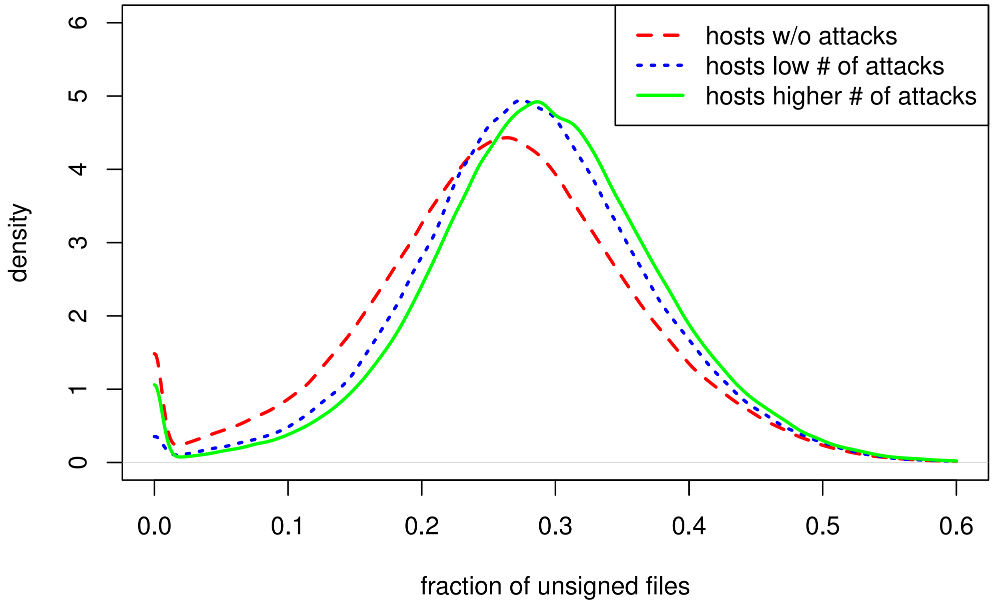


Fig. 9. Kernel-density plots of fraction of unsigned binaries

FINDING 4. *For all categories of users, Table VI clearly shows that users whose fraction of downloaded binaries is over the median value experience far more attacks than users whose fraction of downloaded binaries is below the median.*

The by-decile data of Figure 10 shows that an especially sharp increase in attack numbers for the 9th and 10th decile.

The KD plot in Figure 11 shows a much better separation of the density curves for the three classes of hosts than the plots for most other IVs, suggesting that the hosts that are not attacked exhibit a different download behavior than the other hosts. In particular, for hosts with no attacks, the most frequent value in the download count distribution is 0: refraining from downloading any executable files from the Internet makes users safer from cyber attacks.

6.6. User Travel History and Risk

We studied the hypothesis that increased travel by a host machine increases the risk and number of attacks of that machine. Measuring travel is hard — we used the number of Internet Service Providers (ISPs) that a host has connected through as a proxy for the amount of travel by the user of that machine.

Table VII. Average number of attacks for hosts with greater or less than the median of the number of ISPs. (diff. significant at $p < 0.001$)

ISPs	Gamer	Pro	SW-dev	other	All
< Median	4.9	3.7	7.5	2.6	3.0
> Median	5.9	4.6	8.7	3.1	3.8

Median Table VII shows the number of attacks for machines above/below the median number of ISPs that a machine connected to.

