COMPUTERS & SECURITY 70 (2017) 335–350

Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

A novel privacy preserving user identification approach for network traffic

N. Clarke [a,b], F. Li [a,*], S. Furnell [a,b,c]

[a] Centre for Security, Communications and Network Research, University of Plymouth, Plymouth, United Kingdom
[b] Security Research Institute, Edith Cowan University, WA, Australia
[c] Centre for Research in Information and Cyber Security, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

ARTICLE INFO

Article history:
Received 23 January 2017
Received in revised form 30 May 2017
Accepted 26 June 2017
Available online 10 July 2017

Keywords:
Biometrics
Digital forensics
Network forensics
Network metadata
Traffic analysis
User identification

ABSTRACT

The prevalence of the Internet and cloud-based applications, alongside the technological evolution of smartphones, tablets and smartwatches, has resulted in users relying upon network connectivity more than ever before. This results in an increasingly voluminous footprint with respect to the network traffic that is created as a consequence. For network forensic examiners, this traffic represents a vital source of independent evidence in an environment where anti-forensics is increasingly challenging the validity of computer-based forensics. Performing network forensics today largely focuses upon an analysis based upon the Internet Protocol (IP) address – as this is the only characteristic available. More typically, however, investigators are not actually interested in the IP address but rather the associated user (whose account might have been compromised). However, given the range of devices (e.g., laptop, mobile, and tablet) that a user might be using and the widespread use of DHCP, IP is not a reliable and consistent means of understanding the traffic from a user. This paper presents a novel approach to the identification of users from network traffic using only the meta-data of the traffic (i.e. rather than payload) and the creation of application-level user interactions, which are proven to provide a far richer discriminatory feature set to enable more reliable identity verification. A study involving data collected from 46 users over a two-month period generated over 112 GBs of meta-data traffic was undertaken to examine the novel user-interaction based feature extraction algorithm. On an individual application basis, the approach can achieve recognition rates of 90%, with some users experiencing recognition performance of 100%. The consequence of this recognition is an enormous reduction in the volume of traffic an investigator has to analyse, allowing them to focus upon a particular suspect or enabling them to disregard traffic and focus upon what is left.

1. Introduction

During the past 15 years, Internet usage has experienced explosive growth and technological evolution – from a simple data network with around 500 million users to a multipurpose and multiservice platform with almost 3.2 billion users (Internetlivestats, 2015). Indeed, with the prevalence of various broadband network technologies, mobile devices, and the web, users can utilize a wide range of services to complete

* Corresponding author.
E-mail address: fudong.li@plymouth.ac.uk (F. Li).
http://dx.doi.org/10.1016/j.cose.2017.06.012

various personal and business tasks both in the office and on the move 24/7. Services include (but are not limited to) entertainment (e.g., watching online videos), communication (e.g., making VoIP calls), finance (e.g., online banking), data storage (e.g., cloud services), office applications (e.g., Google Docs), and even Operating Systems (e.g., ZeroPC). It is evidenced that these activities generate a tremendous amount of IP traffic – 60 Exabyte's globally per month in 2014, 40% of which originated from non-PC devices (Cisco, 2015).

While people and business take the full advantage of the Internet, malicious attackers use the same infrastructure to plot various cyberattacks (e.g., hacking, Denial of Service (DoS), insider misuse, and phishing) against user's information, corporation's servers, and Internet services. Indeed, there is an ever-increasing volume of literature reporting the scale and nature of the computer-related crime (both cyber and computer-assisted). For example, the FBI Internet Crime Complaint Center reported 269,422 self-reported incidents of cyberattacks (mainly fraud related) in 2014, with a total estimated loss of $800 million (FBI, 2015). The Verizon's 2015 Data Breach Investigations Report shows that almost 80,000 security incidents were discovered by 70 organizations around world in 2014, causing them an estimated financial loss of $400 million (Verizon, 2015).

