Title

Uncovering Footprints of Malicious Traffic in Wireless/Mobile Networks

Authors

Raghuramu, Arun
Pathak, Parth
Zang, Hui
et al.

Publication Date

2016-04-01

Peer reviewed

Uncovering the Footprints of Malicious Traffic in Wireless/Mobile Networks

Arun Raghuramu, Parth H. Pathak

Department of Computer Science, University of California, Davis

Hui Zang

Huawei Inc.

Jinyoung Han, Chang Liu, Chen-Nee Chuah

Department of Electrical & Computer Engineering, University of California, Davis

Abstract

This paper presents a measurement study that analyzes large-scale traffic data gathered from two different wireless scenarios: cellular and WiFi networks. We first analyze packet traces and security event logs generated by over 2 million devices in a major US-based cellular network, and show that 0.17% of mobile devices are affected by security threats. We then analyze the aggregate network footprint of malicious and benign traffic in the cellular network, and demonstrate that statistical network features (e.g., uplink data transfer volume, IP entropy) can be effectively used to distinguish such malicious and benign traffic. We next investigate over 2.4 TB of WiFi traffic data, which are generated by 27 K distinct users, in a university campus network. Based on the lessons learned from a comprehensive exploration of a large feature space consisting of over 500 statistical attributes derived from network traffic to/from malicious and benign domains, we propose a novel, in-house traffic screening method, which has the capability of effectively identifying potential malicious domains. Our method achieves over 90% accuracy with only using a small set of simple statistical

Email addresses: [araghuramu,phpathak]@ucdavis.edu (Arun Raghuramu, Parth H. Pathak), huizang@gmail.com (Hui Zang), [rghan,cchliu,chuah]@ucdavis.edu (Jinyoung Han, Chang Liu, Chen-Nee Chuah)

network features, without using any additional specialized datasets (e.g., geo-location database) or resource-intensive solutions (e.g., DPI boxes to collect HTTP traffic.).

---

1. Introduction

The pervasive use of mobile devices such as smartphones to access an array of personal and financial information makes them rich targets for malware writers and attackers. Studies have revealed threats and attacks unique to mobile platforms, such as SMS and phone call interception malwares [1]. The claims about prevalence of mobile malware were recently disputed when Lever et. al [2] showed that mobile malware appears only in a tiny fraction of devices in their dataset: 3492 out of 380 million (0.0009%), concluding that mobile application markets are providing adequate security for mobile device users. However, their work did not provide a comprehensive view of malicious network traffic since their analysis was limited to the threats that issue DNS requests to known malicious domains. Also, they did not quantify the prevalence of specific types of threats affecting the network in their characterization study.

In this paper, we perform a detailed characterization of malicious traffic generated by mobile devices using packet traces and security event logs from a major US-based cellular network. Our analysis reveals that 0.17% of over 2 million devices in the cellular network triggered security alerts. This fraction, while still small, is much higher than the previous infection rate reported in [2] and is in agreement with recent direct infection rate measurements focusing on the Android platform [3]. This alarming infection rate calls for a more careful and thorough study of malicious traffic in the mobile ecosystems.

A second area of our focus deals with the problem of 'detecting' malicious hosts/URLs. Previous studies such as [4, 5] treat this as a supervised learning problem where a classifier learns on a combination of DNS, WHOIS, lexical, and other features associated with a given host to decide whether it is malicious or benign with high accuracy. Other studies such as [6, 7] exclusively

utilize lexical features to achieve similar goals. A different approach, Nazca [8], was proposed recently to detect malware distribution networks by tracking web requests associated with malware downloads and installations.

Instead of focusing on features associated with the malware or hosts (e.g., URLs), we examine features based on network traffic to/from malicious domains/hosts associated with the detected threats in a cellular network. We observed that there are distinctive network access patterns that can be leveraged to distinguish between benign and malicious sites.

We then present an in-depth analysis on the network-level features of malicious domains using a large-scale WiFi traffic dataset from a university campus. Based on the lessons learned, we design an in-house domain screening technique that can accurately detect malicious domains by mining network traffic. Such a technique can be used by operators to augment their existing security capabilities (such as firewalls, IDS etc.) or to complement other detection methods such as those based on lexical features. Also, the domains screened via our proposed technique can be reported to third-party systems that can further scrutinize them with more advanced techniques and/or specialized auxiliary datasets (e.g., geo-location database).

To summarize, the contributions of our work are three-fold:

a) We provide a large-scale characterization of malicious traffic by analyzing traffic records and security alerts of over 2 million devices in a US-based cellular network. In addition to revealing higher infection rate, we show that four classes of threats - privacy-leakage, adware, SIP attacks and trojans - are the most prevalent in mobile devices. Also, we find that 0.39% of Android devices are infected, while the infection rates of BlackBerry and iOS devices which are commonly considered more secure are observed to be comparatively high (0.32% and 0.22% respectively).

b) We analyze the aggregate network-level features of cellular traffic for malicious and benign domains accessed by user devices. We demonstrate that the network traffic based features are complementary to lexical features

and hold promise to add to the capabilities of existing malicious domain detection rules.

c) By analyzing over 2.4 TB of WiFi traffic from a university campus network, we perform a comprehensive exploration of a large feature space consisting of over 500 statistical attributes derived from network traffic to/from malicious and benign domains. Using an enhanced feature set and methodology, we design an effective machine-learning classifier that is capable of identifying malicious domains with an accuracy of over 90% utilizing only 20 network traffic features. These results can provide important implications on mobile network operators since they can leverage network traffic to perform effective malicious domain screening without the need for specialized datasets (e.g., geo-location database) or resource-intensive solutions (e.g., DPI boxes to collect HTTP traffic.).

The remainder of the paper is organized as follows. Section 2 provides an overview of our datasets and methodology. In Section 3, we present the findings of our characterization study of mobile threats. Sections 4 and 5 investigate how to detect the malicious traffic by exploring their nature of network footprints in a cellular and WiFi networks, respectively. After discussing related work in Section 6, we conclude the paper in Section 7.

2. Data Summary & Methodology

We utilize datasets obtained from two different operational wireless network environments for our analyses: (a) A US based cellular carrier network environment and (b) A large university campus WiFi network. We now describe each of these datasets in more detail.

2.1. Cellular Network Data

This dataset, collected at a distribution site operated by a US cellular service provider, is multiple terabytes in size and logs HTTP activities of over two million subscribers for a week-long period in summer 2013. What makes

the dataset more interesting is the associated security alert logs generated by commercial systems deployed in the network.

Specifically, the following traces are contained in our dataset:

- Deep Packet Inspection (DPI) Records: These records log HTTP activity of subscribers in the network and contain flow level information associated with each HTTP request, such as, the timestamp, duration, bytes transmitted in each direction, source IP address, URL, and User Agent of the flow.

- Intrusion Detection System (IDS) and Anti-Virus (AV) Alert Logs: These logs contain threatname (usually vendor specific), subscriber IP address, timestamp, destination HTTP domain, and destination port of the alerted activity.

- IP Assignment Records: These records map dynamically assigned IP addresses to anonymized subscriber device IDs.

- VirusTotal, McAfee scan results: We performed additional scans on certain domains and IP's in the IDS and AV logs to obtain additional information about the threats and number of malware detection engines flagging it as positive (malicious).

2.1.1. Identifying Cellular Devices and Platforms

The events in our malicious traffic alert database could have been triggered by either mobile devices such as smartphones and tablets or laptops and desktops that connect to the cellular network via hotspots/modem devices. We were provided with the registered make, model and operating system information for about half of the anonymized subscribers in the trace. For the other subscribers, we infer the device type, make, and OS type using the User-Agent fields from their DPI records with the help of an in-house tool[1]. The devices

[1]This utility analyzes every User-Agent string in the DPI trace associated with the unknown device to make an estimate of its make, model and platform.

