An Epidemiological Study of Malware Encounters in a Large Enterprise

Ting-Fang Yen
E8 Security
tyen@e8security.com

Victor Heorhiadi
University of North Carolina at Chapel Hill
victor@cs.unc.edu

Alina Oprea
RSA Laboratories
alina.oprea@rsa.com

Michael K. Reiter
University of North Carolina at Chapel Hill
reiter@cs.unc.edu

Ari Juels
Cornell Tech
ajuels@gmail.com

ABSTRACT

We present an epidemiological study of malware encounters in a large, multi-national enterprise. Our data sets allow us to observe or infer not only malware presence on enterprise computers, but also malware entry points, network locations of the computers (i.e., inside the enterprise network or outside) when the malware were encountered, and for some web-based malware encounters, web activities that gave rise to them. By coupling this data with demographic information for each host's primary user, such as his or her job title and level in the management hierarchy, we are able to paint a reasonably comprehensive picture of malware encounters for this enterprise. We use this analysis to build a logistic regression model for inferring the risk of hosts encountering malware; those ranked highly by our model have a $> 3\times$ higher rate of encountering malware than the base rate. We also discuss where our study confirms or refutes other studies and guidance that our results suggest.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection—*Invasive Software*; C.2.0 [**Computer-Communication Networks**]: General—*Security and Protection*

Keywords

Malware encounters, enterprise security, measurement, logistic regression

1. INTRODUCTION

In this paper, we present the first epidemiological study of malware encounters within a large enterprise. Formally, epidemiology "deals with the incidence, distribution, and control of disease in a population" [1] — the disease, in our setting, being malware. We explore in this sense the patterns and causes of malware encounters within a population, namely the hosts and employees of the enterprise under study. Our work benefits from privileged access to security logs generated within this enterprise, as well as databases containing information about the employees using the hosts on the enterprise network. By agreement with the security administrators of the enterprise, we omit its name from this paper.

Malware spread is a well-studied problem in consumer environments. It is far less well studied, however, in enterprise settings, which may differ in important ways from consumer contexts. Corporate computing resources are subject to tighter security policies and benefit from more expert security administration than most consumer devices. Corporations possess digital assets, however, including financial data and intellectual property, far more valuable than the data on most consumer devices. Enterprise data might thus be subject to highly sophisticated targeted attacks in which malware often plays a pivotal role.

Enterprise host security also merits study in its own right. Enterprises are interesting microcosms in which users (employees) are assigned highly specific demographic classifications in the form of job titles, business-unit placement, and level in the management hierarchy. Employees are also subject to more or less uniform security policies across an enterprise and use similar (if not identical) O/S and software versions, creating a controlled environment amenable to scientific study. As employee behavior on an enterprise network is subject to monitoring by the enterprise, and many hosts are instrumented with software (e.g., anti-virus) that generates internal reports, enterprises additionally have data about the behavior of their users unavailable in many consumer settings or whose use would infringe upon users' legal rights.

Our study and results: We rely on reporting by anti-virus software (McAfee), with which 85,000+ hosts owned and managed by the enterprise are instrumented. As these are cases where malware was detected and presumably prevented from executing, we borrow from existing terminology [12] and refer to them as malware *encounters* rather than infections.

We study these encounters from several distinct perspectives. First, by examining the file system locations of detected malware instances, we characterize the vectors by which malware gains access to hosts (e.g., external drive, web). Second, we quantify the frequency of "inside" encounters—those occurring while hosts are connected to the enterprise network, either over VPN or on the LAN—and "outside" encounters (when employees take hosts home or to customer sites). Third, we correlate malware reports with demographic information about hosts' users, including their job roles, levels in the management hierarchy of the company, and geographic locations. Finally, we examine encounters resulting from host visits to malicious web sites, drawing on enterprise web proxy

CCS'14, November 3–7, 2014, Scottsdale, Arizona, USA.
ACM 978-1-4503-2957-6/14/11.
http://dx.doi.org/10.1145/2660267.2660330 .

logs to ascertain the browsing behavior underlying these encounters.

Many of our data sources do not provide direct indications of the host and user behaviors of interest, and inferring these behaviors in some cases presents interesting technical challenges. For example, McAfee anti-virus software reports do not indicate whether hosts are inside the enterprise network ("on-network") or outside ("off-network") at the time of a malware encounter. We thus classify encounters based on reporting delay times.

We present a number of interesting findings, including several that, to the best of our knowledge, reflect previously unstudied phenomena. Some key findings in the enterprise under study are:

- Off-network encounters are roughly three times more common than on-network encounters.
- Encounter rates are lowest at the upper levels of the corporate management hierarchy.
- External drives are the most common vector of malware encounters, and especially prevalent in low per-capita GDP countries.
- Roughly 31% of the web-based malware encounters (around 554 over four months, as extrapolated from encounters traceable to proxy logs) originate from websites classified by the enterprise web proxy into the "business" category, and 15% (around 266, similarly extrapolated over four months) originate from the "travel" category.

We present these and other results and offer some conjectural explanations below.

Finally, drawing together the data sources examined in our study, we identify features (demographic and behavioral) that correlate significantly with a host encountering malware. We use these features to construct a logistic regression model that estimates the encounter risk associated with individual hosts. This model successfully identifies high-risk hosts: the encounter rate among the top 1,000 identified hosts is 51%, significantly more than three times the base rate for the enterprise as a whole.

Our contributions: Our main contributions are:

- We present the first large-scale epidemiological study of enterprise malware encounters, based on analysis of a repository of sensitive enterprise security data.
- We correlate malware encounters on hosts with a variety of demographic and behavioral features for users (job roles, web browsing behavior, etc.), many of them previously unstudied. Our study yields a number of significant findings.
- We build a logistic regression model across a subset of these features that successfully identifies hosts with a significantly elevated risk of encountering malware.

While our findings are illuminating (and sometimes counterintuitive) in their own right, we believe they can also help shape enterprise security policy, select effective security tools, and create targeted security education programs for employees. We emphasize that our study treats a single (albeit, large) enterprise; its wider applicability, of course, requires further investigation. We believe that case studies of this kind are individually informative and also important steps toward the creation of a portfolio of real-world research results that can illuminate broad practices and trends.

2. DATA

We leverage multiple data sources to study malware encounters in the enterprise under study. We give details in this section on the five major data sources in our study (Section 2.1-Section 2.5) and review the ethical and privacy considerations affecting our work (Section 2.6).

2.1 Anti-Virus Reports

We examine reports collected from McAfee anti-virus agents deployed on 85,000+ hosts owned and managed by the enterprise. In addition to matching known virus signatures, the McAfee agent also detects suspicious files using a cloud-based reputation service (McAfee Global Threat Intelligence).[1] Detection occurs in on-access mode, where a file object is scanned as it is read into memory, after which the suspicious file is deleted or quarantined.

Reports generated by the McAfee agent are sent from the end hosts to a centralized data collector within the enterprise network immediately upon generation. If the host is outside the corporate network or the collection server is otherwise inaccessible, the agent will buffer reports on the host and attempt to re-send every five minutes. Due to storage and bandwidth constraints, only reports for detected malware are collected, which excludes data about, e.g., results of virus scan on "clean" end hosts or time of the last signature update.

Table 1 shows the fields available in each McAfee report. In addition to the host name, virus name, and file path, each report also includes two timestamps: One indicating the time when the malicious file was detected on the end host, and another the time when this report was received at the data collection server. We note, with confirmation from the enterprise's IT team, that no buffering is performed during the data collection process that would affect the timestamps we use.

Field	Description
Host name	A fully-qualified domain name that serves as an unique identifier for the end host on the enterprise network.
Virus name	The name of the identified threat (according to McAfee).
File path	The full path of the malicious file.
Detection time	Time of detection on the end host.
Reporting time	Time of collection at the enterprise data collection server.

Table 1: Fields in McAfee reports.

Over a four-month period, from July 10 to November 10, 2013, the centralized data collector received a total of 569,967 reports. However, many of those reports appear to be redundant, with the same {host name, virus name, file path} tuple repeated within an interval of seconds. 87.76% of them appear less than one minute apart from a previous, identical report. Most of these redundant reports are due to the McAfee agent attempting to delete or quarantine read-only files. For reports with the same {host name, virus name, file path} tuple, we retain the report with the earliest detection timestamp and exclude the remainder from consideration.

We also discovered an outlier host that generated over 10,000 reports in four hours. The McAfee agent detected and deleted files repeatedly created by malware on the host (but not the actual malware binary). There are also a small number of "false-positive" reports (2,132 reports) whose detected malicious files were sample malware from security certification courses or research activities; e.g., the file paths included directories names like "PenTesting" or

[1] `http://www.mcafee.com/us/threat-center/technology/global-threat-intelligence-technology.aspx`

"CEHv8 Module". We also filter these reports out. Our filtered dataset includes 120,161 reports from 10,941 distinct hosts.

2.2 Employee Database

The enterprise also stores information about each employee that includes the employee's name, the employee ID number, office location, business unit, job title, and manager ID number. We infer additional information about each employee based on this dataset. From the job title, we categorize a user's job type as the last word in the job title after stripping away level indicators (e.g., "engineer I" and "engineer II" are both considered "engineers"). Given each employee's manager ID, we build the organization tree with the company CEO as root. This allows us to assign a "level" to each employee based on the number of steps down from the tree root.

