Computer Communications 000 (2016) 1–13

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Uncovering the footprints of malicious traffic in wireless/mobile networks

Arun Raghuramu[a,*], Parth H. Pathak[a], Hui Zang[b], Jinyoung Han[c], Chang Liu[c], Chen-Nee Chuah[c]

[a] Department of Computer Science, University of California, Davis, United States
[b] Huawei, Santa Clara, CA, USA
[c] Department of Electrical & Computer Engineering, University of California, Davis, United States

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Security
Mobile traffic analysis
Data mining

ABSTRACT

This paper presents a measurement study that analyzes large-scale traffic data gathered from two different wireless scenarios: cellular and Wi-Fi 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 Wi-Fi 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 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

[*] Corresponding author. Tel.: +15306013589.

E-mail addresses: araghuramu@ucdavis.edu (A. Raghuramu), phpathak@ucdavis.edu (P.H. Pathak), huizang@gmail.com (H. Zang), rghan@ucdavis.edu (J. Han), cchliu@ucdavis.edu (C. Liu), chuah@ucdavis.edu (C.-N. Chuah).

http://dx.doi.org/10.1016/j.comcom.2016.04.011
0140-3664/© 2016 Elsevier B.V. All rights reserved.

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 Wi-Fi 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 Wi-Fi 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. Section 4 and 5 investigate how to detect the malicious traffic by exploring their nature of network footprints in a cellular and Wi-Fi 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 Wi-Fi 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 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. Wi-Fi network data

As noted earlier, in addition to the cellular traffic data, we collect network traces from Wi-Fi controllers that connect and control the Wi-Fi 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 Wi-Fi users over a three days period from nearly 1000 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 Wi-Fi session start and end times.

[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.