With the aim to search evidence and testify against cybercriminals, images of suspects' digital devices are forensically created and examined by investigators. For instance, statistics from the FBI's Regional Computer Forensics Laboratory show that they had completed 6322 examinations and investigators testified in court and/or at hearing 88 times in 2014 (FBI, 2014). Due to the volatile nature of computer memory and the availability of anti-forensic techniques (e.g., data wiping), critical evidence can vanish when the computer is switched off or purposely destroyed by the suspects as they have direct access to their digital devices. As a result, a complete picture of how an attack is conducted might not be achievable. As such, independent sources of data, such as network traffic, provide investigators with an invaluable source of evidence, where the chance of the evidence being tampered or destroyed is minimized.

Many tools (both commercial and open source) have been designed and developed to assist network forensic examiners to conduct investigations. These tools include NIKSUN's NetDetector Suite (NIKSUN, 2016), RSA's Netwitness Suite (RSA, 2016), Wireshark (Wireshark, 2016), PyFlag (Cohen, 2008), and Xplico's Open Source Network Forensic Analysis Tool (Xplico, 2016). They all rely upon the IP address of the suspect's machine as a basis for the investigation, assuming IP is static and linkable to an individual. However, IPs are increasingly unreliable due to the mobile nature of devices and the use of dynamic allocation of IP addresses. As a result, beyond the detected attack (often a single IP packet or flow), it is a challenge to investigate the question of *what* has happened in terms of the wider attack and *who* was actually involved. Indeed, to identify and extract a specific user's traffic from the wider organization over a prolonged period is a particularly challenging task.

Biometrics is a proven method that identifies individuals based upon their physiological or behavioural traits. Several biometric techniques, such as face recognition, fingerprint identification, and speaker recognition, are already in wide use within the forensic domain (Vezzani et al., 2013). However, little research has been undertaken on identifying individuals using biometric design techniques within the network forensic domain. Existing research has largely focused upon merely providing network data (either in packet or flow forms) to identify anomalous behaviour (a two-class problem of benign and malicious traffic) rather than looking to identify particular individuals (which is an n class problem, where n is the size of the user population). Studies into behavioural profiling on desktop and mobile platforms have demonstrated the ability to verify an individual; however, deriving application-level interactions (such as which websites users visit and more importantly what they do whilst visiting – posting, chatting, listening to music or watching video) from low-level encrypted packet-based data has proven challenging. Furthermore, using these application-based interactions for identification rather than verification introduces a need for stronger discriminative information. To this end, this paper describes an experimental study which proposes and investigates user identification through user's application-level activities based solely on the metadata of network packets.

The remainder of the paper is structured as follows. Section 2 reviews existing network traffic analysis methods and the prior art in behavioural profiling. Section 3 presents a novel feature extraction approach to deriving user oriented application level activities. The research methodology and the formation of the user activity dataset are presented in Section 4, followed by a full experimental study to evaluate the approach in Section 5. Section 6 discusses the proposed approach and its impact, while the conclusions and future work are highlighted in Section 7.

2. Prior art in network and behavioural profiling

In order to fully understand the relationship between user's application level interactions and their corresponding network signals, a detailed review on existing network traffic analysis is discussed. The work into network traffic analysis can be traced back to 1990s (Claffy et al., 1995; Debar et al., 1999) and is utilized by network administrators in various domains, including management, prioritization, performance, accounting, application behaviour analysis, and security (Hofstede et al., 2014). Depending upon the granularity of the analysis, network traffic can be analysed by two approaches: packet based (finer grained) or flow based (coarser grained). The packet based method is mainly used to examine the content (i.e., the payload) of individual IP packets, while flow based approach is utilized to analyse the summary of multiple IP packets that share similar characteristics over a period of time (i.e., IP flows). Details of these two methods, including their working principles, existing research, and advantages and disadvantages, are described. In addition, whilst behavioural profiling has not been applied to network traffic, an analysis of the prior art is presented to provide a baseline understanding of the technique and the typical levels of performance that can be expected.