2.3 Windows Authentication Logs

While the McAfee reports and employee data each contain much useful information, correlating the two is a non-trivial task — McAfee reports are associated with *hosts*, while employee data is about *users*. Lacking documentation about machine assignments in the enterprise, we draw on a third data source to bridge this gap: Windows authentication logs from domain controllers (DC).

The DCs are responsible for validating authentication requests to access resources on a Windows domain. For example, when a user logs on to her corporate machine, the request is sent to the DC, where her credentials are verified. Each authentication log includes the user name, the host where access was requested, the timestamp, and other fields indicating the type of logon and whether the logon was successful.

To infer the primary user of a host, we examine Windows authentication logs over one month. For every host, a list is kept documenting the users that successfully authenticated to the host. After this month, the user responsible for a large majority (80%) of the logons on the host is assumed to be the primary user. If no such user exists for a host, it is assumed to be a multi-user server and removed from further consideration.[2] In this way, we determined the primary user for 62,884 enterprise-managed hosts that are instrumented with the McAfee client, of which 9,625 generated malware reports during our four-month observation period. In our study, we focus on the hosts for which a primary user can be identified.

2.4 Web Proxy Logs

In addition to anti-virus software, the enterprise network deploys a variety of security tools to prevent unwanted software and intrusions. One is a Cisco IronPort web proxy that filters HTTP and HTTPS requests. The proxy vendor provides reputation scores and category information (e.g., business, news, sports) for known sites, and the filtering policy blocks connections to websites with low reputation or in non-business-related categories.

In cases where a web request is made to a previously unknown website lacking reputation and category, the proxy instead displays a warning page to the user, stating that the site is considered higher risk. The user is asked to acknowledge that access to the site adheres to the company's security policies before being allowed to proceed. Once the user has acknowledged, her consent is valid for one hour. During this time, visits to other non-categorized websites are allowed without further prompting from the proxy.

Part of our study examines the effectiveness of the web proxy's filtering policy at preventing malware infections. For this, we make use of logs generated by the web proxy, which include the timestamp, the destination URL and domain, the source IP address, the web referer, user-agent string, the website reputation and category, and the filtering policy applied to that connection.

2.5 VPN Logs

One focus of our study is on *where* McAfee detection occurred, i.e., whether "inside" the corporate network (in which case the malware penetrated the network's security perimeter to arrive on the victim host) or "outside." As part of that investigation, we use Virtual Private Network (VPN) logs to examine employees' accesses to corporate resources while physically outside the company.

A VPN allows remote employees to establish a secure communication channel to the enterprise network. For each VPN session, the Cisco VPN server records the username that logged in, the fully-qualified domain name of the host used to log in, the time of login, the duration of the VPN session, the number of bytes sent and received during the session, and the external IP address from which the login was made. This gives us an approximation of how often a corporate laptop is brought outside of the enterprise network, and how it is used while outside.

2.6 Ethical and Privacy Considerations

As a matter of enterprise policy, employees are notified that they must consent to logging / monitoring of their activities by the enterprise IT security department in order to use enterprise networks and computers. The supervisors of the enterprise IT security department consented to the use of all of the datasets described above for the purposes of this study and also specifically to release of the summary data contained in this paper. Authors were granted access to the data only while on site at the enterprise, and not permitted to handle any data outside the enterprise network beyond the summary results presented here. Authors under the jurisdiction of Institutional Review Boards were not provided identifiable data and so did not require IRB approval.

3. MALWARE ENCOUNTER STATISTICS

Among enterprise-managed hosts whose primary user could be identified (see Section 2.3), 15.31% encountered malware over our four-month observation period. The hosts that generated McAfee reports, however, are not spread uniformly across the enterprise, as previous works have also observed in different datasets and types of entities (e.g., the geographic locations of spamming bots [16]).

Figure 1 shows the number (and fraction) of hosts in each country from which McAfee reports were received, where the country was determined by the corresponding employee's office location, as described in Section 2.2. Only the top 20 countries with the highest number of employees are plotted. The *encounter rate* (on the right Y-axis) is defined as the fraction of enterprise managed hosts in that country (for which the primary user could be identified, see Section 2.3) that generated McAfee reports. The encounter rate varies widely across geographic locations — the average encounter rate by country is 25.68%, with a standard deviation of 27.26%.

In this section, we attempt to understand better the threat landscape in enterprise environments and to expand upon differences between victims affected by malware. First, we examine the file system location of detected malicious files to identify potential methods by which malware arrived on victim hosts. Second, by observing the time difference between malware detection and reporting, we quantify the relative frequency of malware encounters taking place within and outside the enterprise network. Third, we study demographic characteristics of users whose hosts encountered malware. Finally, we correlate McAfee reports with web

[2] Not all of the enterprise servers are instrumented with McAfee anti-virus, unlike end hosts. Moreover, as they are multi-user, their behaviors could not be attributed to a primary user for inclusion in the demographics aspects of our study.

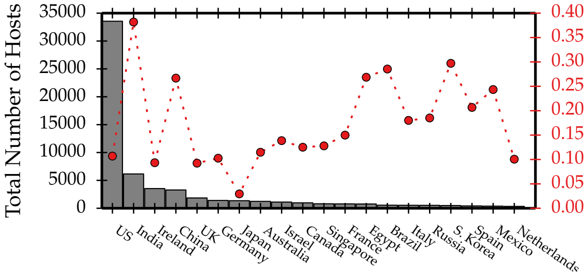

Figure 1: The number of hosts, and the malware encounter rate, in each country.

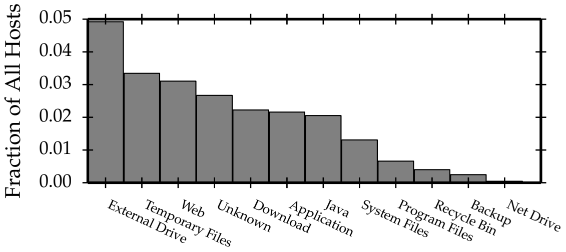

Figure 2: The file system location where malware was found.

proxy logs to understand the browsing behavior underlying web-based malware encounters. In Appendix A, we report the prevalence of various malware types identified by the McAfee agent.

Results in this section focus on the ten countries in our dataset with the most hosts that encountered malware: U.S., India, China, Ireland, Egypt, the U.K., Brazil, Israel, South Korea, and Germany.

3.1 Malware Location

In this section we investigate where the detected malicious file was found on the victim's file system. Lacking the ability to monitor host-level activities and to collect additional information at the time of detection, we instead leverage the directory structure of the Windows operating system to infer how (or why) a file ended up at that location. Appendix B describes the malware locations and how we categorize them.

Figure 2 shows the fraction of hosts reporting malware at a location indicated on the horizontal axis, out of all hosts instrumented with McAfee and whose primary user could be identified (see Section 2.3). External drives are by far the most prevalent location for malicious files, associated with 4.92% of all hosts (one-third of the hosts that encountered malware). Temporary folders and the browser cache follow, which likely correspond to secondary malware downloads by an initial exploit and web drive-by downloads.

This result is somewhat surprising. According to the Symantec Internet Security Threat Report from 2012 [19], web attacks are the favorite method for malware authors to gain access to victim hosts. The 2011 Microsoft Security Intelligence report [11] also found that most of the malware infections analyzed (45%) required user interactions to compromise the host, i.e., by visiting malicious webpages or installing malicious software, much higher than those accounted for by infected USB or removable drives (26%). By contrast, the detected malicious files are most commonly found on external drives in our dataset. This discrepancy can perhaps be partially attributed to the enforcement of web filtering policies on outbound requests (apparently motivated by the prevalence of web attacks). As described in Section 2.4, connections are blocked if they are destined for a blacklisted site, or if the remote site is non-categorized or has low reputation. As a result, known web attacks (such as those detected by anti-virus like McAfee) are prevented, and other infection vectors (e.g., external drives) dominate.

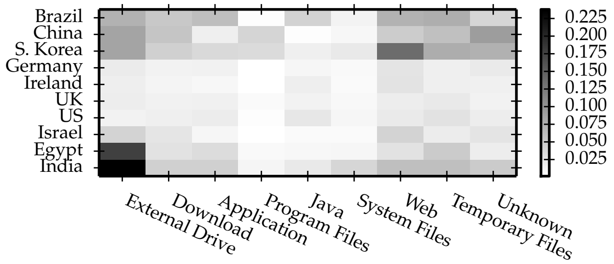

Figure 3: Distribution of malware locations by country. The shading of a box represents, out of the hosts in that country, the fraction where malware was found at that file system location.

Figure 3 shows, for each location category, the fraction of hosts in each country on which malware was found at that file system location. Rows (and columns) are sorted according to a hierarchical clustering algorithm to minimize the Euclidean distance between adjacent rows (and columns). It is clear that India and Egypt stand out as having a high fraction of hosts encountering malware on external drives (23.79% and 17.86%, respectively). Also obvious is the dominance of malware under the web browser cache for hosts in South Korea, perhaps due to the high number of phishing, malware hosting, and drive-by-download sites in the country (80, 172, and 54 times higher than that in the U.S.[3]).