Day	Unique Users	Unique Devices	Size
Day 1	22,453	32,088	954GB
Day 2	21,930	31,255	919GB
Day 3	18,698	26,158	615GB
Total	**27,292**	**41,397**	**2.42TB**

Table 1: Summary of WiFi Network Traffic Traces.

in our alert datasets are then classified manually as one of the four general categories: phones, tablets, hotspots/modems and other devices.

We would like to note that the availability of the data from the carrier's network was limited due to its proprietary nature. We also note that we can map traffic records to devices generating them uniquely using anonymized device registration identifiers and NAT (Network Address Translation) logs provided by the carrier.

2.2. WiFi Network Data

As noted earlier, in addition to the cellular traffic data, we collect network traces from WiFi controllers that connect and control the WiFi Access Points (APs) at a large university campus. The controllers connect the APs to the campus backbone network allowing the wireless devices (laptops, smartphones, tablets etc.) to access Internet. The network traces were collected from controllers dedicated for different locations such as residential dormitories, offices, classrooms, cafeterias etc. We collect over 2.4 TB of packet captures generated by 27,292 distinct WiFi users over a three days period from nearly 1,000 campus APs in April 2014. Table 1 provides a summary of the network capture data we use in our analysis. Also, we obtain an auxiliary set of network session logs: each entry in a session log represents a user session with information about the user-name, device MAC address, IP address, and WiFi session start and end times.

Note that all the network traffic traces and logs are anonymized and any personally identifiable information are removed. We have anonymized the IP

Data source	Alert triggering event(s)
IDS-1	DNS requests seen to known malicious domains
IDS-2	(a) The HTTP request header contains a known malicious user agent string or URI (b) Leakage of IMEI, IMSI, Phone number or location information through a HTTP message. (c) Attempts to connect to a known C&C server. (d) DNS request to a known malicious domain (Utilizes a different set of malicious domains from IDS-1). (e) Known malicious behavior. Eg. Attempt to trigger a DDoS, replay attack, etc.
AV-1	Known malware detected on a device through a signature.

Table 2: Security data sources and their alert triggering mechanism.

and MAC addresses, user names, and device IDs. For anonymizing IP addresses, we use a prefix-preserving anonymization as proposed in [9]. We also note that there is no impact of NAT on our study since we can uniquely identify each device connecting to the network with the anonymized session log.

2.3. Building Ground Truth for Malicious Traffic

We now discuss how we generate the ground truth for malicious traffic in cellular and WiFi datasets.

2.3.1. Cellular Network Malicious Traffic

As mentioned earlier, the cellular carrier deploys two separate commercial IDS's in its premises. Each IDS passively monitors different characteristics of traffic and flags security events without initiating any 'block' actions. We utilize logs produced by these appliances in our characterization study. We also use records logged at AV scanners deployed at select end-client devices as an additional auxiliary source of security evidence. Table 2 describes the alert triggering mechanism of these IDS and AV systems.

We collect IP's and URL's associated with the alert events and submit them to commercial URL scanners such as VirusTotal [10] to eliminate false positives and to gather detailed information about the threats associated with these alerts.

The VirusTotal service scans a submitted domain over a corpus of 61 different website/domain scanning engines and datasets (at the time of data collection) in its backend systems, and responds with the aggregated scan result. Scanning a domain through VirusTotal is logically equivalent to checking the status of the domain from multiple security data sources. This was one of the main motivations behind choosing the VirusTotal service. In addition to using this commercial service, we manually group the most prominent threats in the network into four general categories or "Threat classes" as: Trojans, Privacy leakage threats, Potentially Unwanted Applications(PUA) and SIP threats based on the common characteristics and infecting behavior of the threats.

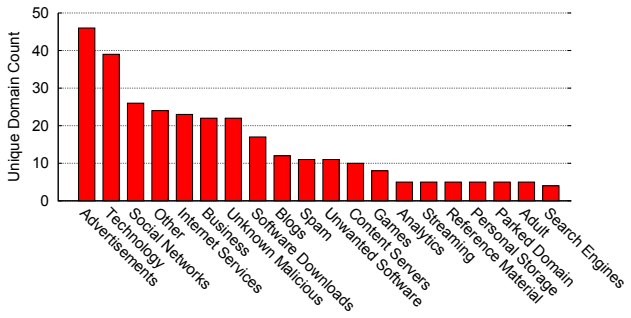


Figure 1: Top web categories of known-malicious domains.

2.3.2. WiFi Network Malicious Traffic

To build the ground truth of whether a domain captured in our WiFi network trace is malicious or not, we submit all the domains captured in our traffic to VirusTotal and capture the response. Note that this is different from Section 2.3.1 above since we do not have data from IDS systems in the campus WiFi network.

We identified a set of 305 "known malicious" domains based on the aggre-

gated responses from VirusTotal. To minimize the chances of a chosen sample being a false positive, we only consider the domains that have at least two detections (or confirmations) from VirusTotal engines. The 305 domains in this set consist of 193 malware domains, 15 phishing domains, and 97 domains known to be engaging in other malicious activities such as involvements with web spam campaigns, adult malvertisements etc. Figure 1 presents the web categories of domains in the "known malicious" set. We also form a set of 20,000 "known benign" domains chosen from the samples with zero detections from VirusTotal. We use these labeled data along with the WiFi network traces and the auxiliary network logs for our supervised classification experiment.

3. Characterizing Mobile Threats

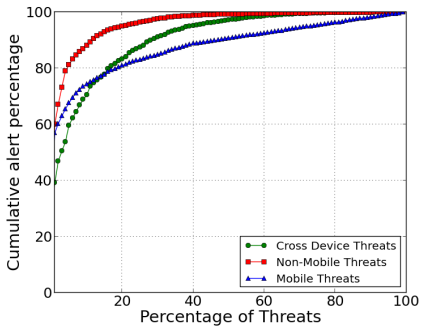
3.1. Prevalence of Malicious Traffic in the Cellular Network

Unlike Lever et. al [2] we do not limit our characterization study to HTTP/ HTTPS traffic generating DNS requests to malicious domains. The security systems in the cellular network in addition to observing web and DNS traffic, also monitor security events generated by VoIP traffic (running over the SIP protocol on ports 5060, 5061) and a number of non-standard ports used by known malware/trojans (eg. Ports 22292, 21810 used by the ZeroAccess Trojan [11], Port 7776 used by Backdoor.Remocy Trojan [12], Port 8080 used by worms such as MyDoom [13]).

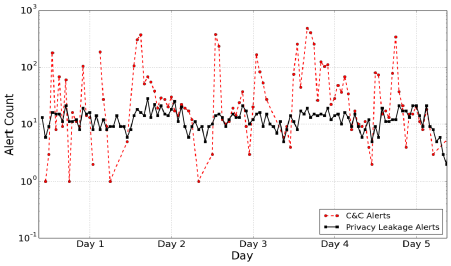
A total of 7849 out of 3.38 million uniquely identifiable devices were observed to trigger security alerts in our combined alert datasets. 73.2% of these events originated from a mobile device such as a smartphone or tablet while the rest are triggered by devices behind a wireless hotspot, and hence cannot be uniquely identified as being mobile or non-mobile. This puts the lower bound of the overall infection rate of mobile devices at 0.17% (5751 out of 3.29 million mobile devices), which is orders of magnitude higher than those reported in the most recent work by Lever et al [2]. Also, our observed infection rate is in agreement

with the reported rate in a recent study focusing on Android malware infection rates [3].

We rank the infected devices based on the total number of security alerts generated over the course of the week, and found that the top 20% of the devices account for more than 80% of the security alerts. Interestingly, the top 20% of the infected devices primarily consisted of Android and iOS based phones/tablets. Next, we investigate the nature of security threats that trigger these alerts.