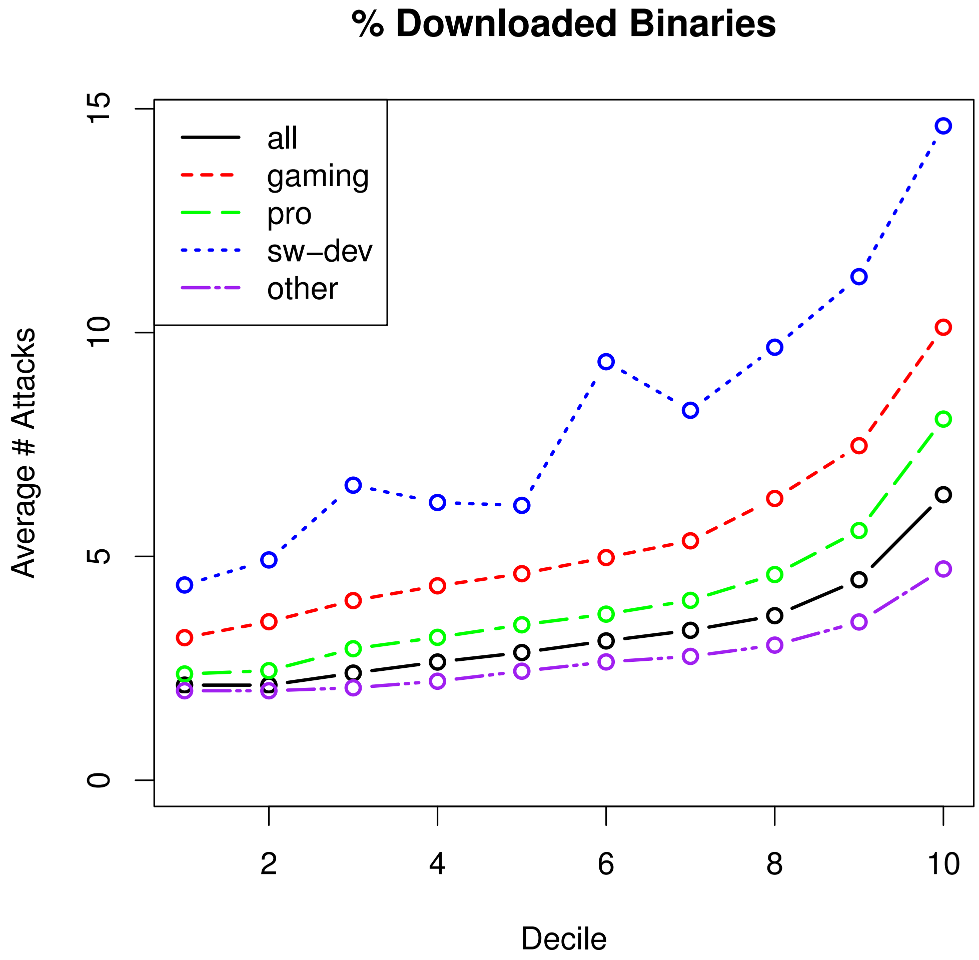


Fig. 10. Plot showing number of attacks for host machines in the i'th decile by fraction of downloaded binaries for $i = 1, \ldots, 10$

FINDING 5. *We see* `SW-dev` *hosts receive more attacks than* `gamer` *hosts, which receive more attacks than* `pro` *hosts. Moreover, there is a clear increase in the number of attacks on hosts that are above the median in terms of the number of ISPs that they connected from.*

As in the previous sections, we investigated this further by looking at the number of attacks on a decile by decile basis w.r.t. the number of ISPs that hosts connect to — the result in shown in Figure 12 below.

The trend here is not as clear as in previous cases, and the risk does not increase as much for hosts with above-median travel histories (up to 24% more attacks, for professional users). Nonetheless, there is a discernible upward trend (especially when we get to the higher deciles). In particular, for all categories of hosts, we see that when we get to the eighth decile, there is a marked increase in the number of attacks. This is further confirmed by the kernel density plot in Figure 13—but the connection between number of ISPs and a higher risk of attack is weaker than in preceding sections.

As the influence of the ISP count on the number of attacks is rather low and may also reflect the usage intensity of a host, we do not claim a strong connection between travel frequency and attacks.