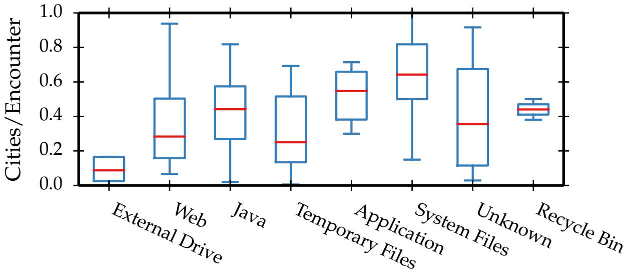

Figure 4: The rate at which malware of different file system location categories are encountered across cities.

We also examine the temporal and spatial locality of malware encounters in each file system location category. The intuition is that, even though file system location is only a rough indicator of how malware propagates, certain file system locations commonly associated with human-driven activities should show a "slower" malware spread across geographic regions than those associated with automated events.

To measure this behavior, we first identified each distinct malware (as identified by McAfee) that is primarily found in one file system location category (i.e., accounting for over 95% of its encounters), and that is encountered more than 10 times within our dataset and in multiple cities (inferred from the office location of the employee associated with the reporting host). For each such malware, we computed its rate of geographic spread or, more

[3] `http://blogs.technet.com/b/security/archive/2011/07/18/a-very-active-place-the-threat-landscape-in-the-republic-of-korea.aspx.`

specifically, the number of cities in which it is encountered, divided by the number of encounters for that malware. So, a smaller value for a particular malware indicates that its encounters are concentrated in fewer cities. We then grouped the malware by their file system location categories and summarized each group's rates of geographic spread using a boxplot, as shown in Figure 4. As expected, malware found on external drives spread the slowest, likely due to them involving human interactions to move across geographic locations. Malware under the browser cache (also likely associated with human activities) and temporary folders (potentially secondary downloads caused by web drive-bys) are slightly faster, while those under the system or application directories spread the fastest.

Findings: Malware on external drives are found on 4.92% of the hosts instrumented with the McAfee client, while malware under temporary folders are found on 3.34% of the hosts, followed by the web cache on 3.11% of the hosts. Compared to consumer contexts, hosts in the enterprise seem to be less exposed to web attacks (as seen by the relative dominance of external drives and temporary folders to the web cache). This is likely due to the enforcement of web filtering policies at the enterprise network border.

There is a geographic difference in the malware locations on the file system — hosts in India and Egypt are more likely to encounter malware on external drives, while a high fraction of hosts in South Korea encounter web malware. This suggests that targeted user education may be helpful in reducing specific risk factors in certain regions.

3.2 Inside vs. Outside Enterprise Network

As mentioned in Section 2.1, a McAfee agent attempts to communicate with the centralized data collector immediately upon report generation, i.e., at detection time. When the host is connected to the enterprise network (either over VPN or on the LAN), we expect the difference between reporting time and detection time to be small, e.g., on the order of seconds or minutes. By contrast, if the host is outside the enterprise network when a report is generated, the reporting time is delayed until the host is brought back inside, which is likely to be much later (e.g., hours or days).

To estimate the fraction of McAfee reports that are generated on the corporate network — which means that the host would potentially become infected inside the enterprise, even with the deployment of various security products on the enterprise network — we examine the difference between reporting time and detection time. Figure 5 plots the cumulative distribution of this time difference across all McAfee reports, as well as for reports generated by hosts in the five countries where the most hosts encountered malware. Overall, only 19.13% of the reports are received by the data collector within five minutes of generation, and 23.06% within 10 minutes. This suggests that the large majority of reports were generated when the host was *outside* the corporate network.

However, we do observe that the fraction of McAfee reports generated inside (or outside) the corporate network varies widely across countries. As shown in Figure 5, 52.46% of the McAfee reports from hosts in Ireland are collected within 10 minutes of detection, while this is true for only 10.91% of the reports from hosts in India. Rather than suggesting that hosts in Ireland are exhibiting more risky behaviors "inside," we believe this reflects differences in culture and working style across regions. For example, some employees only access company resources during regular working hours, while others bring corporate laptops home for both work and personal use.

To investigate this diversity further, we examine VPN logs collected from the enterprise VPN servers (see Section 2.5). Specifically, the VPN logs allow us to infer: (1) the frequency at which corporate machines are being taken outside, and (2) how those hosts are used while outside. Figure 6 shows the cumulative distribution of the number of VPN logins per user and the duration of VPN sessions for the five countries in Figure 5.

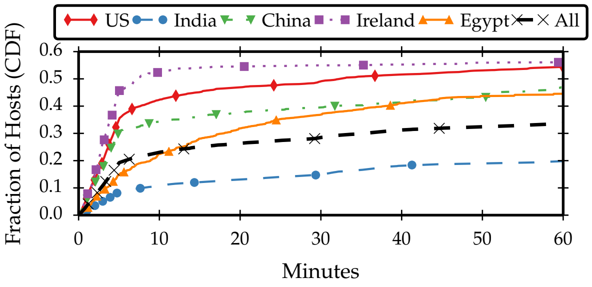

Figure 5: The cumulative distribution of the difference between reporting time and detection time, for the top five countries with the most number of malware-encountering hosts.

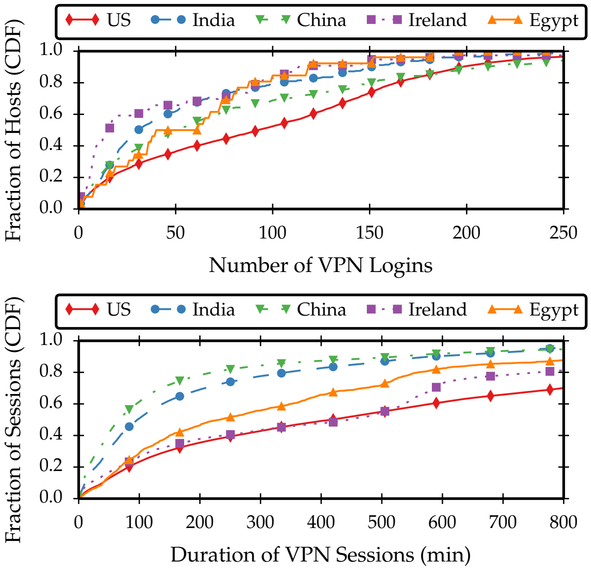

Figure 6: Cumulative distribution of the number of VPN logins per user and the duration of VPN sessions, for the top five countries with the most number of malware-encountering hosts.

We can make several observations by looking at the two extreme countries in Figure 5, Ireland and India. First, users in Ireland login to VPN less frequently than other countries. Half of the users logged in less than 15 times during our four-month observation period, while this number is 31 for India, and 93 for the U.S. This perhaps indicates that employees in Ireland are less likely to bring corporate machines outside of the enterprise network, hence the higher fraction of "inside" McAfee reports in Figure 5. Secondly, while users in India also have a relatively low number of VPN logins, they logon for much shorter durations (e.g., half of the users logged off after 1.5 hours, compared to 7 hours for users in Ireland). Relating this to the fact that around 89% of the McAfee detections in India likely occurred "outside," we conjecture that these employees tend to bring corporate laptops outside the enterprise, though for personal use. Another explanation for the relatively short VPN duration of users in India may be that home Internet connections are less stable, and so employees who bring their laptops home tend to work offline.

In addition to the fraction of McAfee reports generated "inside" and "outside," we are also interested in whether there is a difference between the file system location of the detected malicious file in the two cases. Figure 5 shows a sharp knee in the curves at around five minutes. To be conservative, we use 10 minutes as a threshold; we treat reports whose difference between reporting time and detection time is within 10 minutes as "inside," and the remainder as "outside." Figure 7 shows the distribution of malware locations separately for the "inside" and "outside" cases. The fractions are computed as the number of hosts that reported malware at that file system location while "inside" (or "outside") over all hosts instrumented with McAfee and whose primary user was identified.

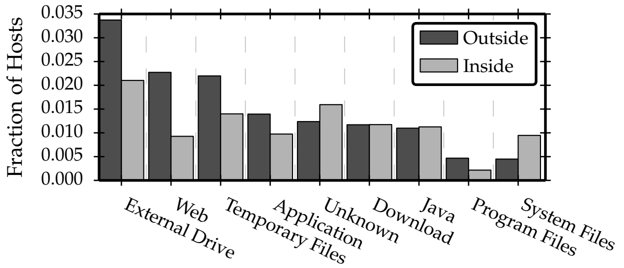

Figure 7: Malware locations for the "inside" and "outside" cases. The fractions are computed as the number of hosts that reported malware at that file system location while "inside" (or "outside") over all hosts instrumented with McAfee and whose primary user was identified.

External drives are the most prevalent malware location in both "inside" and "outside" cases. The fact that the fraction for "outside" is higher, in addition to a non-trivial fraction (around 8%) of those files being multimedia files (based on the file extension, e.g., mp3, avi, jpg), suggests that users do bring corporate laptops home for personal use, as noted above. If true, this result further highlights the risks of mobile devices being physically brought out and back into the enterprise network.

The following two popular malware locations, web browser cache and temporary files, also affected more hosts "outside" than "inside." Since they are commonly associated with drive-by or secondary downloads by initial infections, their relative popularity outside the corporate network is likely caused by the enforcement of strict filtering policies when the host is on the enterprise network.