(a) Threat Alerting Behavior



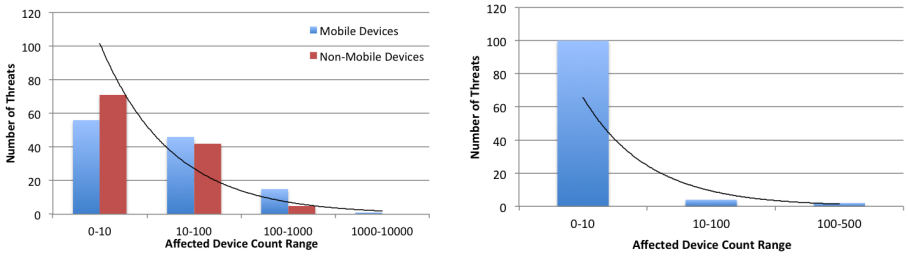
(b) Timeseries of Privacy Leakage and Botnet Communication Alerts

Figure 2: Macroscopic characterization of alert data

Based on the methodology described in Section 2, we extracted detailed information about the threat associated with each security event by leveraging

commercial virus-scanning tools, and through manual inspections. We found 327 unique threats in our *malicious traffic groundtruth* dataset that spans over the course of one week. After performing device classification, we further categorized these 327 threats into three classes with 75% confidence intervals as follows: (a) mobile-only threats that infect mobile devices (97 threats) (b) non-mobile threats that infect non-mobile devices (107 threats), and (c) cross-device threats that infect both types of devices (123 threats). Figure 2a characterizes the macroscopic alerting behavior of the three classes of threats in the network. The x-axis in this graph represents the top n% of threats in terms of the total number of alerts generated. In general, a small fraction of threats (5-15%) are responsible for a major proportion (over 80%) of the observed alert traffic. However, we note that mobile threats in general tend to generate less number of alerts than their non-mobile counterparts. This might indicate that attackers have adapted mobile malware to be stealthier and harder to detect on the network. Moreover, some mobile-specific threats (e.g., privacy leakage) generate less network footprints and hence trigger less number of alerts.

Exploring this further, we see that the number of alerts observed to be generated per threat is a function of the threat family (e.g. botnet, data leakage, etc.) and the number of devices affected by the threat. Privacy leakage threats such as threats responsible for leaking IMEI or location information from a device generally do not generate as many alerts as devices affected by a botnet threat (as shown in Figure 2b). A 'zombie' bot device makes regular call-backs to command and control servers for downloading instructions, data exfiltration and so on, hence generating a much larger footprint in the security alert logs. This implies that mining alert logs generated by network access activities could be effective in early detection and prevention of botnet-like threats. However, similar methodology will be ineffective for other threats, such as data leakage, that leave very little footprints. We incidentally observe that three botnet threats (1% of all observed threats) generated 49.3% of the observed alerts in the data.



(a) Mobile and Non-Mobile Threats

(b) Mobile Only Threats

Figure 3: Infection Effectiveness of Threats

Threat Class	Threat Description	Unique Threats	# Mobile	# Non-mobile & Unknown	# Associated IPs	Associated Ports
Trojans	Malware which utilizes techniques of social engineering, drive-by download & advanced rootkits to affect user devices	8	1669	470	159	53
Privacy leakage	Leakage of sensitive information such as IMEI number & user location	2	1277	418	77	8080, 80
Adware & PUA	HTTP Requests to known adware domains & Requests with known malicious UA strings	3	1179	368	45	80
SIP threats	Illegal session information modification & Replay attacks on SIP protocol	2	161	98	21	5060, 5061

Table 3: Top categories of prevalent mobile malware

3.1.1. Top Mobile Threats

Next, we perform detailed characterization of the top threats that infected the most number of *mobile* devices. Malware writers often aim to infect as many devices as possible in order to maximize their financial or other gains. Therefore we use the number of devices affected by a threat to quantify its success in the cellular network.

Figure 3 plots the infection effectiveness of two categories of threats: cross-device threats and mobile-only threats, respectively.The y-axis in these graphs plots the number of threats affecting a given range of devices in the x-axis. Notice here that only a few threats are able to successfully affect a large number of devices (either non-mobile or mobile). To better analyze the nature of these prominent threats, we further classify the top 15 threats (either mobile-only or cross-device threats) affecting the most number of mobile devices in the

Type of data	Affected Devices
IMEI number	757
Device Location	603
Phone number	14
Call Logs	5
SMS Logs	1

Table 4: Types of Privacy Leakage

network into four different classes based on unique characteristics exhibited by each threat as shown in Table 3. We now describe the characteristics of each of these malware categories and how they affect end users:

Trojan Threats: These programs deliberately cause harm to a user device while posing to be a benign application such as a free anti-virus solution. The harm can be either in terms of allowing unauthorized remote access to the device, hijacking device resources, turning the device into a bot/proxy, stealing user information etc. This class of malware is observed to be the most effective form of threat currently affecting mobile devices. Interestingly, through the course of our analysis, we detected instances of the Zeus trojan affecting 82 distinct iOS based mobile devices in the network. Although mobile variants of this threat affecting other platforms such as Windows Mobile and Android have been seen in the wild, to the best of our knowledge, this is the first time a variant of this threat was identified affecting iOS devices [14]. Unfortunately, we were not able to explore characteristics of this malware further due to limitations in the dataset.

Privacy Leakage Threats: Threats which maliciously leak the IMEI (International Mobile Equipment Identity) number and device location information affect over 1200 unique mobile devices, making this one of the most successful attacks targeting mobile devices. Although traditional desktop malware which leak sensitive user data exist, this problem is more pronounced in the mobile ecosystem. This may be due to the sensitive nature of data stored on user-

devices which attackers deem valuable, issues of application over-privilege in some mobile platforms, and the availability of third party app stores which makes deploying such malicious applications easy to do. Table 4 presents the breakdown of the various types of privacy leakage issues observable in our ground truth data. Clearly, information such as those presented above would potentially allow an attacker to uniquely observe a targeted user and his activities, making this a serious violation user privacy.

Adware & Potentially Unwanted Applications (PUA): This class of applications sneak into a device deceptively and get installed in such a way that it can be difficult to detect and remove. The primary motive of these programs is to display unwanted advertisements to users, often in the form of pop-up ads. While some of these apps may just be a minor irritant to the user, they may, in some cases, also act as dangerous spyware that monitor user behavior and collect data without consent.

SIP Threats: The Session Initiation Protocol (SIP) is widely used for controlling multimedia communication sessions such as VoIP calls over the internet. Our results indicate that vulnerabilities in this protocol is seen to be a popular target for attackers seeking to exploit mobile devices. These are alarming trends since such vulnerabilities can potentially give attackers the ability to listen-in on confidential voice communications or launch denial of service attacks as reported in previous studies [15, 16].

3.1.2. Infection rates of popular mobile platforms

The question of which mobile platforms are most vulnerable to security threats has been a hot topic of debate for several years. We attempt to answer this question by utilizing ground truth data obtained from the operational cellular network. Table [5] presents the following data points: *a*) The proportion of devices belonging to each identifiable mobile platform in our dataset. *b*) The proportion of devices of a given platform which are affected by threats or the infection rate and, *c*) The proportion of alerts observed in the ground truth originating from a given platform.

Device Platform	% Total devices	% Infection Rate	% Mobile Alerts
iOS	40.57%	0.22%	53.12%
Android	20.09%	0.39%	45.74%
Windows	0.2%	0.12%	0.76%
RIM OS	0.08%	0.32%	0.15%
Custom Feature Phone OS & Others	39.06%	0.0009%	0.21%

Table 5: Affected mobile platforms

We observe from the second column of the table that Android is the most vulnerable platform with a 0.39% infection rate (or 2631 out of 662,089 devices which are infected). This infection rate is slightly higher than those claimed by the most recent independent study of malware infection rates in Android by Truong et al [3] who measure it to be in the range of 0.26-0.28% and three times the rate reported by Google [17].

Android is followed closely by Blackberry with an infection rate of 0.32% (9 out of 2,739 devices) and iOS with 0.22% (3055 out of 1,336,853 devices). These figures show that the walled garden approach / security through obscurity as employed by these platforms are failing to ensure against malware spread. Blackberry devices are often used for business purposes due to their security capabilities. However, the nature of data stored on these devices may induce attackers specifically target this platform which can explain its high infection rate. Attackers are however failing to affect a large proportion of users with devices running Windows based mobile platforms as noted by recent industry reports [18].