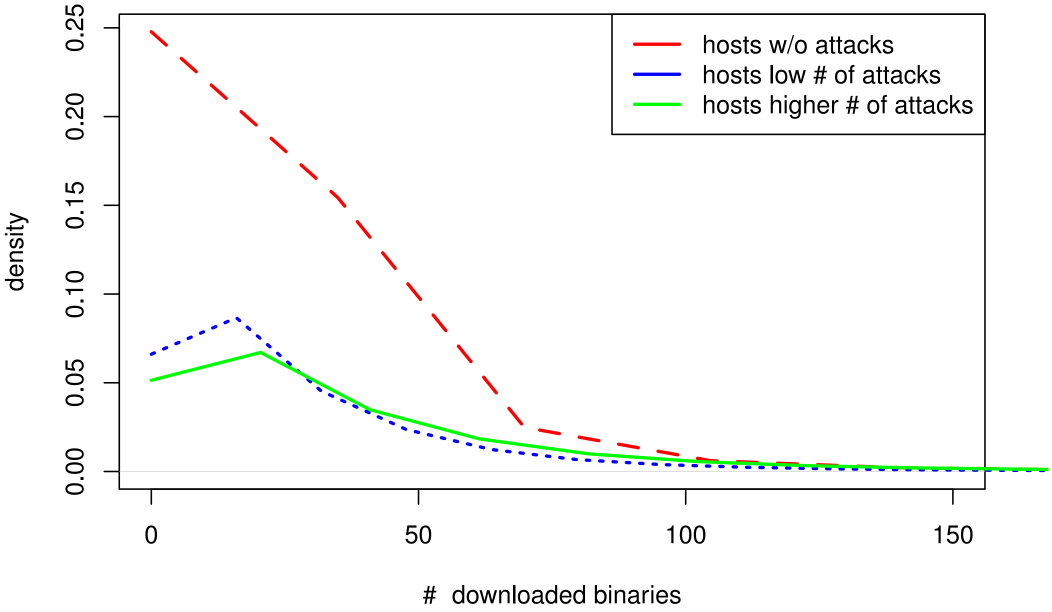


Fig. 11. Number of downloaded files density

7. DISCUSSION AND CONCLUSION

Though humans are believed to be one of the weaker links in cyber-security [Schneier], hardly any work to date has focused on the relationship between the behavior of human users of machines and malware attacks on those machines. In this paper, we report the results of an extensive analysis of Symantec's WINE dataset in which we studied 1.6M machines host over an 8-month period. To our knowledge, this is the most extensive, data-centered study of this important topic. We identified a set of machine features (number of binaries, fraction of unsigned, downloaded, low prevalence, and unique binaries and number of ISPs the user connected from) related to behaviors of the user. For instance, these features are proxies for the tendency of users to download lots of binaries, to travel a lot, and to download rare pieces of code. We grouped all users into 5 categories: `gamers`,`pros`,`SW-dev`, `other` and all.

Our results show that all of these variables are related to the number of pieces of malware found on their host machines at a statistically significant level ($p < 0.001$). Of these statistically significant results, the ones that we deem the most solid are ones showing that the number of malware infections on a machine are related to the number of downloaded, unsigned, and low prevalence binaries for all categories of users. Moreover, the number of binaries on hosts are linked to number of malware infections on hosts in the `SW-developer` category also at the statistically significant $p < 0.001$ level even when we account for the fact that software developers may generate binaries by compiling code they are developing.

In addition, we identified 5 groups of users (gamers, software developers, professionals, other not in any of the preceding categories, and all users) and saw that software developers appear to be the most prone to malware attacks. An interesting possibility for future work would be to perform a clustering of users who are prone to malware attacks and see what kinds of properties are common to users within a cluster.

A major next step is to see if we can predict which machines will be infected by a given piece of malware and/or how many hosts in a given population of hosts will be infected. This