More significant "inside" than "outside" is malware found under "System Files" and "Unknown" (the fallback category for file paths that match no other location categories). These locations suggest that these McAfee reports are a result of intentional user actions, e.g., to install custom software. Security policies may thus be insufficient in preventing such malware, and targeted user education a more promising approach.

Findings: The large majority of McAfee reports were generated when the host was *outside* the enterprise network. However, the fraction of "outside" reports varies widely by geographic location, possibly caused by differences in culture and working style. The dominant malware locations on the file system are different between the "inside" and "outside" cases, which might be attributed to both user behavior (i.e., bringing corporate laptops home) and the enforcement of security policies on the enterprise network.

3.3 User Demographics

In addition to the malware location and place of detection, we also examine the relationship between user demographics and the likelihood of a host being affected by malware. As described in Section 2.2, the enterprise employee database allows us to infer the user's level in the organizational tree and job type.

Figure 8 shows the fraction of hosts (i.e., whose corresponding user is) at each level in the organizational tree that encountered malware, for all hosts as well as for each country. The company CEO is assigned the root of the tree (level 0), hence the larger the level, the lower down the user is in the management hierarchy. Levels with fewer than five hosts are crossed out. The countries are ordered by the number of hosts in the country.

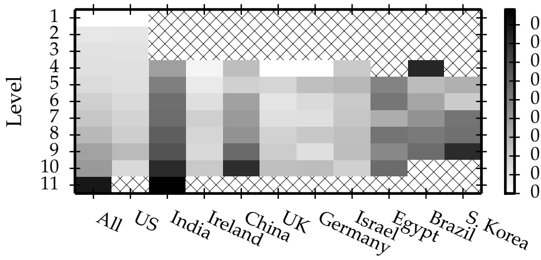

Figure 8: Malware encounter rate by level in the organizational tree, both for all hosts (the first column) and per country. The countries are ordered by the number of hosts in that country. The company CEO is assigned the root of the tree (level 0).

In general, the shading of the boxes becomes darker with the organizational level, particularly for several countries (India, China, South Korea, and except for one outlier at the top of the column, Brazil). We conjecture that a reason for this phenomenon is that employees higher up the organizational tree (i.e., smaller levels) are likely to assume managerial roles and hence exhibit less intensive computer use.

Given an employee's job title, we also categorize his or her job type by taking the last word in the job title after stripping away level indicators (e.g., "engineer I" and "engineer II" are both considered "engineers"). There are 12 job types that have more than 500 employees, covering 84% of all employees whose machine(s) are instrumented with the McAfee client. Appendix D describes each of these job types.

Figure 9 shows the number of hosts (whose corresponding user is) of each job type, as well as the fraction of hosts with that job type that encountered malware. It appears that job types requiring greater technical expertise also have a greater likelihood of encountering malware. Similar results were reported by Lévesque et al. [9] from a small study of 50 users. It is possible that technically savvy users may be more exposed to malware by spending more time with computers or the Internet, or they are potentially less careful about unknown files.

Findings: The likelihood of encountering malware increases the further the employee is from the top of the enterprise organizational tree, and also increases with technical proficiency.

3.4 Web Malware

The web is reportedly the most prevalent vector for attackers and malware authors to gain access to victim hosts [19]. In our dataset, 3.11% of the hosts likely encountered malware by visiting malicious websites, i.e., the detected malicious file was found in the browser cache (see Section 3.1). Among those hosts, 29.85% (583 hosts) generated the McAfee report while they were connected to the corporate network. We are interested in investigating where those malicious files came from, and in particular, why the web proxy that filters web connections from the enterprise failed to block them.

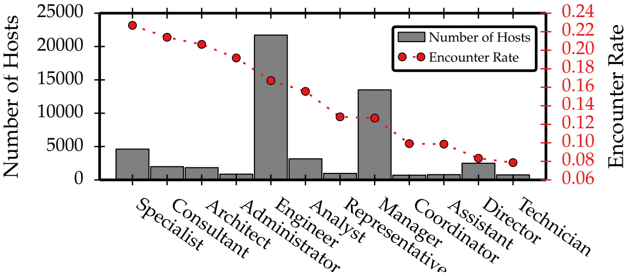

Figure 9: Malware encounter rate by job type. The encounter rate is the fraction of hosts (whose corresponding user has) the job type on the horizontal axis that encountered malware.

The web proxy logs collected in the enterprise contain fields in the HTTP request header, including the destination URL and domain, web referer, user-agent string, as well as auxiliary information like the website reputation and category (provided by the web proxy vendor), and the filtering policy that was applied to that connection (see Section 2.4). Since we know the name of the detected malicious file, we can identify matching URLs in the web proxy logs to obtain details of that connection.

This method only works for files that are stored as-is with the same filename. For some browsers, including Chrome[4] and Firefox[5], the cache is structured as a layered hash table for fast access, where cached files are stored in multiple local files. The name of the detected malicious file, as recorded in McAfee reports, is hence an index to the actual data location, and difficult to reverse-engineer without knowledge of the cache structure and addresses local to the host. In addition, the web proxy logs only store the IP address of the host, while McAfee reports store only the hostname. We further analyze DHCP logs in the enterprise to obtain an IP-to-hostname mapping, and correlate it with the web proxy logs to look up the host assigned that IP during that time (though clock skew and missing or out-of-order logs sometimes cause this lookup to fail). Even with these limitations posed by imperfect data, we were able to match 390 McAfee reports (22% of those found in the browser cache and from hosts inside the enterprise network at the time of detection) with the corresponding web proxy logs.

Table 2 lists the top website categories associated with matched McAfee reports. Surprisingly, none of these categories seem to be intuitively suspicious or malicious. The top category, "Business," is a rather general one that encompasses sites related to marketing, commerce, business practices, human resources, transportation, payroll, and a dozen other services.[6] "Search" includes search engines and portal sites, and "Communities" are sites associated with special interest groups, web newsgroups, and message boards.

Of particular interest is the "Non-Categorized" category, which are new sites that have not yet been given a category label. If an uncategorized site lacking a reputation score is contacted, the web proxy requests that the user acknowledge the enterprise security policies before proceeding. Once the user gives consent, it remains valid for one hour for that host, during which the user can visit other uncategorized sites without further prompt from the proxy.

Website category	Fraction of matching hosts	Fraction of matching reports
Business	29.51%	31.28%
Communities	11.48%	8.46%
Search	9.02%	4.36%
Travel	8.19%	14.87%
Non-Categorized	6.56%	4.62%
SaaS	6.56%	2.31%
Sports	4.92%	3.59%
Food	4.92%	8.46%
Entertainment	3.28%	3.08%
Computers	3.28%	1.03%

Table 2: Top website categories for matching McAfee reports in the "Web" category.

Table 3 shows the web proxy filtering policy that was applied to connections associated with matching McAfee reports. An overwhelming majority (95%) of the matching McAfee reports were allowed through the web proxy, while 2.56% were blocked. This means that although the proxy prevented the connection while the host is inside the corporate network, the malicious file was downloaded when the host was brought outside the enterprise. More interestingly, among the 2.31% matching reports whose connection required user consent, over half (55.56%) were allowed to proceed because of a previous user acknowledgment to a *different* site. Given the small number of encounters we observed under this policy, more data is needed to draw conclusions on its effectiveness.

Policy	Fraction of matching hosts	Fraction of matching reports
Allowed	93.44%	95.13%
Require user consent	6.56%	2.31%
Blocked	0.82%	2.56%

Table 3: Web filtering policies applied to matching McAfee reports in the "Web" category.

Findings: The majority of web-based encounters that we correlated with the web proxy logs are from sites deemed business-appropriate under enterprise policy. 31% of the web-based encounters (around 554 over the four-month duration of our study, extrapolated from the 22% of correlatable encounters) originate from websites in the "business" category and 15% (around 266, similarly extrapolated over four months) originate from the "travel" category.

4. INFERRING THE RISK OF INFECTION

We have observed in Section 3 that there are multiple factors related to the likelihood of a host encountering malware, including demographic features of its user as well as various aspects of its user's behavior. In this section, we develop a statistical model to infer the risk of a host encountering malware proactively. We evaluate the model accuracy and discuss applications to detection and remediation of malware infections in early stages.

4.1 Logistic Regression Model

Motivated by results in Section 3, we extract three categories of features to be used in the model: 1) *Demographic features* capturing information about the user, 2) *VPN activity features* to infer user behavior outside the corporate network (e.g., number of VPN logins, duration of VPN sessions), and 3) *Web activity features* including information about the user's browsing behavior (e.g., categories of web sites visited, web traffic volume). Some of the features are numeric values (e.g., number of VPN logins), and some are categorical (e.g., the country where the user is located). A subset of the features are static (e.g., country, job type), but most vary over time (e.g., number of domains visited or VPN logins).

We are interested in building a predictive model that estimates the conditional probability of encountering malware given the fea-

[4] `http://www.chromium.org/developers/design-documents/network-stack/disk-cache`

[5] `https://code.google.com/p/firefox-cache-forensics/wiki/FfCacheRead`

[6] Although not the same as "malicious", a study of Android apps by Sounthiraraj et al. [18] also found that "business" is the top app category with the most number of vulnerable apps.

ture values at a particular moment in time. We investigated a number of statistical models based on regression (linear, logistic, Poisson, and proportional hazards regression) and found logistic regression to be the most suitable for constructing the predictive model. Logistic regression is used to estimate a conditional probability $Pr(Y|X)$ of a binary response variable Y given a set of input variables $X = (X_1, \ldots, X_n)$. The model assumes that $Pr(Y|X)$ is the logistic function and estimates unknown parameters using the maximum likelihood method.