4. Network Footprints of Cellular Threats

In this section, we investigate if network access patterns associated with malicious domains/hosts contacted by infected user devices exhibit distinct statistical features when compared to accesses to their benign counterparts. There are many existing studies that target accurate detection of malicious domains/URL's by using different methodologies. Some of these studies utilize a combination of DNS and WHOIS features, host based features, content of the webpage, etc in order to achieve their goals [4, 5] while some other studies such as [6] and [7] exclusively use lexical features. The motivation of our experiments with cellular network traffic however is to investigate if statistical network features can complement existing detection rules such as the lexical features. This methodology can be helpful in situations where other data such as WHOIS, webpage content etc. which are useful for the malicious domain classification task is infeasible to obtain or is otherwise unavailable.

4.1. Feature Extraction and Selection

In order to perform our classification experiment, we first build a set of known malicious domains using information from the ground truth alert database. We then create a set of benign domains by randomly choosing domains visited by subscriber devices which are otherwise not listed in the ground truth database. We further verify they are benign by running the domains through commercial URL scanners. For these set of known malicious and benign domains, we extract lexical and statistical network features as follows:

a) *Lexical features*: Each target hostname in our labeled benign/malicious domain set is broken down into multiple 'tags' or 'tokens' based on the '.' delimiter. We identify 6,729 such unique lexical tags through this process over a set of 1200 benign and malicious domains. We then utilize the frequency of occurrence of each tag in a given domain name as the lexical features of the target. This approach to represent lexical information is commonly referred to as the bag-of-words model. Variants of this model have been used to generate

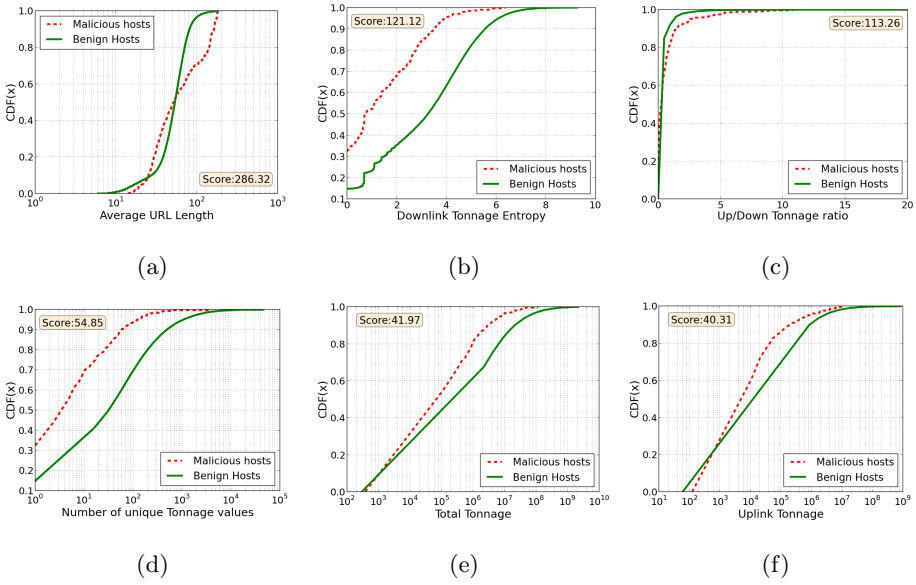


Figure 4: Network Footprint of Malicious Domains

lexical features for use in detecting malicious URL's in previous studies such as [7], [4].

b) *Statistical Network Features*: Using the DPI records from the cellular carrier we extract the following 12 heuristic features for each target domain: Uplink data transfer volume(or uplink tonnage in bytes), downlink data transfer volume (or downlink tonnage in bytes), ratio of uplink / downlink tonnage, total tonnage, proportion of failed connections, average URL length, number of connections, number of unique source IP's connecting to the domain, number of failed connections, entropy of destination IP addresses,downlink tonnage entropy and the number of unique tonnage values. We will discuss the utility of these features in more depth in Section 5.

We start our analysis by identifying specific network and lexical features that contribute towards distinguishing between malicious and benign hosts. In order to select such features, we utilize the raw set of attributes described above and apply the Chi-squared statistic evaluation [19]. The Chi-squared score essentially measures the difference between the conditional distributions of a network

Data set	ROC Area-α	ROC Area-β	ROC Area-γ
600 malicious 600 benign	0.843	0.744	0.9
540 malicious 600 benign	0.83	0.737	0.897
480 malicious 600 benign	0.838	0.732	0.895
420 malicious 600 benign	0.84	0.73	0.891
360 malicious 600 benign	0.852	0.746	0.897
300 malicious 300 benign	0.84	0.703	0.885
240 malicious 600 benign	0.813	0.703	0.885
180 malicious 600 benign	0.796	0.771	0.857
120 malicious 600 benign	0.824	0.76	0.869
60 malicious 600 benign	0.763	0.728	0.876

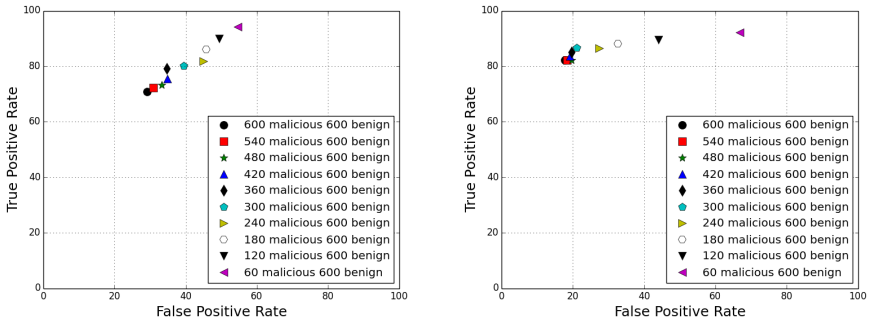
Table 6: Comparing ROC Areas

feature associated with the two classes: malicious vs. benign domains/hosts. On the basis of the results of this exercise, we narrow down our feature set to 53 distinct attributes associated with each malicious/benign domain after removing attributes which have a score of zero. This reduced feature set includes 10 statistical network features and 43 distinct lexical features. We note that we also experimented with other feature selection methods such as those based on information gain and subset based feature selection [20], but, obtained the best classification results with features selected using the Chi-squared statistic score.

Figure 4 shows the cumulative distribution function (CDF) of six selected network features associated with malicious and the benign hosts that exhibited the highest chi-squared scores. It is visually apparent that there is significant difference between the conditional distribution for malicious vs. benign domains/hosts for these network features. Other network features which were selected but not shown include the connection entropy, the destination IP entropy and the downlink tonnage.

4.2. Classification of Malicious/Benign Domains

Many of the statistical network features we have considered have complex non-linear relationships. This makes the task of classification of domains/hosts into malicious and benign categories non-trivial. To tackle this problem, we



(a) Only Lexical features

(b) Lexical & Network Features

Figure 5: Cross-Validation Results

use a machine learning approach which can handle such dependent features efficiently. In particular we use the "Random Forest" ensemble learner [21] to create a model with the individual features. This classification method operates by constructing multiple decision trees at a time (15 in our case) and predicts a class by aggregating the predictions of the ensemble. In addition, we use the n-fold cross validation technique to evaluate the accuracy of our model (setting n=10).

We run our classification experiments on varying proportions of malicious and non-malicious hosts employing *a*) Statistical network features alone (α), *b*) Lexical Features alone (β) and *c*) Statistical network features in addition to lexical features (γ). Figure 5a and Figure 5b present the receiver operating characteristic(ROC) for two of our cross-validation experiments. The ideal ROC would lie close to the upper-left corner with false positive rate close to 0% and true positive rate close to 100%. Note that with the addition of statistical network features to simple lexical features, we obtain a better true positive rate at lower false positive rates for most combinations of malicious and benign hosts. We observe from Table 6 that the ROC area is higher in the case where we utilize statistical network features along with lexical features (column 3) to perform classification as compared to using the lexical features alone (column 2) or statistical network features alone (column 1) for all proportions of malicious

and benign domains. These preliminary results show that statistical network features are indeed complementary to lexical features and hold promise to add to capabilities of existing detection rules to help solve the malicious domain detection problem.

