You Are How You Query: Deriving Behavioral Fingerprints from DNS Traffic

Dae Wook Kim(✉) and Junjie Zhang

Wright State University, Dayton, USA
{kim.107,junjie.zhang}@wright.edu

Abstract. As the Domain Name System (DNS) plays an indispensable role in a large number of network applications including those used for malicious purposes, collecting and sharing DNS traffic from real networks are highly desired for a variety of purposes such as measurements and system evaluation. However, information leakage through the collected network traffic raises significant privacy concerns and DNS traffic is not an exception. In this paper, we study a new privacy risk introduced by passively collected DNS traffic. We intend to derive *behavioral fingerprints* from DNS traces, where each behavioral fingerprint targets at uniquely identifying its corresponding user and being immune to the change of time. We have proposed a set of new patterns, which collectively form behavioral fingerprints by characterizing a user's DNS activities through three different perspectives including the domain name, the inter-domain relationship, and domains' temporal behavior. We have also built a distributed system, namely *DNSMiner*, to automatically derive DNS-based behavioral fingerprints from a massive amount of DNS traces. We have performed extensive evaluation based on a large volume of DNS queries collected from a large campus network across two weeks. The evaluation results have demonstrated that a significant percentage of network users with persistent DNS activities are likely to have DNS behavioral fingerprints.

Keywords: Domain Name System · Behavioral fingerprints · Privacy

1 Introduction

The Domain Name System (DNS) plays an indispensable role in the Internet by providing fundamental two-way mapping between domains and Internet Protocol (IP) addresses. Its practical usage has gone far beyond the domain-IP mapping service: it supports many critical network services such as traffic balancing [1] and content delivering [2]; it is also leveraged by attackers to build agile and robust malicious cyber infrastructures, where salient examples include fast-flux [3], random domain generator [4], and covert channels [5]. The importance and prevalence of DNS signifies the demand of its traces collected from real networks, which are essential for many DNS-relevant designs by serving as benchmark data or ground truth. For instance, DNS traces have been collected

B. Thuraisingham et al. (Eds.): SecureComm 2015, LNICST 164, pp. 348–366, 2015.
DOI: 10.1007/978-3-319-28865-9_19

to evaluate DNS cache algorithms [6] and to train statistical models for malicious domain detection [7,8]. Although the specific type and granularity of information extracted from DNS traces may vary for different applications, the demand for DNS traces is generally increasing.

Despite their practical values, DNS traces may introduce significant privacy concerns. For example, DNS queries that are triggered by the prefetching mechanisms of popular browsers can leak users' search engine queries [9]; DNS queries can also reveal the types of operating systems [10]. In this project, we study a new privacy risk introduced by passively collected DNS traffic: to which extent network users can be uniquely identified merely based on the way they issue DNS queries? In other words, we intend to derive *behavioral fingerprints* from DNS traces, where each behavioral fingerprint targets at uniquely identifying its corresponding user and being immune to the change of time. Such DNS-based behavioral fingerprints, once successfully derived, have strong privacy implications. For example, they can be used to de-anonymize the DNS traces with anonymized sources. To be more specific, when DNS traces are shared, the source (e.g., the IP address) that issues the DNS query is usually anonymized (e.g., by obscuring the IP address using hash functions). However, one can learn behavioral fingerprints from un-anonymized DNS traces and use the acquired fingerprints to reveal the presence of specific users in (other) anonymized traces. In addition, if one can get access to DNS traces collected from multiple access networks (e.g., through open DNS services or collecting traces from multiple networks), he/she can track users' locations across different networks by using behavioral fingerprints to reveal users in DNS traces.

This paper aims at investigating the extent to which behavioral fingerprints can be derived and measuring their accuracy on identifying the presence of corresponding network users. As a means towards this end, we have proposed a set of new patterns, which collectively form behavioral fingerprints. We also built a distributed, scalable system, namely *DNSMiner*, to automatically derive DNS-based behavioral fingerprints from a massive amount of DNS traces. Specifically, we make the following contributions in this paper.

- We have designed five new patterns including *domain set*, *domain sequence*, *window-aware domain sequence*, *period behavior*, and *hourly behavior*, which collectively form behavioral fingerprints. These patterns systematically characterize DNS behaviors from three aspects including the domain name, the inter-domain relationship, and the temporal behavior. Although more patterns might be discovered to enhance behavioral fingerprints, our proposed patterns serve as a lower bound of the capabilities to use DNS behaviors to fingerprint network users.
- We have built a system, namely *DNSMiner*, to automatically mine behavioral fingerprints from a massive amount of DNS traces. The design of the system leverages the MapReduce distributed infrastructure to scale up the system performance. After being deployed in a 15-nodes Hadoop platform, *DNSMiner* can process more than 467 million DNS queries using approximately 4 hours.