More precisely, let $p(\vec{x}) = Pr(Y = 1|X = \vec{x})$, for $\vec{x} = (x_1, \ldots, x_n)$. Logistic regression assumes that:

$$\log \frac{p(\vec{x})}{1 - p(\vec{x})} = \alpha + \beta \cdot \vec{x},$$

where α is called *intercept*, $\beta = (\beta_1, \ldots, \beta_n)$ are regression coefficients for the features and $\beta \cdot \vec{x}$ denotes the scalar product of vectors β and $\vec{x}$.

We model the response variable Y as a random variable with value 1 if the host encountered malware, and 0 otherwise. Input variables $X = (X_1, \ldots, X_n)$ denote features modeling various aspects of user demographic information and behavior, whose selection process is described in detail in Section 4.2.

To demonstrate the model's effectiveness, we randomly split the entire host population into two equal-size training and testing data sets. The parameters of the logistic regression model are estimated using the training set, which are used to compute risk scores for hosts in the testing set. We finally demonstrate that among the hosts with highest risk score, a large fraction (more than 50%) encountered malware. Our evaluation results are presented in Section 4.3.

4.2 Feature Selection

We employ a two-stage feature selection process to identify the most relevant features for the model. First, we build a logistic regression model separately for each category of features with the goal of finding the "significant" features to predict malware encounters. Second, we combine the statistically significant features selected in the first stage to build the final model. Here we describe the feature selection process in more detail.

4.2.1 Details on Statistical Model

We use the `glm` function in R for implementing logistic regression. Based on the training data, `glm` outputs estimates of the intercept α and regression coefficients β_i, as well as standard errors for estimation. For each feature i, `glm` also computes the p-value for the hypothesis test that β_i is zero, implemented using the standard Wald test. A low p-value indicates that the null hypothesis can be rejected with high confidence, implying that the feature is relevant in the model. Significance levels of 0.001, 0.01 and 0.05 are denoted by ***, **, and *; a dot (.) denotes a 0.1 significance level; and no star or dot means the feature is not found significant.

For categorical (i.e., discrete) variables, R employs the following encoding scheme. Assume that a variable V takes m possible values $v_1, \ldots, v_m$. Then R encodes this with $m - 1$ binary variables $Z_1, \ldots, Z_{m-1}$. Value $V = v_i$ for $i \in \{1, \ldots, m-1\}$ is encoded with $Z_i = 1$ and all other Z_j binary variables set at 0, for $j \neq i$. Value $V = v_m$ is encoded with all variables Z_i set at 0, for $i \in \{1, \ldots, m-1\}$. v_m is called the *reference value* for V.

4.2.2 Demographic Features

The user demographic features we consider include Gender (inferred user gender), Country (country of user's office), Level (level in the management hierarchy) and Technical (technical level of the user's job type). While the employee dataset does not include user gender, we infer this information from the employees' first names using data from the U.S. census bureau.[7] The gender for 65.59% of the employees was determined this way, with the remaining users labeled as "unknown." Technical level is a binary variable inferred from the job title, set to 1 for "Engineer," "Architect," "Specialist," and "Administrator," and 0 for all other job types. There are over 70 unique countries in our employee dataset. Ordering them by the number of employees, we focus on the top countries that cover 95% of the employee population.

Table 4 shows the estimated coefficients for each feature, the standard error, the p-value for the hypothesis that the regression coefficient is zero, and the significance level.[8] The low p-values for the Gender, Level, Technical variables, as well as for the majority of the binary variables encoding Country, demonstrate that all demographic features considered are significant. The estimated coefficient is correlated with the infection risk of that feature, confirming that India has the highest infection risk, while Japan has the lowest. Six countries (Japan, Ireland, Netherlands, Germany, UK, US) have negative coefficients indicating negative correlation with malware encounters. Another six countries have coefficients close to 0, suggesting no statistical significance regarding infections.

Feature	Value	Est.	Error	p-value	Signif.
Gender	Male	0.23	0.04	1.35e-09	***
	Unknown	0.39	0.04	2e-16	***
Country	Brazil	1.01	0.13	6.82e-15	***
	Canada	0.05	0.13	0.67	
	China	0.92	0.1	2e-16	***
	Egypt	0.86	0.12	1.36e-12	***
	France	0.24	0.13	0.07	.
	Germany	-0.22	0.13	0.08	.
	India	1.21	0.1	2e-16	***
	Ireland	-0.24	0.11	0.03	*
	Israel	0.09	0.13	0.48	
	Italy	0.37	0.14	0.01	**
	Japan	-1.64	0.18	2e-16	***
	Korea	1.02	0.13	2.29e-14	***
	Mexico	0.87	0.15	5.02e-09	***
	Netherlands	-0.23	0.20	0.26	
	Russia	0.52	0.14	3e-04	***
	Singapore	0.19	0.14	0.18	
	Spain	0.57	0.15	e-04	***
	UAE	0.62	0.17	2e-04	***
	UK	-0.24	0.12	0.047	*
	US	-0.02	0.09	0.79	
	Other	0.52	0.10	3.23e-07	***
Levels		0.15	0.01	2e-16	***
Technical		0.08	0.02	7e-04	***

Table 4: Significance of demographic features.

4.2.3 VPN Activity Features

Section 3.2 shows that, with a threshold of 10 minutes between detection and reporting times, around 77% of all malware encounters occurred outside the corporate network. While we have no visibility into users' activities outside the enterprise, VPN usage is an approximate quantitative metric. We extracted the following features to model a user's VPN usage: VPN_conn (total number of connections over the monitoring period), VPN_dur (total duration of all VPN connections in seconds), VPN_sbytes (sum of bytes sent in VPN connections), VPN_rbytes (sum of bytes received in VPN connections), and VPN_extip (number of distinct external IP addresses from which VPN connections are initiated). Intuitively,

[7] `https://raw.github.com/Bemmu/gender-from-name`

[8] For categorical features Gender and Country, values *Female* and *Australia*, respectively, were chosen as the reference values by the `glm` function (and as such are not explicitly included in the table).

users connecting from many different external IPs visit multiple networks, and can be exposed to more attack vectors. These features were aggregated over a subset of the monitoring period, from August 1 to November 10, 2013.

Table 5 shows that almost all features are highly significant in estimating the conditional probability of malware encounters. Surprisingly, VPN_dur is the only feature negatively correlated with malware encounters (i.e., users exhibiting less total time in VPN sessions are at higher risk). One conjecture is that users who bring their machines outside often, but spend less time on VPN, are more exposed to threats since they lack protection by enterprise security products while on those external networks.

Feature	Est.	Error	p-value	Signif.
VPN_conn	4.47e-03	3.22e-04	2e-16	***
VPN_dur	-8.54e-08	8.21e-09	2e-16	***
VPN_sbytes	9.65e-12	1.41e-12	6.96e-12	***
VPN_rbytes	1.38e-13	2.08e-12	0.947	
VPN_extip	1.73e-02	6.62e-04	2e-16	***

Table 5: Significance of VPN activity features.

4.2.4 Web Activity Features

Various aspects of users' web behavior are potentially correlated with malware encounters. We investigate features related to categories of web sites visited, aggregate volumes of web traffic, and connections to blocked or low-reputation sites.

Categories of web sites visited. The web proxy vendor classifies web sites into categories. For each host, we count the total number of HTTP connections to each category of interest, including chat, entertainment, file transfer, filtering, freeware, gaming, gambling, online storage and backup, peer-to-peer, social networks, online mail, streaming, business, travel and non-categorized sites. As described in Section 3.4, non-categorized sites are those that are new and yet to receive a category label. The results in Table 6 show that only seven website categories are relevant in building the statistical model. Among these, four categories in particular (chat, file transfer, social networks and non-categorized sites) contribute significantly to estimating the risk of hosts encountering malware.

Category	Est.	Error	p-value	Signif.
Chat	1.1e-05	2.8e-06	8.84e-05	***
Entertainment	2.01e-08	7.4e-07	0.98	
File transfer	3.62e-06	1.1e-06	9.8e-04	***
Filtering	2.12e-04	2.03e-04	0.3	
Freeware	2.05e-06	6.44e-07	1.4e-03	**
Gaming	1.27e-05	6.48e-06	0.05	.
Online storage	-3.7e-08	e-07	0.71	
Peer-to-peer	2.09e-07	2.57e-06	0.94	
Social networks	4.54e-06	1.04e-06	1.36e-05	***
Online mail	6.45e-07	5.05e-07	0.20	
Streaming	4.48e-07	1.93e-07	0.02	*
Business	2.44e-07	2.59e-07	0.35	
Travel	1.74e-06	1.11e-06	0.12	
Non-Categorized	4.49e-06	8.98e-07	5.65e-07	***

Table 6: Significance of website category features.

Web usage features. We also consider a set of features measuring the aggregate volume of web traffic generated by each host. Intuitively, higher Internet exposure could potentially result in higher likelihood of encountering web-based malware. These features include: No_conn (total number of web connections over the monitoring period), No_doms (number of distinct domains visited by the host), rbytes (sum of the bytes received in all web connections), and sbytes (sum of the bytes sent in all web connections). Table 7 shows that only the number of distinct domains visited by the host is strongly correlated with the probability of encountering malware.