Agg.-level	**Network traffic features**	**HTTP-header features**
Global	**Popularity Characteristics:** Number of Source IP's, Number of Connections **Client Characteristics:** Uplink tonnage, Total tonnage, Source IP entropy, User Entropy, Count of Unique uplink Tonnage Values. **Destination/Downlink Characteristics:** Destination IP entropy, Downlink Tonnage Entropy, Unique Downlink Tonnage Values, Downlink tonnage, Statistics S over Destination port	**HTTP Conversation Characteristics:** Statistics S over the Request URI, HTTP-Referer, HTTP-Accept, HTTP-Accept Encoding, HTTP-Accept Language, User-Agent and Cache-Control fields; Number of requests/responses with cache-control field set to no-cache or max-age=0. **HTTP Content Characteristics:** Statistics S over the Content-length and Content-Type fields; Number of requests/responses with content-types of 'application','image','multipart' and 'text'.
Device, User	Minimum, maximum and average over each network feature described above along with counts of number of unique accessing devices/users aggregated at the PDA/PUA level.	Minimum, maximum and average over each http-header feature described above aggregated at the PDA/PUA level.

Table 7: Summary of Network traffic and HTTP-header features

5. Screening Malicious Domains by Mining Network Traffic

In this section, inspired by the analysis above, we perform a deeper exploration of the utility of network traffic-based features in identifying malicious domains. As mentioned before, due to limited availability of the cellular network data, we utilize a large-scale dataset obstained from a campus WiFi network for this purpose. The goal of this exploration is to design an in-house traffic screening technique that can accurately identify potentially malicious domains. To this end, we classify a given FQDN (Fully Qualified Domain Name) as being either malicious or benign by mining of network traffic characteristics at three

aggregation levels: device-level, user-level, and network-wide. For each aggregation level, we explore possible information available through network traffic and HTTP headers, which will be detailed in the following subsections.

5.1. Supervised Learning and Classification

To detect whether a domain is malicious or not, we perform a supervised learning based classification using statistical features of network and HTTP-header information extracted from the secondary set of WiFi traffic traces. We build a classification model as before using the Random Forest ensemble learning algorithm [21], and apply SMOTE (Synthetic Minority Over-sampling TEchnique) [22] to deal with the class-imbalance issue. SMOTE allows us to learn with a combination of under-sampled instances from the majority class (i.e. benign domains in our case) and over-sampled instances from the minority class (i.e. malicious domains in our case) and helps deal with the classical problem of the model overfitting to the majority class instances.

5.1.1. Feature Exploration

We conjecture that malicious and benign domains show different characteristics due to differences in terms of (i) popularity (e.g. number of connections), (ii) access patterns of clients (e.g. uplink tonnage values or source IP entropy), and (iii) response patterns of domains (e.g. downlink tonnage.). We illustrate and explain the intuition behind choosing these characteristics in greater detail by considering representative features in Section 5.2.

In addition to the network features, we consider HTTP-header information that may capture the differences between malicious and benign domains in terms of (i) HTTP conversation characteristics between users and domains (e.g. average length of the request URI, user agent entropy, HTTP referer length etc.) and (ii) Content metadata, e.g. we can infer whether the payload is a binary executable file (which is often likely to be malicious) with the content-type field. Table 7 provides a summary view of these network and HTTP-header features.

Note that exploring the above features in a large-scale network trace whose size is over 2.4 TB is computationally expensive. To address this problem, we use the Spark[2] cluster computing framework [24] in four server machines that consist of a total of 88 CPU cores.

5.2. Feature Aggregation Dimensions

We suggest and discuss three different traffic aggregation levels over which the network and HTTP statistical features described above are extracted:

(i) Global-network aggregate (GNA): The GNA (also referred to as the network-wide aggregate) captures characteristics of the overall status of network traffic to and from a given domain. For instance, the count of the total number of connections to a given domain is an example of a GNA feature. For attributes that cannot be captured by a single numerical value such as those involving destination ports of domains, HTTP headers or Request URI Length values, we introduce a simple set of statistical measures S as our feature subset:

$S :=< min., max., avg., std.dev., var., entropy, median, 25th\%ile, 75th\%ile,$ $count\ of\ unique\ values >$

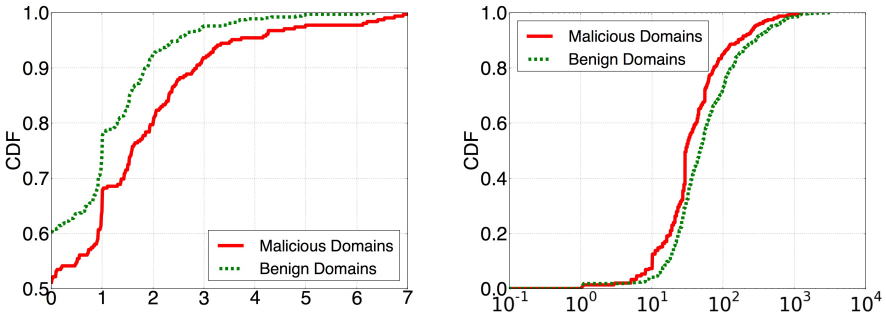
Note that if a field from the HTTP-header is not a numerical value, e.g. strings in the 'content-type' field in the HTTP-header, we use the length of the strings of the given field to calculate S.

As an representative example feature obtained via GNA, Figure 6a shows the CDF of *destination IP entropies* of malicious and benign domains. Entropy is often used to capture the degree of dispersal or concentration of a distribution. We compute the entropy with a process similar to one described in [25]. We build an empirical histogram $X = \{n_i, i = 1, ..., N\}$, where flows to destination IP i occurs n_i times. Then, the destination IP entropy of a given domain is defined as:

$$H(X) = -\sum_{i=1}^{N} \left(\frac{n_i}{S}\right) \log_2 \left(\frac{n_i}{S}\right) \quad (1)$$

[2]Spark's in-memory primitives are well suited for machine learning tasks and can produce up to 100x faster performance compared to Hadoop in certain applications [23].

Here $S = \sum_{i=1}^{N} n_i$ is the total number of observations in the histogram. We observe in Figure 6a that the malicious domains tend to have higher entropy values than the benign domains. This is consistent with the attacker behavior of associating multiple IP addresses with a given domain and fluxing between these addresses to avoid IP blacklisting attempts[26]. We note that many domains have only 1-2 IP's associated with them resulting in low entropy values ($\leq$1). Specifically, around 50% of malicious domains and 60% of benign domains only have a single IP address associated with them leading them to have a destination IP entropy of zero.



(a) Destination IP entropy (b) 25th%ile Request URI Length

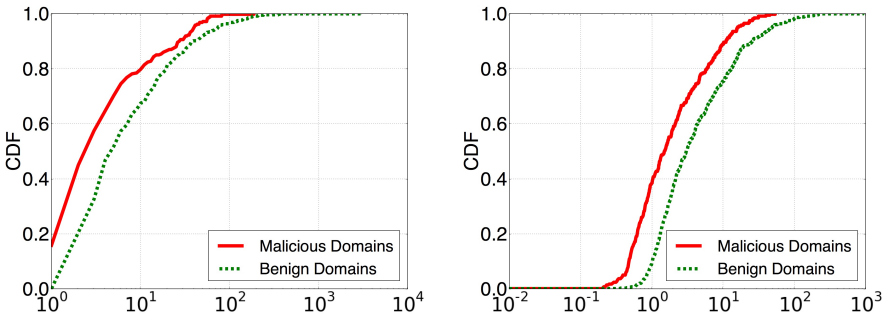
Figure 6: Global Network Aggregate (GNA) features

We also show the distribution of another example feature, *25th %ile of the request URI length*, obtained by GNA in Figure 6b. As shown in Figure 6b, the benign domains tend to show higher values of this attribute than the malicious domains. For example, the URI of a benign domain *play.google.com/store/-apps/details?id=air.com.insomniacgames.release.f3.omo* is sufficiently verbose and informative to the users to guess that the request is related to a game application. On the other hand, the URI of a malicious domain *www.962.net/wz/-63863.html* is much shorter than the benign URI in length, which can hide what the request leads to. Similar obfuscation effects and their security implications have been studied by recent work such as [27] in the light of URL shorten-

ing services. We note however that the dataset we consider does not have many shortened URL's. We also note that the length of the URI of a malicious domain can sometimes be abnormally long. We capture this characteristic in separate attribute which is the *Max. Request URI Length*.

