Journal of Network and Computer Applications 47 (2015) 72–84

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Accurate DNS query characteristics estimation via active probing

Xiaobo Ma [a], Junjie Zhang [b], Zhenhua Li [c], Jianfeng Li [a], Jing Tao [a,*], Xiaohong Guan [a,c], John C.S. Lui [d], Don Towsley [e]

[a] MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an, China
[b] Dept. Computer Science and Engineering, Wright State University, Dayton, OH, US
[c] Dept. Automation, School of Software and TNLIST, Tsinghua University, Beijing, China
[d] Dept. Computer Science and Engineering, The Chinese University of Hong Kong, China
[e] Dept. Computer Science, University of Massachusetts, Amherst, MA, US

ARTICLE INFO

Article history:
Received 27 November 2013
Received in revised form
12 July 2014
Accepted 11 September 2014
Available online 12 October 2014

Keywords:
DNS
DNS query characteristics
Active probing

ABSTRACT

As the hidden backbone of today's Internet, the Domain Name System (DNS) provides name resolution service for almost every networked application. To exploit the rich DNS query information for traffic engineering or user behavior analysis, both *passive capturing* and *active probing* techniques have been proposed in recent years. Despite its full visibility of DNS behaviors, the *passive capturing* technique suffers from prohibitive management cost and results in tremendous privacy concerns towards its large-scale and collaborative deployment. Comparatively, the *active probing* technique overcomes these limitations, providing broad-view and privacy-preserving DNS query analysis at the cost of constrained visibility of fine-grained DNS behavior. This paper aims to accurately estimate DNS query characteristics based on DNS cache activities, which can be acquired via active probing on a large scale at negligible management cost and minimized privacy concerns. Specifically, we have made three contributions: (1) we propose a novel solution, which integrates the renewal theory-based DNS caching formulation and the hyper-exponential distribution model. The solution offers great flexibility to model various domains; (2) we perform a large-scale real-world DNS trace measurement, and demonstrate that our solution significantly improves the estimation accuracy; (3) we apply our solution to estimate the malware-infected host population in remote management networks. The experimental results have demonstrated that our solution can achieve high estimation accuracy and outperforms the existing method.

1. Introduction

The Domain Name System (DNS) provides two-way mapping between domains meaningful to humans and IP addresses associated with networking services. It has become an indispensable component for the Internet since the vast majority of network applications rely on DNS to establish connections with Internet services. The salient examples include web servers, data centers, content delivery networks (CDNs) (Adhikari et al., 2012), and cloud computing (Bernstein et al., 2009). In addition, DNS plays an increasingly critical role in improving the robustness and agility of malicious services such as botnet command and control (C&C) servers (Conficker, 2014; Shin et al., 2009), phishing websites (Zhang et al., 2011), and spamming campaigns (Egele et al., 2013). Therefore, it becomes a natural way to study network behaviors of a variety of network applications by monitoring and investigating DNS traffic. For instance, the number of hosts that query a botnet C&C domain reveals the bot population in the monitored network (Abu et al., 2006).

As depicted in Fig. 1, DNS is mainly composed of two types of components including *authoritative* servers (a.k.a., *A-DNS* servers) and *recursive* servers (a.k.a., *R-DNS* servers) (Rfc1034, 2014). Since R-DNS servers directly interact with applications in end users' hosts, the DNS activities observed between end users' hosts and their corresponding R-DNS server characterize the DNS-relevant activities at the finest granularity. Hence, R-DNS servers represent a perfect vantage point for network monitoring. In fact, a number of methods (Bilge et al., 2011; Sato et al., 2010; Choi et al., 2009; Villamarín-Salomón and Brustoloni, 2009; Dagon and Lee, 2009; Jiang et al., 2010; Antonakakis et al., 2010) have been proposed to study network activities by *passively* monitoring DNS queries between hosts and R-DNS servers, which are usually deployed in networks belonging to one management network range (e.g., an enterprise network or an ISP network). As the dynamics and diversity of network applications are rapidly

* Corresponding author. Tel.: +86 29 82664603; fax: +86 29 82664603
E-mail address: jtao@xjtu.edu.cn (J. Tao).

http://dx.doi.org/10.1016/j.jnca.2014.09.016
1084-8045/

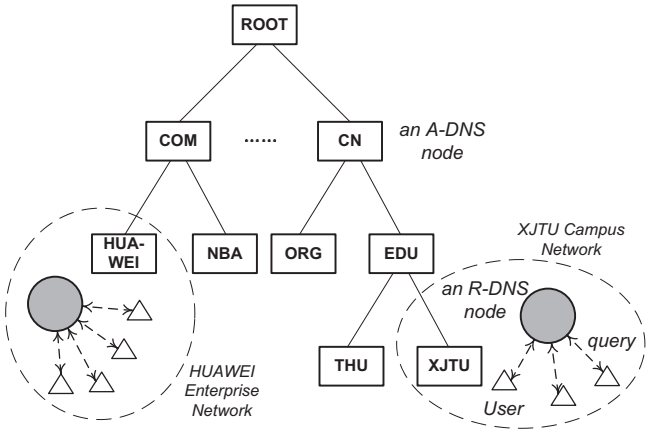
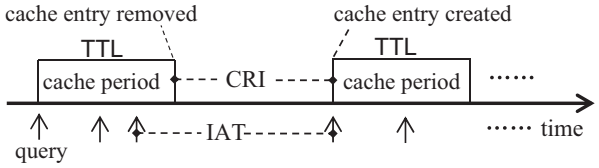


Fig. 1. A simplified DNS hierarchy.

a



b

Domain	Type	Resource Record	Remaining Time
www.google.com	A	74.125.128.147	223 (secs)
www.xjtu.edu.cn	A	202.117.1.13	458 (secs)
.....	...	.....	.....