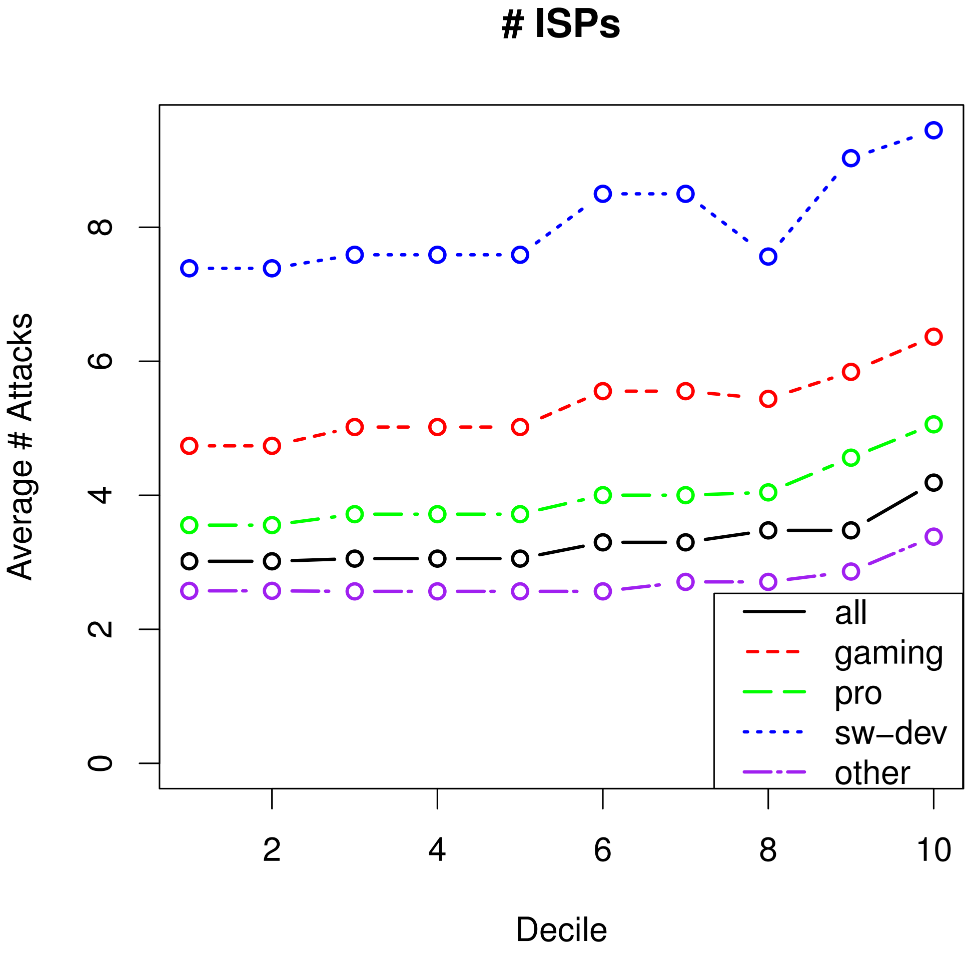


Fig. 12. Plot showing number of attacks for host machines in the i'th decile by number of ISPs that the machine connected to for $i = 1, \ldots, 10$

is a challenging problem which is significantly different from that studied in the paper — though many of the findings in this paper could feed into just such a predictive model. We have developed an ensemble based predictive model [Kang et al. 2016] with high correlations between predicted number of infections and true number of infections, but much further work remains.

Acknowledgments

This work may have been partly supported by ARO under Grants W911NF11103, W911NF09102, W911NF1410358, W911NF1110344, ONR grant N00014-15-1-2007 and by the Maryland Procurement Office under Contract No. H98230-14-C-0137.

We thank Symantec for providing access to the WINE platform. Other researchers may reproduce and verify our results by analyzing the reference data set we recorded in WINE (WINE-2013-001) after signing a research agreement with Symantec.