(ii) Per-device aggregate (PDA): The PDA captures characteristics of network and HTTP features for each device that has accessed a given domain. For instance, if 10 distinct devices visit a domain 'bad.com', we calculate 10 sets of features, each of which represents the characteristic of traffic generated by each device. We then calculate meta-statistics (such as the minimum, maximum, and average) over the distributions of the 10 sets of features associated with the chosen domain 'bad.com'. These meta-statistics are then used as our PDA features for the domain considered. Note that this PDA abstraction can capture more fine-grained (device-level) characteristics of the given domain than the GNA that considers the overall status of network traffic for the domain. For clarity, let us consider a scenario where we find that there are 10,000 connections to 'bad.com' (This is an example of a popularity based GNA feature). It is difficult to *concretely* conclude from this single feature without the IP-device mapping if (i) One device in the network produces all the 10K connections, or (ii) If the requests to 'bad.com' are spread over many distinct devices. In the case where a single device is found to connect 10K times to a domain in a short time period is likely to be an anomalous occurrence, whereas observing the same number of connections from the entire network to a given domain can be normal in a large network. Thus, in this example scenario, we gain additional behavioral information by considering the device level aggregate statistics of the *Max. number of connections*.

We plot two representative features obtained via PDA in Figures 7a and 7b. Figure 7a presents the CDF of average number of connections for malicious and benign domains. We observe that the benign domains tend to have a larger *average* number of connections than the malicious domains, which indicates benign domains tend to leave much bigger footprints in the traffic compared to malicious domains. This is because there are more frequent interactions



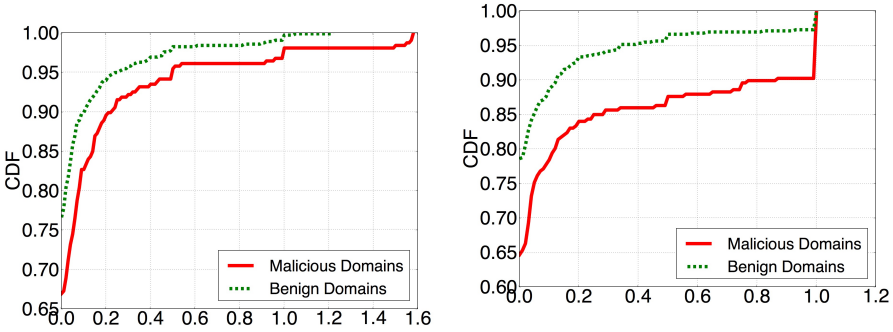
(a) Avg. number of connections (b) Avg. uplink tonnage (in KB)

Figure 7: Per-device Aggregate (PDA) features

between the benign domains and users (e.g. checking social network feeds). Figure 7b shows the CDF of average uplink tonnage for malicious and benign domains. We find that malicious domains tend to have lesser uplink traffic to them than benign domains, which is due to the fact that there are a lot more uplink interactions between benign domains and users, e.g. uploading video, sharing files etc.

(iii) Per-user aggregate (PUA): The PUA computes the feature based on user-level aggregation. In our dataset, we find that there exist many users who use more than one device (smartphone, laptop, tablet). Hence, aggregating along the devices of a user might reveal user-specific characteristics of accessing malicious domains. Figure 8a, as an example network feature, shows the CDF of the average source IP connection entropy values. We find that malicious traffic tends to show higher values for this network attribute in our trace. This can be explained by the observation that malware sometimes spoof source IP's before performing malicious actions (e.g. click fraud) to avoid detection or imitate legitimate user traffic. Next, we consider the CDF of a HTTP-header attribute: the average response count where the HTTP-cache control field set to 'no cache' or 'max-age=0' in Figure 8b. Attackers may want to force connecting clients to reload content from the malicious server and not use cached browser content by setting the HTTP-cache control field to 'no cache' or 'max-age=0'. Accordingly,

we observe that the values of this attribute tend to be slightly higher in the case of malicious domains compared to benign domains.



(a) Avg. source IP connection entropy

(b) Avg. no-cache or max-age=0 response count

Figure 8: Per-user aggregate (PUA) features

5.3. Feature Selection

We explore a large feature space of over 500 features in different aggregation dimensions (i.e. GNA, PDA, and PUA) and different information sources (i.e. network and HTTP-headers). However, considering all of these attributes would make the feature extraction process significantly computationally expensive, and in addition, many of these attributes may not contribute positively towards detection of malicious domains. To identify the features that contribute most effectively toward differentiating malicious and benign domains, we use the Chi-squared statistic evaluation [19]. We prune our initial feature space by excluding features with Chi-squared scores of zero, and use the resultant reduced feature set (<10% of all extracted features) to build our classifiers. Figure 9 presents the normalized Chi-squared statistic score of the top 5 traffic attributes from the aggregation dimensions of network, device and user. It is clear that device level aggregate features have relatively high statistic scores compared to the user and global aggregates. We explore the performance of these three aggregation levels in more details in Section 4.2.

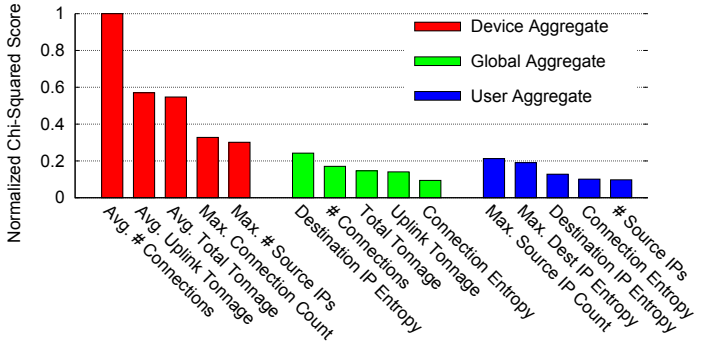


Figure 9: Comparing feature importance of the top network traffic features.

5.4. Classifier Evaluation

Abstraction	**# Features**	**Top χ^2 selected features**	**Malicious**	**Benign**	**ROC Area, Accuracy**
Network	7	Destination IP Entropy, Num. of Connections, Total tonnage, Uplink tonnage, Connection Entropy, Num. of unique tonnage values, Num. of Source IP's	F-score:0.87 TPR: 0.844 FPR: 0.096	F-score:0.88 TPR: 0.904 FPR: 0.156	**0.94**, 87.41%
http-header	61	Statistics S of Accept- Encoding, Accept-Language, Cache-control, Request URI, User Agent and http-referer fields.	F-score:0.82 TPR: 0.818 FPR:0.15	F-score:0.836 TPR: 0.85 FPR:0.182	0.898, 83.25%
Overall	68	Features listed above	F-score:0.85 TPR: 0.836 FPR: 0.13	F-score:0.856 TPR:0.87 FPR:0.163	0.926, 85.21%

Table 8: A comparison of three classifiers based on Network, HTTP-header, and Overall (Network + HTTP-header) features under the GNA aggregation.

In this section, we report the performance of our supervised classification in terms of the following four well-known metrics: (a) Accuracy, (b) True Positive Rate (TPR) and False Positive Rate (FPR) [28], (c) F-score / F-measure (F) [29], and (d) ROC area [30]. We evaluate our models by performing 10-fold cross validation.