Fig. 2. The TTL-based DNS caching mechanism demonstration. (a) Caching dynamics for a specific domain (IAT, interarrival time, CRI, cache refresh interval), Cache entry snapshot (type A: IP address).

increasing, there is a growing demand to design effective and scalable methods that can facilitate collaborative DNS traffic monitoring across different management network ranges. A large-scale, collaborative DNS monitoring framework can be tremendously beneficial. For example, it can reveal the propagation patterns of botnets and subsequently lead to effective mitigation solutions. Unfortunately, deploying passive monitor methods in a large-scale, collaborative manner is extremely challenging for at least two reasons. First, passive monitor methods mandate the installation and maintenance of traffic collection from various networks and thus may incur huge computational and management cost. In addition, the captured network traffic might contain sensitive information such as IP addresses, which might introduce significant privacy concerns.

An alternative DNS monitoring strategy is to actively probe R-DNS servers by taking advantage of their "*TTL-based caching mechanism*". Figure 2a shows the caching dynamics for a specific domain. Specifically, if a host (say $host_A$) issues a domain request to its R-DNS server and the R-DNS server does not have the record of this domain in its cache, it will retrieve the record and then return it to $host_A$, where the record contains the IP address(es) for the queried domain and a *time-to-live* (TTL) value. Meanwhile, the R-DNS server will cache this record for TTL units of time (i.e., cache period). Any host (say $host_B$) in the same network can obtain the cached record for a domain within the cache period by querying its R-DNS server. Note IAT represents the interarrival time between two successive DNS queries, and CRI represents the cache refresh interval between the expiration of one cache period and the start of the subsequent cache period. Figure 2b illustrates examples of cached records and the remaining TTL time of several domains. If we can model the correlation between the activities of the cached records and DNS queries from the end hosts, we can observe the activities of cached records and then estimate the DNS query activities. The activities of cached records can be obtained by actively and continuously probing the cache of an R-DNS server.

Despite the fact that the activities of cache records provide less fine-grained information compared to passively captured DNS queries, the active monitoring strategy offers several unique advantages. First, it drastically reduces the management cost since the monitoring system does not need to get access to the traffic of an R-DNS server. Instead, an arbitrary host in the network, which is willing to cooperate with the monitoring system by relaying cache probes to its R-DNS server(s), is sufficient for the collection of cache behavior. Second, it fundamentally minimizes the privacy concerns since all data collected are accessible to all hosts in the monitored network. Third, there might be an "honest-but-curios" A-DNS server that would like to infer the popularity of some of its domains with respect to the number of clients querying them from within a certain (possibly sensitive) network. The A-DNS server will only see queries from the R-DNS server of the target network, and essentially enables measuring the activities of cached records.

In spite of its promise, the fundamental challenge for active monitoring is how to deduce the query behavior of a domain based on the patterns of its cache entries. In practice (Akcan et al., 2008; Rajab et al., 2008), the query behavior of a domain is usually represented by its DNS query *interarrival times* (IATs) while its cache behavior is characterized by *cache refresh intervals* (CRIs), a sequence of time intervals between the removal and the immediately successive creation of its cache entry. Therefore, the specific design target is twofold:

- We need to design a model that can accurately characterize the relationship between DNS query IATs and CRIs for any given domain.
- Based on the proposed model, we need to profile DNS query IATs based on the observation of CRIs.

Accomplishing such design target is a challenging task since different domains may exhibit distinct DNS query patterns. As a result, the model has to be (i) sufficiently flexible to cope with the high diversity of DNS query patterns and meanwhile (ii) generic enough to simplify the design of the estimation algorithm. Existing methods (Akcan et al., 2008; Rajab et al., 2008) make the assumption that DNS query IATs for any domain follow the *exponential* distribution and therefore fail to profile DNS query patterns with high diversity. In fact, our empirical study based on large-scale datasets collected from real-world networks has demonstrated that the exponential distribution can only fits a small proportion of around 35% domains.

In this paper, we investigate the general distribution relationship between observed CRIs and the IATs, and design a novel model based on hyper-exponential distribution to characterize the correlation between DNS query IATs and CRIs for a given domain. The hyper-exponential distribution can profile the distributions of IATs with high diversity because it can control the distribution function according to Bernstein Theorem (Schilling et al.) by tuning the number of components to describe the high dispersion (i.e., variation) of the IATs for a domain. Our model is adaptive since it can automatically estimate the optimal number of components for different domains. Based on the new model, we then design an estimation algorithm that can accurately estimate the IATs based on observed CRIs. Specifically, we have made the following contributions:

1. Validating assumption of existing work using real-world DNS traffic:
   (a) Instead of simply assuming the distribution models of DNS query arrivals, we performed the goodness-of-fit test of distribution models based on extensive real-world DNS