Category	Est.	Error	p-value	Signif.
No_conn	2.14e-08	1.82e-08	0.24	
No_doms	1.84e-06	2.405e-07	2.07e-14	***
rbytes	3.83e-13	9.47e-13	0.69	
sbytes	-6.82e-13	5.91e-13	0.24	

Table 7: Significance of web usage features.

Blocked and low-reputation domains. Accessing blocked or low-reputation sites might be indicative of risky activity. For each host, we count the number of web connections blocked by the proxy (Blocked), the number of connections to non-categorized sites that required explicit user agreement (Challenged), and the number of connections to non-categorized sites to which the user explicitly consented (Consented).

In addition, we maintain a history of all external destinations visited by internal hosts in the enterprise over an interval of three months. This history is updated daily to account for newly visited domains. Connections to *new domains*, i.e., that have not been visited before by any host in the organization, are also possible indicators of suspicious activity. For each host, we count the number of new domains visited each day, and then aggregate these values over the monitoring period into a feature called New_domains.

All of these features are highly significant in the logistic regression model, as shown in Table 8, but the most significant are visits to new domains (New_domains) and number of non-categorized sites requiring user agreement (Challenged).

Category	Est.	Error	p-value	Signif.
Blocked	1.02e-06	3.05e-07	8.1e-04	***
Challenged	7.75e-06	1.59e-06	1.04e-06	***
Consented	2.85e-03	3.46e-04	2e-16	***
New_domains	8.25e-04	1.6e-04	2.56e-07	***

Table 8: Significance of domain reputation features.

4.2.5 Combining Relevant Features

We combine in our final logistic regression model all features found significant in the above analyses, listed in Table 9. We also ran a χ^2 goodness-of-fit test to test the hypothesis that the final model fits the training data set, and obtained a very high p-value (close to 1), implying that the null hypothesis can not be rejected. This finding gives us confidence that the model is a good fit to the features modeling user demographics and behavior, and we present our evaluation results on the predictive power of the model next.

Category	Feature	Description
Demographic	Gender	Gender of user
	Country	Country of user's office
	Level	Level in management hierarchy
	Technical	Technical level
VPN	VPN_conn	Total no. VPN connections
	VPN_dur	Duration of VPN connections
	VPN_sbytes	Bytes sent in VPN connections
	VPN_extip	No. external IPs connecting to VPN
Web	Chat	No. chat sites visited
	File transfer	No. file-transfers sites visited
	Freeware	No. freeware sites visited
	Games	No. gaming/gambling sites visited
	Social networks	No. social-networking sites visited
	Streaming	No. streaming sites visited
	Non-Categorized	No. non-categorized sites visited
	No_doms	No. distinct domains visited
	Blocked	No. connections blocked by proxy
	Challenged	No. connections challenged by proxy
	Consented	No. connections consented by proxy
	New_domains	No. new domains visited

Table 9: Features included in final logistic regression model.

4.3 Evaluation

To evaluate the final model, we split the set of all hosts randomly into equal-sized sets, training and testing. We estimate the parameters of the logistic regression model with the training set, and compute the risk scores of hosts in the testing set. Figure 10 shows the cumulative distribution (CDF) of scores for hosts in the testing set. It is clear that the CDF for malware-encountering hosts is distinct from that for "clean" hosts, with the former having (on average) higher scores than the latter.

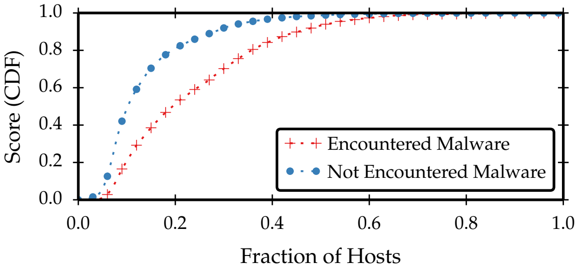

Figure 10: Cumulative distribution of scores for hosts that encountered malware, and those that did not.

This suggests that the user demographic and behavior features can be used to infer the likelihood of malware encounters. As an example application of this result, hosts can be ranked according to their risk score, under a model tuned to the particular organization. Proactive measures can then be applied to hosts with the highest scores so as to detect and remediate potential malware infections.

How well will such a prioritization approach work? To answer this question, we ordered the hosts in the testing set based on the risk score output by the model, and computed the malware encounter rate for the top n hosts. Figure 11 shows our results, averaged over ten independent runs, where each run splits the hosts into training and testing datasets, and builds the model on the training set while computing risk scores for hosts in the testing set. Results for models built with all features and those with each feature category are shown as separate lines in Figure 11.

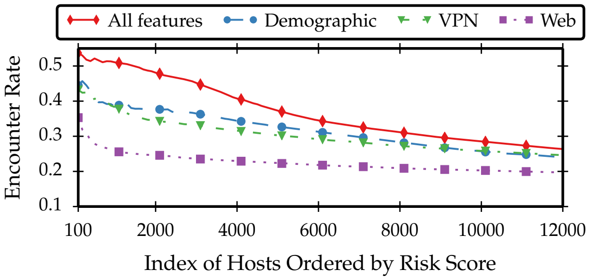

Figure 11: Ordering hosts by their risk score. The malware encounter rate decreases with the hosts' score ranking. Among the top 1,000 hosts, the malware encounter rate is 51% (well more than 3× higher than the overall encounter rate of 15%).

With all 20 features combined, the malware encounter rate is 51% among the top 1,000 hosts—well more than 3× higher than the overall malware encounter rate in the enterprise (15.31%), as described in Section 3. Among the three feature categories, user demographics is the most powerful at indicating risk, followed by VPN behavior. Counter-intuitively, web activity contributes marginally to the overall model. One explanation for this surprising result is that, in our study, only 3.11% of the hosts encountered malware from the web (see Section 3.1), and among those, we have visibility into only a small fraction that happened inside the corporate network.

5. PREVIOUS STUDIES

Several prior works have studied the relationship between the likelihood of malware encounters and users' online behavior or demographic information. The datasets used in previous works vary, as well as the methods by which they identify malware encounters. In this section, we compare our results to those reported in prior studies where possible, and highlight instances in which our findings corroborate or refute theirs. We also summarize some more distantly related works in Appendix C.

Malware encounter rate: The 2013 Microsoft Security Intelligence Report [12], involving over 600 million hosts installed with Microsoft security products in the first half of 2013, reported a worldwide malware encounter rate of around 18%. Our enterprise dataset, consisting of 62,884 hosts instrumented with the McAfee client collected over four months, has a similar encounter rate of 15.31%. On average, our encounter rate during any week is 1.26%.

Other smaller datasets consisting of network packet captures exhibit a similar encounter rate to ours. Maier et al. [10] found, over a period of 14 days, that 1.23% of the users subscribed to an European ISP exhibited scanning or spamming behavior or contacted known malware sites. Carlinet et al. [4] observed 3.04% of the customers of the Orange ISP in France generating traffic that triggered Snort alerts during the course of three hours.

User studies that surveyed or observed users showed a much higher malware encounter rate. Among 295 university students surveyed by Ngo and Paternoster [14], 46% reported encountering malware at least once in the last year, while 38% of 50 participants in a study by Lévesque et al. [9] were found to be infected over the course of four months.

The lower encounter rate in our dataset compared to prior works may be due to the network policies enforced in the enterprise. A whitelist and blacklist are applied to outbound network connections, and employees are not allowed by default to install software on enterprise-managed hosts. It is also possible that Lévesque et al. and Ngo and Paternoster observe higher encounter rates due to the populations they study (primarily students).

User Demographics: Ngo and Paternoster found that both age and race are significantly related to the likelihood of encountering malware, but not gender, marital status, or employment status. While agreeing that those factors do not contribute to the risk of infection, Lévesque et al. also found age to be irrelevant. The only significant factor they identified is technical expertise.

Our employee dataset does not include personal information (age, gender, etc.), and obviously the users in our data were employed. That said, we did corroborate the observation that technical expertise correlates with the likelihood of encountering malware.

User Behavior: Both Maier et al. and Ngo and Paternoster examined the relationship between the use of security products and malware encounters. The former found that neither the installation of anti-virus scanners nor regular O/S and blacklist updates reduced malicious activities. However, the latter found that having security software led to a higher probability of encountering malware (perhaps an artifact of the user study, as users would otherwise not know that they were infected). The nature of our dataset prevents us from observing hosts with varying software configurations, since enterprise hosts are centrally configured and managed in our case.

One might expect that more exposure to the Internet would result in higher likelihood of malware encounters. This was the case in the studies of Carlinet et al. and Lévesque et al., as well as of Canali et al. [3] in terms of traffic volume and the number of unique websites visited. It also holds true in our dataset. In addition, we find some categories of websites to be correlated with higher malware encounter rates than others, as did Carlinet et al. and Lévesque et al., although the categories do not always agree. Canali et al., on the other hand, use logistic regression to show that browsing time, duration and number of distinct domains are better indicators of a user being at risk than the domain category. However, their definition of *at risk* is laxer and does not rely on documented malware encounters.