Aggregation	**# Features**	**Top χ^2 selected features**	**Malicious**	**Benign**	**ROC Area, Accuracy**
Global	7	Network features listed in Table 3	F-score:0.87 TPR:0.844 FPR:0.096	F-score:0.878 TPR:0.904 FPR: 0.156	0.94, 87.41%
Device	13	Avg. and Max. values of number of Connections, Uplink and total tonnage, connection entropy, destination IP entropy, Number of source IP's and Avg Number of Unique Tonnage values	F-score:0.892 TPR:0.884 FPR:0.083	F-score:0.897 TPR:0.917 FPR: 0.12	0.959, 89.46%
User	7	Avg. and Max. values of Connection Entropy, Source IP count and Destination IP Entropy; Avg. of number of unique tonnage values	F-score:0.664 TPR:0.687 FPR:0.44	F-score:0.593 TPR:0.552 FPR:0.313	0.663, 63.19%
Global + Device	20	Features listed in Global and Device aggregates above.	F-score:0.898 TPR:0.885 FPR: 0.08	F-score:0.903 TPR: 0.92 FPR:0.115	**0.959**, **90.03 %**

Table 9: 10 fold cross-validation results using Network features of different aggregation granularity.

5.4.1. Network vs. HTTP-header

We build two separate machine learning classifiers for network and HTTP-header features. Table 8 compares the performance of the two classifiers along with a third classifier "Overall" which is built using both the network and HTTP features together. We report the results under the GNA aggregation as it can be easily calculated and obtained by most network operators (since IP to device and device to user mappings are not required for GNA). As shown in Table 8, the classifier based on network features outperforms the classifier based on HTTP-features (ROC area: 0.94 vs. 0.898). This can have important implications for network operators as network traffic information is often more easily obtainable than HTTP information which requires specialized DPI boxes for data collection. We also observe that some of the HTTP-header features

introduce noise to the detection, which results in lower performance in terms of the ROC area when we utilize the HTTP-header features along with the network features.

5.4.2. Impact of Aggregation Levels

Next, we investigate if finer aggregation granularity of network traffic information (i.e. PUA or PDA aggregations) could improve the accuracy of detecting malicious domains. Table 9 shows the performance of classifiers based on network features with different aggregation dimensions – global, device, and user levels. We find that the device-level aggregation performs slightly better (accuracy: 89.46%) than global-level aggregation (accuracy: 86.41%). These results imply that both device-level and global-level aggregate statistics are effective in detection of malicious domains, but more fine-grained device-level information can better capture the network traffic characteristics compared to the global aggregate. There is a cost-benefit trade-off involved in choosing features from different aggregation levels: the device-level aggregation requires more computations than the global-level aggregation, but the former achieves higher accuracy than the latter. Network operators can choose the aggregation dimension according to their detection goals and available mapping information. Note that combining global and device-level aggregation features further improves the performance of the classifier to 90.03%. We also observe from Table 9 that the performance of user-level aggregation is worse than the other aggregation dimensions. This is because users in our network traces utilize multiple devices (smartphones, laptops, tablets) belonging to different platforms (e.g. Android, iOS, Windows). As shown in [31], these platforms have different infection rates and hence we observe that aggregating along devices with different infection rates does not strongly contribute towards detecting malicious domains.

5.4.3. Analyzing Misclassifications

We now analyze the domains that are misclassified by our best performing classifier built using network features under global and device-level aggregation

(i.e. the last row of Table 9). A domain is identified as a false positive if our classifier reports the domain to be malicious, but we know it is in-fact benign by our ground truth database. We investigate the characteristics of such false positive domains by issuing active queries to them, and examining their HTTP response codes and response content. We observe that 33.33% of these false positive domains return 404 (Not found) or 403 (Forbidden) responses, 25% of them are redirect domains, 22.91% of them either return empty pages or failed to respond, and only 18.75% of them appear to be benign landing pages. These results suggest that a lightweight pre-processing step for checking domain response codes can help to reduce the false positive rates.

Furthermore, we find that 33.3% of the phishing domains were reported as false negatives (i.e. malicious domains reported as benign). We believe the high false negative rate of the phishing domains is due to the fact that phishing domains masquerade as benign domains and mimic their behavior. This makes it difficult to distinguish the network features of malicious phishing pages and benign domains, resulting in higher misclassification rate. We however note that phishing domains only constitute 5% of all malicious domains in our data. We leave a more detailed analysis of this issue to our future work.

6. Related Work

There have been significant efforts in recent years towards understanding the characteristics of mobile malware. One class of research in such efforts has focused on characterizing the nature of mobile applications available in official and third party mobile application markets. Zhou et al [32] characterized android malware and performed an evolution-based study on representative threats. In [33], the authors examined a broad range of security concerns that can affect applications on the Android platform. Egele et al [34] studied privacy threats affecting iOS users whereas TaintDroid [35] used a system wide taint tracking to identify privacy leaks in Android applications. Some studies have surveyed the types of mobile malware seen in the wild, and evaluated different techniques

that can be leveraged in detecting and preventing such threats [36, 37, 38].

Although these studies have provided good insights on the 'inner workings' of mobile malwares, the vulnerabilities affecting smartphone platforms, and the security issues that application markets have to deal with, little has been known about (i) the details of the prevalence of such malwares in cellular network traffic and (ii) the trends of the evolving landscape of threats actively infecting users in an operational network. In particular, questions about (i) what kinds of threats affect today's cellular networks, (ii) what are the traffic characteristics of such threats, and (iii) how users are being affected by such threats, have not been thoroughly explored in detail.

Recent commercial reports and press articles covering mobile threats have reported different infection rates of mobile devices. For instance, Alcatel-Lucent's recent report [39] showed that >0.5% of mobile devices are infected by malware whereas McAfee [40] reported a much higher infection rate (>5%). Independent academic studies are therefore important to clarify these measurements to comprehensively understand the current threat landscape. Lever et al [2] took a first step to address such issues by performing a network level analysis on mobile malware using DNS traffic data from a major US based cellular carrier. However, their analysis is limited only to HTTP/HTTPS based threats which issue DNS requests prior to communicating with malicious URL's. Also, they did not quantify the prevalence of specific types of threats affecting the network and track their temporal growth behavior. Truong et al [3] performed a measurement on the infection rate in the Android platform (as of 2013). They estimated that between 0.26% to 0.28% of Android devices are infected, by means of direct measurements from 55,000 devices. They also pointed out concerns about disparities in commercial vendor reports about mobile malware in their study. They did not, however, perform a detailed characterization of the most prominent families of threats affecting Android or compare infection rates of Android against other popular mobile platforms such as iOS, BlackBerry, and Windows.

Another set of research efforts relevant to our work has dealt with the prob-

lem of 'detecting' malicious hosts/URLs. Some studies such as [4, 5] cast this as a supervised learning problem where a classifier learns on a combination of DNS, WHOIS, lexical, and other features associated with a given host to decide whether it is malicious or benign. Other studies such as [6, 7] exclusively utilized lexical features to achieve similar goals. Angiulli et al [41] studied an unsupervised intrusion detection technique by mining HTTP information using the n-gram model. Lakhina et al in [25] and [42] focused on the problem of diagnosing network anomalies (eg. port scans, volume anomalies etc.) based on network measurements. However, to the best of our knowledge, this is the first study that comprehensively examines network-level features of malicious domains associated with the threats in wireless data networks, and the application of such statistical network features to the malicious domain identification problem, in these networks.

7. Conclusions and Future Work

In this paper, we presented a study of malicious mobile traffic by using data obtained from a major US based cellular carrier and a large campus WiFi network. Our investigation revealed that 0.17% of mobile devices connecting to the cellular carrier are affected by security threats. This infection rate while still small, is much higher than the last reported infection rate of 0.0009% making this a worrisome problem. We combined multiple disparate data sets to uncover details about the threats affecting mobile devices in the cellular network and their unique characteristics. We also performed a detailed analysis of infection rates in various popular mobile platforms. Our results showed that platforms deemed to be more secure by common opinion as iOS and BlackBerry are not as secure as we think. However, Android still remains the most affected platform with an infection rate of 0.39%. We characterized the aggregate network footprint of malicious and benign domains associated with the threats observed in the cellular network dataset, and showed that network features are complementary to lexical features.