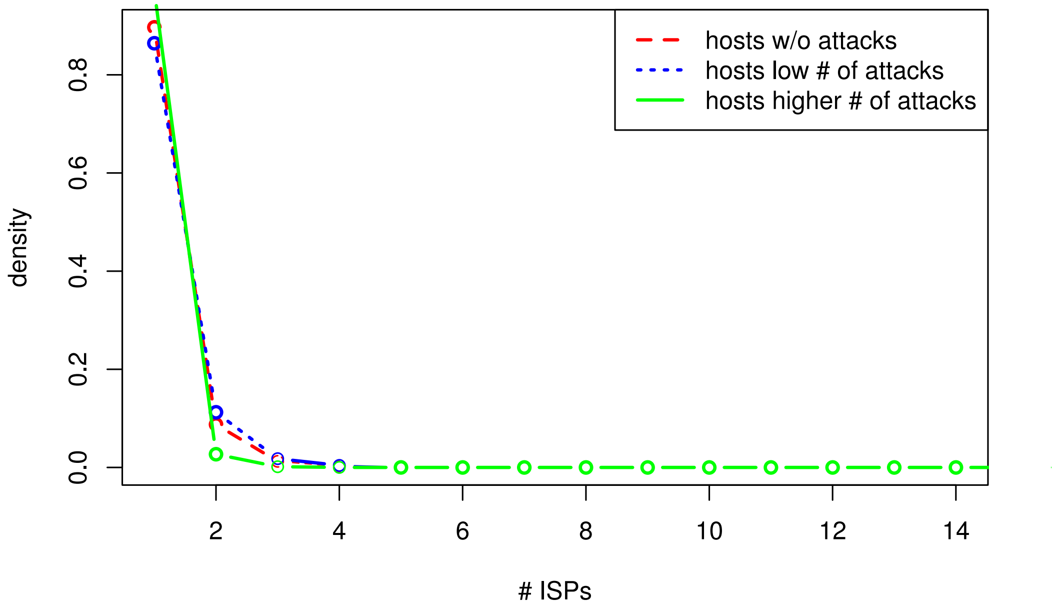


Fig. 13. Number of ISP density

REFERENCES

Sherly Abraham and InduShobha Chengalur-Smith. 2010. An overview of social engineering malware: Trends, tactics, and implications. *Technology in Society* 32, 3 (2010), 183 – 196. DOI:http://dx.doi.org/10.1016/j.techsoc.2010.07.001

Mustaque Ahamad, Dave Amster, Michael Barrett, Tom Cross, George Heron, Don Jackson, Jeff King, Wenke Lee, Ryan Naraine, Gunter Ollmann, and others. 2008. Emerging cyber threats report for 2009. (2008).

Ross J. Anderson. 1993. Why Cryptosystems Fail. In *ACM Conference on Computer and Communications Security*, Dorothy E. Denning, Raymond Pyle, Ravi Ganesan, Ravi S. Sandhu, and Victoria Ashby (Eds.). ACM, 215–227.

Stephen Bono, Dan Caselden, Gabriel Landau, and Charlie Miller. 2009. Reducing the attack surface in massively multiplayer online role-playing games. *Security & Privacy, IEEE* 7, 3 (2009), 13–19.

L. Carlinet, L. Me, H. Debar, and Y. Gourhant. 2008. Analysis of Computer Infection Risk Factors Based on Customer Network Usage. In *Emerging Security Information, Systems and Technologies, 2008. SECURWARE '08. Second International Conference on.* 317–325. DOI:http://dx.doi.org/10.1109/SECURWARE.2008.30

Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, and Christos Faloutsos. 2010. Polonium: Tera-scale graph mining for malware detection. In *Proceedings of the second workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2010), Washington, DC*, Vol. 25.

Sandy Clark, Travis Goodspeed, Perry Metzger, Zachary Wasserman, Kevin Xu, and Matt Blaze. 2011. Why (special agent) Johnny (still) can't encrypt: a security analysis of the APCO project 25 two-way radio system. In *Proceedings of the 20th USENIX conference on Security*. USENIX Association, 4–4.

Crispin Cowan. 2013. Windows 8 Security: Supporting User Confidence. USENIX Security Symposium, invited talk. (August 2013).

Mache Creeger, Charles Reis, Adam Barth, Carlos Pizano, Niels Provos, Moheeb Abu Rajab, Panayiotis Mavrommatis, Thomas Wadlow, and Vlad Gorelik. 2010. CTO Roundtable: Malware Defense Overview. *Queue* 8, 2 (2010), 50.

Robert E Crossler, Allen C Johnston, Paul Benjamin Lowry, Qing Hu, Merrill Warkentin, and Richard Baskerville. 2013. Future directions for behavioral information security research. *computers & security* 32 (2013), 90–101.

Tudor Dumitraş and Darren Shou. 2011. Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Environment (WINE). In *EuroSys BADGERS Workshop*. Salzburg, Austria.

Chris Grier, Lucas Ballard, Juan Caballero, Neha Chachra, Christian J. Dietrich, Kirill Levchenko, Panayiotis Mavrommatis, Damon McCoy, Antonio Nappa, Andreas Pitsillidis, Niels Provos, M. Zubair Rafique, Moheeb Abu Rajab, Christian Rossow, Kurt Thomas, Vern Paxson, Stefan Savage, and Geoffrey M. Voelker. 2012. Manufacturing compromise: the emergence of exploit-as-a-service. In *ACM Conference on Computer and Communications Security*, Ting Yu, George Danezis, and Virgil D. Gligor (Eds.). ACM, 821–832.

AVG Viruslab Research Group. 2013. AVG Insight: 90% of game hacks infected with malware. `http://blogs.avg.com/news-threats/avg-insight-90-game-hacks-infected-malware/`. (Apr 2013).

Paul S Henry and Hui Luo. 2002. WiFi: what's next? *Communications Magazine, IEEE* 40, 12 (2002), 66–72.