6. LIMITATIONS

By the nature of the data available to us, our conclusions are subject to a number of caveats. Perhaps most importantly, since we relied on McAfee reports to indicate malware encounters, our results do not reflect any potential encounters detected and eliminated prior to reaching McAfee analysis—e.g., by firewalls/IPS, web browsers, or any of the other myriad security defenses commonly employed in IT infrastructure or leveraged by this enterprise in particular—or that reached McAfee but evaded it. It is well known that no anti-virus solution detects all malware; at best, it identifies a large fraction of "mass-market" malware. As such, we assume that our encounters are biased toward mass-market malware that entered the enterprise via poorly defended vectors.

Another caveat related to our data is that we do not have ground truth for many aspects of our investigation, requiring us to leverage indirect indicators, instead. So, for example, we used McAfee reporting delays to infer whether each malware encounter occurred while the computer was on the corporate network (Section 3.2), and we leveraged job titles as a surrogate for an objective measure of technical proficiency (Section 3.3). It is important to bear in mind that all such inferences come with a level of error to them, though our belief is that the relatively large amount of data in our study provides some statistical evidence for the correlations that we observe.

Unknowns will be a factor in virtually any study of this type, introducing questions about the extent to which any single study—including this one—will be representative more broadly. We believe that it is necessary to assemble a broad set of such studies to reveal their common elements and differences. (We have attempted to draw out such similarities and differences with other studies in Section 5.) Only through repetition can lasting trends be identified.

7. RECOMMENDATIONS FOR THE ENTERPRISE

Our study suggests recommendations for reducing the rate of malware encounters in enterprise settings:

Proactive scanning, alert prioritization. Enterprises often deploy memory analysis tools on hosts. As use of these tools is labor-intensive, they must be deployed selectively. Our logistic regression model identifies hosts with a highly elevated risk of malware encounters (51% among the top 1,000 such hosts). This allows an enterprise to apply memory-scanning tools proactively, facilitating early detection of malware infection. Similarly, an enterprise can prioritize investigation of alerts (e.g., generated by log-analysis tools) based on modeled host risk.

User education. User education has the potential to reduce behaviors responsible for a significant fraction of malware encounters. A number of studies have affirmed the successes of carefully targeted educational campaigns, e.g., [7]. Our study highlights several opportunities for such targeting, e.g., educating users in low GDP countries to avoid the use of USB sticks with company laptops.

Caution with site categories. In the studied enterprise, most web-based malware encounters permitted by the web proxy originate from sites categorized as business-appropriate under enterprise policy. This suggests that currently deployed website content categories are an inadequate basis for access restrictions. Tightening access policies by, e.g., refining website categories and/or applying per-user (or per-group) rules, may reduce web-based encounters.

Understanding the practical implications and effectiveness of any recommendation would require in-field studies, and likely differ across enterprises based on their policies and compliance requirements. We hope our study will motivate future research in these areas.

8. CONCLUSION

We have presented the first large-scale epidemiological study of malware encounters in the setting of a large enterprise. Our study offers several key findings, including the preponderance of malware encounters outside the enterprise network, differences across geographies in the most important vectors of malware propagation, differences in encounter rates according to employee position in the management hierarchy, and a significant risk of web-based malware encounters that originated with sites categorized as safe and business-appropriate by the enterprise web proxy. While our study corroborates some findings in earlier research, it also sheds new light on the special characteristics of enterprise malware penetration and shows how these characteristics can be combined in a logistic regression model to achieve accurate identification of at-risk hosts. Finally, our study suggests promising, concrete policy- and education-based approaches to driving down malware encounter rates in enterprise environments.

Acknowledgments

We are grateful to the anonymous enterprise who permitted us access to their data for the purposes of this study. We are also grateful to James Lugabihl, Robin Norris, Todd Leetham, Christopher Harrington, Garrett Schubert, Justin Lamarre, Andrew Rutkiewicz, Michael Blanchard, Ronald L. Rivest and members of RSA Laboratories for their many useful comments and suggestions on designing the experimental framework in the paper. We thank our shepherd Davide Balzarotti and anonymous reviewers for their feedback on drafts of this paper. This work was supported in part by NSF grant 0831245.

9. REFERENCES

[1] "epidemiology". In *Merriam-Webster.com*, 15 May 2014.

[2] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring pay-per-install: The commoditization of malware distribution. In *20th USENIX Security Symposium*, Aug. 2011.

[3] D. Canali, L. Bilge, and D. Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *9th ACM Symposium on Information, Computer and Commmunications Security*, June 2014.

[4] Y. Carlinet, L. Mé, H. Debar, and Y. Gourhant. Analysis of computer infection risk factors based on customer network usage. In *2nd International Conference on Emerging Security Information, Systems and Technologies*, pages 317–325, Aug. 2008.

[5] M. P. Collins, T. J. Shimeall, S. Faber, J. Janies, R. Weaver, M. De Shon, and J. B. Kadane. Using uncleanliness to predict future botnet addresses. In *7th ACM Internet Measurement Conference*, pages 93–104, Oct. 2007.

[6] A. Kleiner, P. Nicholas, and K. Sullivan. Linking Cybersecurity Policy and Performance. Microsoft Trustworthy Computing, 2013.

[7] M. W. Kreuter and R. J. Wray. Tailored and targeted health communication: Strategies for enhancing information relevance. *American Journal of Health Behavior*, 27:S227–S232(6), November 2003.

[8] M. Lee. Who's next? identifying risks factors for subjects of targeted attacks. In *Proc. Virus Bull. Conf*, pages 301–306, 2012.

[9] F. Lévesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *20th ACM Conference on Computer and Communications Security*, Nov. 2013.

[10] G. Maier, A. Feldmann, V. Paxson, R. Sommer, and M. Vallentin. An assessment of overt malicious activity manifest in residential networks. In *Detection of Intrusion and Malware, and Vulnerability Assessment, 8th International Conference*, pages 144–163, July 2011.

[11] Microsoft. Security Intelligence Report. http://www.microsoft.com/security/sir/default.aspx, 2011.

[12] Microsoft. Security Intelligence Report. http://www.microsoft.com/security/sir/default.aspx, 2013.

[13] G. R. Milne, L. I. Labrecque, and C. Cromer. Toward and understanding of the online consumer's risky behavior and protection practices. *Journal of Consumer Affairs*, 43:449–473, 2009.

[14] F. T. Ngo and R. Paternoster. Cybercrime victimization: An examination of individual and situational level factors. *International Journal of Cyber Criminology*, 5(1):773–793, 2011.

[15] K. Onarlioglu, U. O. Yilmaz, E. Kirda, and D. Balzarotti. Insights into user behavior in dealing with internet attacks. In *Network and Distributed System Security Symposium (NDSS)*, 2012.

[16] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *2006 ACM SIGCOMM*, pages 291–302, Sept. 2006.

[17] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *ACM Conference on Human Factors in Computing Systems*, pages 373–382, Apr. 2010.

[18] D. Southiraraj, J. Sahs, G. Greenwood, Z. Lin, and L. Khan. Smv-hunter: Large scale, automated detection of ssl/tls man-in-the-middle vulnerabilities in android apps. In *2014 NDSS Symposium*, 2014.

[19] Symantec Corporation. Internet security threat report. `http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_appendices_v18_2012_221284438.en-us.pdf`, 2013.

[20] M. Vasek and T. Moore. Identifying risk factors for webserver compromise. *Financial Cryptography and Data Security*, 2014.

[21] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: Signatures and characteristics. In *2008 ACM SIGCOMM*, pages 171–182, Aug. 2008.

[22] J. Zhang, Z. Durumeric, M. Bailey, M. Liu, and M. Karir. On the mismanagement and maliciousness of networks. In *2014 NDSS Symposium*, 2014.

APPENDIX

A. MALWARE TYPE

According to McAfee's virus naming convention,[9] malware names consist of a prefix and suffix. The prefix specifies the type of file or platform the malware targets (e.g., JS, PDF, W32), or its "class" (e.g., Adware, Backdoor, Exploit, KeyLog). The suffix(es) designate variants of a malware, the byte size of the binary, or additional information about its type. In our McAfee dataset, there are 9,577 unique malware names. By stripping away the suffix we are left with 1,097 unique malware names.

The top two malware names, encountered by 5.47% and 1.73% of the hosts, respectively, are "Artemis" and "Generic." The former is not a malware family, but McAfee's name for those detected heuristically. The latter seems to be a generic detection that lacks addition information. Excluding these two, the encounter rates of the top 20 malware are shown in Figure 12. The shading of the boxes denote the encounter rate for that malware in that country. Rows (and columns) are sorted according to a hierarchical clustering algorithm to minimize the Euclidean distance between adjacent rows (and columns).

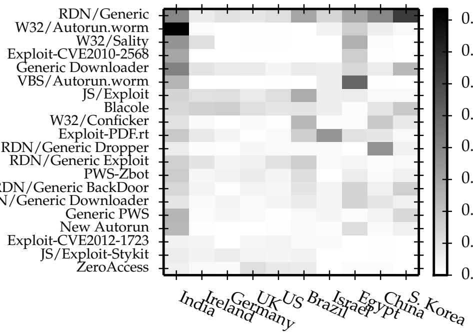

Figure 12: The encounter rate in each country for the top 20 malware. The top ten countries with the most number of malware-encountering hosts are plotted.