Using the over 2.4 TB of WiFi traffic from a university campus network, we next presented a novel, in-house traffic screening technique that can identify malicious domains based solely on statistical properties of network traffic. We demonstrated that our malicious domain classier can achieve over 90% accuracy and 0.959 ROC area.

Our ongoing work includes the exploration of practical system design issues (E.g. finding the optimal training time for the classifier), investigation of fine-grained characteristics of different threat categories (e.g. domains generated by Domain Generation Algorithms (DGAs) [43], phishing etc.) in terms of network traffic patterns, and exploration of different possible ways of evasion by an adversary. Another important direction is to develop a method for detecting malicious domains when they use encrypted channels for communications (i.e. HTTPS) using network traffic features. Methods which use HTTP header data for detection are not applicable when the captured HTTP conversation is encrypted, but if a distinguishing network-level signature for encrypted malicious domains exists, it may be useful for identifying such sophisticated malicious domains.

Acknowledgement

We would like to thank Emiliano Martinez from Virustotal, Dan Nunes from Intel-security and Pankaj Kumar from Guavus Inc. for their help with obtaining the security data feeds. This work was supported in part by the Intel Science and Technology Center for Secure Computing.

References

[1] X. Wei, L. Gomez, I. Neamtiu, and M. Faloutsos, "Malicious android applications in the enterprise: What do they do and how do we fix it?," in *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on*, pp. 251–254, IEEE, 2012.

[2] C. Lever, M. Antonakakis, B. Reaves, P. Traynor, and W. Lee, "The core of the matter: analyzing malicious traffic in cellular carriers," in *Proc. NDSS*, vol. 13, pp. 1–16, 2013.

[3] H. T. T. Truong, E. Lagerspetz, P. Nurmi, A. J. Oliner, S. Tarkoma, N. Asokan, and S. Bhattacharya, "The company you keep: mobile malware infection rates and inexpensive risk indicators," in *Proc. 23rd international conference on World Wide Web*, pp. 39–50, 2014.

[4] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious urls: an application of large-scale online learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 681–688, ACM, 2009.

[5] H. Choi, B. B. Zhu, and H. Lee, "Detecting malicious web links and identifying their attack types," in *Proc. 2nd USENIX conference on Web application development*, p. 11, 2011.

[6] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature based phishing URL detection using online learning," in *Proc. 3rd ACM workshop on Artificial intelligence and security*, pp. 54–60, 2010.

[7] A. Le, A. Markopoulou, and M. Faloutsos, "Phishdef: URL names say it all," in *Proc. IEEE INFOCOM*, pp. 191–195, 2011.

[8] L. Invernizzi, S.-J. Lee, S. Miskovic, M. Mellia, R. Torres, C. Kruegel, S. Saha, and G. Vigna, "Nazca: Detecting malware distribution in large-scale networks," 2014.

[9] J. Fan, J. Xu, M. H. Ammar, and S. B. Moon, "Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme," *Computer Networks*, vol. 46, no. 2, pp. 253–272, 2004.

[10] "Virustotal scanner: https://www.virustotal.com/en/about/,"

[11] "Zeroaccess trojan communication http://www.speedguide.net/-port.php?port=22292,"

[12] "Backdoor.remocy trojan communication: www.speedguide.net/port.php-?port=7776,"

[13] "Uses of port 8080: www.speedguide.net/port.php?port=8080,"

[14] D. Maslennikov, "Zeus in the mobile - facts and theories." www.securelist.com/en/analysis/204792194, 2011.

[15] S. El Sawda and P. Urien, "SIP security attacks and solutions: A state-of-the-art review," in *Information and Communication Technologies, 2006. ICTTA'06. 2nd*, vol. 2, pp. 3187–3191, IEEE, 2006.

[16] D. Geneiatakis, T. Dagiuklas, G. Kambourakis, C. Lambrinoudakis, S. Gritzalis, S. Ehlert, and D. Sisalem, "Survey of security vulnerabilities in session initiation protocol," *IEEE Communications Surveys and Tutorials*, vol. 8, no. 1-4, pp. 68–81, 2006.

[17] S. M. Patterson, "Contrary to what you've heard, Android is almost impenetrable to malware." qz.com/131436/contrary-to-what-youve-heard-android-is-almost-impenetrable-to-malware, 2013.

[18] "Cisco 2014 annual security report." www.cisco.com/web/offers/lp/2014-annual-security-report/index.html.

[19] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, pp. 388–388, IEEE Computer Society, 1995.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, no. 1, pp. 321–357, 2002.

[23] "The apache spark project: https://spark.apache.org/,"

[24] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets,"

[25] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *Proceedings of ACM SIGCOMM 2005*, pp. 217–228, Aug. 2005.

[26] G. Ollmann, "Botnet communication topologies," *Retrieved September*, vol. 30, p. 2009, 2009.

[27] N. Nikiforakis, F. Maggi, G. Stringhini, M. Z. Rafique, W. Joosen, C. Kruegel, F. Piessens, G. Vigna, and S. Zanero, "Stranger danger: exploring the ecosystem of ad-based url shortening services," in *Proceedings of the 23rd international conference on World wide web*, pp. 51–62, International World Wide Web Conferences Steering Committee, 2014.

[28] "Useful statistical definitions: http://www.cs.rpi.edu/ leen/misc-publications/somestatdefs.html,"

[29] Y. Sasaki, "The truth of the f-measure," *Teach Tutor mater*, pp. 1–5, 2007.

[30] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[31] A. Raghuramu, H. Zang, and C.-N. Chuah, "Uncovering the footprints of malicious traffic in cellular data networks," in *Passive and Active Measurement*, pp. 70–82, Springer, 2015.

[32] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *Security and Privacy (SP), 2012 IEEE Symposium on*, pp. 95–109, IEEE, 2012.

[33] W. Enck, D. Octeau, P. McDaniel, and S. Chaudhuri, "A study of android application security.," in *USENIX security symposium*, 2011.

[34] M. Egele, C. Kruegel, E. Kirda, and G. Vigna, "Pios: Detecting privacy leaks in ios applications.,"

[35] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. Sheth, "Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones.," in *OSDI*, vol. 10, pp. 255–270, 2010.

[36] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner, "A survey of mobile malware in the wild," in *Proc. 1st ACM workshop on Security and privacy in smartphones and mobile devices*, pp. 3–14, ACM, 2011.

[37] Y. Zhou, Z. Wang, W. Zhou, and X. Jiang, "Hey, you, get off of my market: Detecting malicious apps in official and alternative android markets," in *Proceedings of the 19th Annual Network and Distributed System Security Symposium*, 2012.

[38] M. Grace, Y. Zhou, Q. Zhang, S. Zou, and X. Jiang, "Riskranker: scalable and accurate zero-day android malware detection," in *Proc. 10th international conference on Mobile systems, applications, and services*, pp. 281–294, ACM, 2012.

[39] "Alcatel-lucent motive security labs malware report 2014: https://www.alcatel-lucent.com/press/2015/alcatel-lucent-report-malware-2014-sees-rise-device-and-network-attacks-place-personal-and-workplace,"

[40] "Mcafee labs threats report 2015: http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q4-2014.pdf,"

[41] F. Angiulli, L. Argento, and A. Furfaro, "Pckad: an unsupervised intrusion detection technique exploiting within payload n-gram location distribution," *arXiv preprint arXiv:1412.3664*, 2014.

[42] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, vol. 34, pp. 219–230, ACM, 2004.

[43] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou II, S. Abu-Nimeh, W. Lee, and D. Dagon, "From throw-away traffic to bots: Detecting the rise of dga-based malware.," in *USENIX security symposium*, pp. 491–506, 2012.