Hao Hu, Steven Myers, Vittoria Colizza, and Alessandro Vespignani. 2009. WiFi networks and malware epidemiology. *Proceedings of the National Academy of Sciences* 106, 5 (2009), 1318–1323.

Princely Ifinedo. 2012. Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory. *Computers & Security* 31, 1 (2012), 83–95.

Allen C Johnston and Merrill Warkentin. 2010. Fear appeals and information security behaviors: an empirical study. *MIS quarterly* 34, 3 (2010), 549–566.

Chanhyun Kang, Noseong Park, B. Aditya Prakash, Edoardo Serra, and V.S. Subrahmanian. 2016. Ensemble Models for Data-Driven Prediction of Malware Infections. In *accepted for publication in: Proc. 2016 ACM International Conference on Web Search and Data Mining*. ACM.

Fanny Lalonde Lévesque, Jude Nsiempba, José M. Fernandez, Sonia Chiasson, and Anil Somayaji. 2013. A Clinical Study of Risk Factors Related to Malware Infections. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &#38; Communications Security (CCS '13)*. ACM, New York, NY, USA, 97–108. DOI:`http://dx.doi.org/10.1145/2508859.2516747`

John Leach. 2003. Improving user security behaviour. *Computers & Security* 22, 8 (2003), 685 – 692. DOI:`http://dx.doi.org/10.1016/S0167-4048(03)00007-5`

Pratyusa K. Manadhata and Jeannette M. Wing. 2011. An Attack Surface Metric. *IEEE Trans. Software Eng.* 37, 3 (2011), 371–386.

Mandiant. 2013. APT1: Exposing One of China's Cyber Espionage Units. Mandiant Whitepaper. (Feb 2013).

L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath. 2011. Malware images: visualization and automatic classification. In *Proceedings of the 8th International Symposium on Visualization for Cyber Security (VizSec '11)*. ACM, Article 4, 7 pages. DOI:`http://dx.doi.org/10.1145/2016904.2016908`

Jarno Niemelä. 2010. It's Signed, therefore it's Clean, right? *CARO 2010* (2010).

Aikaterinaki Niki. 2009. Drive-by Download Attacks: Effects and Detection Measures. In *IT Security Conference for the Next Generation.*

Gavin O'Gorman and Geoff McDonald. 2012. The Elderwood Project. Symantec Whitepaper. (Oct 2012).

Evangelos E. Papalexakis, Tudor Dumitras, Duen Horng Chau, B. Aditya Prakash, and Christos Faloutsos. 2013. Spatio-temporal Mining of Software Adoption & Penetration. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.*

Niels Provos, Dean McNamee, Panayiotis Mavrommatis, Ke Wang, Nagendra Modadugu, and others. 2007. The ghost in the browser analysis of web-based malware. In *Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets*, Vol. 10. 4–4.

Moheeb Abu Rajab, Lucas Ballard, Noé Lutz, Panayiotis Mavrommatis, and Niels Provos. 2013. CAMP: Content-Agnostic Malware Protection. In *Network and Distributed System Security (NDSS) Symposium.* San Diego, CA.

Bruce Schneier. Semantic Attacks: The Third Wave of Network Attacks. `https://www.schneier.com/crypto-gram-0010.html#1`. (????).

Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 373–382.

Frank Stajano and Paul Wilson. 2011. Understanding scam victims: seven principles for systems security. *Commun. ACM* 54, 3 (2011), 70–75.

Stuart Staniford, David Moore, Vern Paxson, and Nicholas Weaver. 2004. The top speed of flash worms. In *Proceedings of the 2004 ACM workshop on Rapid malcode (WORM '04)*. ACM, New York, NY, USA, 33–42. DOI:http://dx.doi.org/10.1145/1029618.1029624

Stuart Staniford, Vern Paxson, and Nicholas Weaver. 2002. How to Own the Internet in Your Spare Time. In *Proceedings of the 11th USENIX Security Symposium*. USENIX Association, Berkeley, CA, USA, 149–167. http://dl.acm.org/citation.cfm?id=647253.720288

Symantec Corporation. 2012. Symantec Internet Security Threat Report, Volume 17. http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_2011_21239364.en-us.pdf. (April 2012).

Alma Whitten and J Doug Tygar. 1999. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Proceedings of the 8th USENIX Security Symposium*, Vol. 99. McGraw-Hill.