Some patterns emerge from Figure 12. First, the malware encountered differs by geographic location. Some can be found in all countries (e.g., "RDN/Generic," "Generic Downloader"), while others are specific to certain locations (e.g., Autorun malware and worms in India and Egypt, and exploits in Western countries). Secondly, the diversity of malware in each country varies widely. Hosts in India encountered more than 300 different malware, while those in the U.K. only encountered 72. This suggests that the difference in the encounter rate per country may be partially due to the abundance of malware, and the types of malware, in that region.

To gain further insight about the categories of malware, we leverage the "class" information assigned by McAfee. The class keyword, e.g., Adware, Backdoor, Exploit, when available, is included in the prefix of the malware name. We find the malware class present in around 29.23% of the McAfee reports. Figure 13 shows the encounter rate of each malware class in each country.

[9] `http://download.nai.com/products/datfiles/4.x/nai/readme.txt`

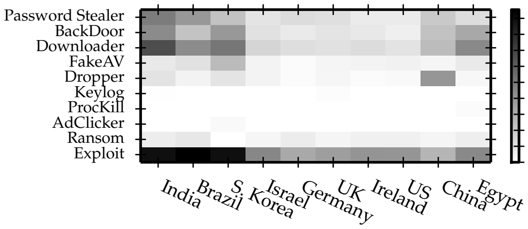

Figure 13: The encounter rate of each malware class in each country. The top ten countries with the most number of malware-encountering hosts are plotted.

The top malware class, "Exploit," is encountered by 3.67% of the hosts. It seems to be especially common in India, Brazil, and South Korea, who share a similar makeup of malware classes. China stands out as having the highest "Dropper" encounter rate, perhaps due to the abundance of custom, free software available online in that region.[10]

Findings: Malware types differ by geographic location, some targeting specific regions while others are common to all countries. Exploits are the most common malware class in our dataset, particularly prevalent in India, Brazil, and S. Korea. Droppers are mostly found in China (possibly related to the abundance of custom, free software available online in that region). Exploits primarily target vulnerabilities in Javascript and Java, though a non-trivial fraction of hosts also encountered PDF exploits and those targeting the Windows Shell.

B. FILE SYSTEM LOCATION CATEGORIES

We group McAfee reports into the following categories, based on the paths of detected malicious files:

- **External drives:** The file was found on high-lettered drives (i.e., F and above) or is named "autorun.inf" located directly in the root directory of the drive. Even though personal machines can have multiple internal, physical drives, enterprise-owned PCs were placed under a relatively restrictive configuration policy. All were shipped with identical configurations except in special cases. Employees were also not given administrator privileges on their machines by default. As a result, the vast majority of hosts have only drives C and D.
- **Temporary files:** The file was found under a temporary folder in the application directory (e.g., "`C:\Users\User Name\AppData\Local\Temp`"). This folder is commonly used to store secondary downloads by an initial infection.
- **Web cache:** The file was found in the browser's cache, e.g., in the "`Temporary Internet Files`" directory, or the "`cache`" folder under the browser's directory. In this case, the malware likely arrived on the victim through drive-by downloads.
- **Download:** The file was found on the user's desktop or in the default folder storing downloaded files (e.g., "`C:\Users\User Name\Downloads`"). This is often associated with intentional downloads performed by the user.
- **Application:** The file was found under the applications data directory, which stores user-specific application information, including configuration files, default templates, etc. An example directory path is "`C:\Documents and Settings\User Name\Application Data\`."
- **Java:** The file has a ".class" extension, or found in the system directory for Java.
- **System:** The file was found in the Windows system folder (e.g., "`C:\Windows\System`). This can also be indicative of secondary downloads performed by an initial infection.
- **Program files:** The file was found under the "Program Files" directory, which stores user-installed applications that are not part of the O/S.
- **Recycle:** The file was found in the recycling bin.
- **Backup:** The file was found in the directory where the O/S stores restore points for recovery purposes. E.g., "`C:\System Volume Information`," or "`\Device\HarddiskVolumeShadowCopy`."
- **Network drives:** The file was found on a network drive, i.e., the path starts with two backslashes.
- **Unknown:** File paths that do not match any of the above. Many of these files are found in directories created by the user.

C. ADDITIONAL RELATED WORK

In addition to the studies discussed in Section 5, several other studies have been conducted that deserve mention, albeit while using different methodologies and providing less directly comparable results.

Regional differences: In a study of malware distributed via pay-per-install services, Caballero et al. [2] witnessed families of malware preferentially delivered to the U.S. and Europe, and others exclusively targeted to a single country. Kleiner et al. [6] examined the impact of socio-economic factors in a country or region on malware infections. They also found countries that implement policies for investigating and prosecuting cybercrime offenses to have a lower infection rate.

Internet-wide studies: Numerous studies have investigated the characteristics of malware proliferation by studying traffic sent by apparently infected machines (scans, spam, denial-of-service packets, etc.) on the public Internet. For example, using the malicious traffic observed at the border of a large network, Collins et al. [5] demonstrated the tendency of malware infections to cluster within the same networks (identified by CIDR blocks) over time and how this tendency can be used to predict where such infections will likely occur in the future. Ramachandran and Feamster [16] and Xie et al. [21] studied spam feeds to quantify, at the granularity of autonomous systems, where spam bots most often arise; the Xie et al. study further noted that the top five autonomous systems by this measure are all Internet service providers that offer residential network access. Zhang et al. [22] also found that "mismanaged" autonomous systems, such as those that have open DNS resolvers, lack egress filtering, allow untrusted HTTPS certificates, etc., are more likely to be responsible for malicious activities.

User behavior: Also distantly related are ethnographic studies focused on other online behaviors and threats. For example, Sheng et al. [17] used Mechanical Turk to evaluate how gender, age, technical knowledge, risk perception for financial investment, and prior exposure to anti-phishing training impacted susceptibility to phishing attacks in a role-playing exercise. Via an online survey of 449 participants, Milne et al. [13] studied the relationships of both participant self-efficacy and demographics (e.g., age) to the participants' online behaviors. Lee [8] conducted an epidemiological study in an academic environment and determined that the employees' department and job type are indicative of their susceptibility to targeted phishing attacks. In an academic study with 164 participants, Onarlioglu et al. [15] found that non-technical users demonstrated an ability comparable to that of security experts in

[10] `http://www.infosecurity-magazine.com/view/35047/googlebacked-filesharing-service-spreads-chinese-malware/`

averting frequently encountered threats, but performed poorly at detecting more unusual and sophisticated ones.

Server compromises: Orthogonal to our study of host malware encounters, Vasek and Moore [20] explored factors relating to the compromise of web servers. Specifically, they quantified the relationship between compromise and the type of web server, type of content management system (CMS), hosting country, and secure administration practices demonstrated on the website using logistic regression models.

D. EMPLOYEE JOB TYPES

In this section, we summarize each job type introduced in Section 3.3, including the fraction of all users with that job type and a brief description of the job type.

- **Engineer (27.88%)** Software engineers comprise 28% of the employees under this job type, and systems engineers another 23%. These are followed by support engineers (16%), service engineers (11%), quality engineers (8%). There are also hardware engineers (2%), solutions engineers (1%), test engineers (0.75%), field engineers (0.54%), performance engineers (0.36%), etc.
- **Manager (19.32%)** Account managers, which are under the sales department, make up a quarter of the "manager" employees, and project, program, and product managers another quarter. Engineering managers make up another 9% of the employees of this type, and customer service and tech support 5%.
- **Specialist (6.52%)** The majority or employees with this job type (70%) are responsible for maintaining special hardware and systems, or providing support for integrating products into customer environments. Another 11% of the employees are associated with business operations, sales, and marketing.
- **Analyst (4.67%)** 55% of the employees with this job type are business operations analysts or financial and revenue analysts. There are also process operations (7%), engineering (6%), maintenance (6%) analysts, and others that deal with processes, inventory, supply chain, logistics.
- **Consultant (4.41%)** About half of the employee with this job type simply have "Consultant" as their job title, hence it is difficult to determine their expertise. However, we do find 15% of the employees with this job type to be practice or advisory consultants, 9% to be business consultants, and 7% to be technical consultants.
- **Director (3.17%)** Directors are typically a level up from managers, and span across all business units in the enterprise. 18% of the employees with this job type are in sales or marketing, 16% in business and products, 15% in engineering.
- **Architect (2.81%)** 84% of the employees of this type are Solutions Architects, who are primarily responsible for designing and integrating hardware and software systems. Another 9% of the employees of this job type simply have "Architect" as their job title. There are also technology, application, information, data, storage, and network architects.
- **Representative (1.58%)** These employees are in the sales department, e.g., account, renewals, and sales representatives.
- **Technician (1.40%)** 45% of the employees with this job type are either test or debug technicians, 12% are engineering technicians, and 11% are customer support technicians. Employees with this job type appear to deal mostly with hardware systems rather than software.
- **Administrator (1.15%)** 41% of the employees with this job type are system administrators, 22% are storage or database administrators. There are also contract administrators (6%), district administrators (5%), account administrators (3%).
- **Coordinator (1.12%)** These employees deal with logistics and processes in the company. 30% of employees of this job type are program or project coordinators, 21% are distribution coordinators, 12% are human resources coordinators, and 11% are inventory coordinators.
- **Assistant (1.07%)** 94% of employees with this job type are executive and administrative assistants that support administrative duties in the office. There is also a small fraction of legal assistants, marketing assistants, and sales assistants.