Thesis Topic Proposal:
Measuring CDN Privacy from diverse geographic vantage points.

Daniel S. Berger

April 2, 2016

1 Introduction

Web-based services – like web shops, or news websites – need to ensure the fast delivery of their web and media content. This is facilitated by content-delivery networks (CDNs). CDNs consist of a network of caches within close vicinity of end-users (cf. [4, 6, 7]). This vicinity is key to having a high-bandwidth connection with short round-trip times to deliver content to a service's user base.

Achieving user vicinity is challenging as it requires maintaining caches in major cities round the world. Therefore, this task is commonly outsourced to commercial CDNs such as Akamai, Amazon CloudFront, Windows Azure, CloudFlare, or Limelight. Many commercial CDNs have a global footprint, and also offer extended services (e.g., the instant mitigation of denial of service attacks[1]). Today, the CDN business is essential to e-commerce and a multi-billion dollar market, which is expected to significantly grow in the coming years[2].

2 The Question of CDN Privacy

Given the central role of commercial CDNs in the business of web-based services, CDNs can be expected to hold valuable (and secret) business information. We seek to expose the extend of which the current CDN infrastructure can leak such information to an outside attacker.

We are a research group at the University of Kaiserslautern working in collaboration with the University of Oxford, Warwick University, Carnegie Mellon University, and the Swiss Federal Office for Defence Procurement (armasuisse). Our group brings together knowledge in network security (e.g., [3, 8]) and network performance (e.g., [1, 2]), which gives us a unique attack vector on challenges like CDN privacy. Recent research projects include the characterization of caching networks used by CDNs [2], and a proof-of-concept attack on CDN business information.

This proof-of-concept is a protoype with components: a web crawler compiles a list of CDN-cached content items and a network of measurement nodes then queries the CDN continuously for these items. While limited in scale and duration, the data gathered by our prototype proved the existence of information leaks. For example, we have been able to compile customer rankings of news websites, we found the most popular items in the web shop of a clothing retailer, and we revealed resource asymmetries than may be exploited by a denial of service attack.

3 Open Problems

Our existing prototype is limited in scope by its simplistic infrastructure.

- the prototype did not exploit the distributed nature of a CDN. This was due to limitations of the PlanetLab infrastructure and we used < 10 vantage points. As CDN nodes feature significant geographic diversity, concurrent observation of different CDN clusters would allow to characterize the geospatial footprint of content items (e.g., "where are most of Apple's smart watches sold");

[1] `https://www.cloudflare.com/features-security`
[2] `https://www.bizety.com/2015/08/15/cdn-market-size-in-2015-and-2019-2/`

- our observations covered a short period of time and few sources of content items. This was caused by stability issues of the measurement nodes and computational limits of a centralized data storage (AWS MySQL instance). Short periods severely restrict the creation and interpretation of popularity profiles. Such profiles can reveal request numbers to individual items or websites (estimates thereof), which then may allow to deduce secret information such as customer or sales numbers;
- the crawler and query nodes suffered from various technical deficiencies. For example, our current prototype focuses only on image files, is not able to interpret JavaScript code, interprets the CDN caching state (which may have up to ten values) as a binary value (cached/not cached), and covers only a single CDN (Akamai).

As a thesis topic, a student can work on building a significantly more potent infrastructure, which enables the further study and quantification of CDN information leaks. Building the infrastructure involves the redesign (various design decisions), the reimplementation, and redeployment of the prototype.

4 Anticipated Thesis Challenges and Tasks

The previous prototype focused on the evolution of cached items over short time scales (which is called the characteristic time [2]). The new prototype shall make available observations over longer periods of time. Here's a relevant scenario.

> A new web item (e.g. a new story on CNN.com, or a new item in the Apple store) is created and appears online. We want to timely **detect** this event and then continuously query its state in CDN caches **world wide**. This enables to study the web item's popularity evolution over time and space in order to (e.g.)
>
> - compare the long-term popularity evolution of two competing products/web services;
> - determine how long does it take to penetrate the US/ the world (geo spatial evolution)
> - study information dissemination (e.g., propagation of news)
> - cross-verify other studies/claims on a service's user base

This involves the following technical challenges

- continuous crawling a set of websites to timely detect new web items and submit them as a task to the measurement network. Target detection delay on the order of one to two hours.
- sending queries to the CDN from a system of distributed nodes (e.g., based on PlanetLab[3]). The sending of queries needs to optimized to cover a large number of items by (possibly) deciding how to prioritize certain objects or randomization
- fine-grained logging of the query response, e.g, Akamai CDN nodes answer with the current cache state[4][5] (TCP_HIT, MEM_HIT, TCP_MISS)
- efficient access of predictive analysis tools to the resulting *structured* measurement data (e.g. in the Microsoft Azure Machine Learning framework[6]).

The implementation framework (platform, programming model, etc) for crawler and query nodes is open.

[3] `https://www.planet-lab.org/` also see the prior work [5]
[4] `http://www.sobstel.org/blog/debugging-akamai/`
[5] `http://stats.wikimedia.org/archive/squid_reports/2013-12/SquidReportMethods.htm`
[6] `http://azure.microsoft.com/en-us/services/machine-learning/`

References

[1] M. A. Beck, S. A. Henningsen, S. B. Birnbach, and J. Schmitt. Towards a statistical network calculus - dealing with uncertainty in arrivals. In *The 33rd IEEE International Conference on Computer Communications (INFOCOM 2014)*, Toronto, Canada, 2014.

[2] D. S. Berger, P. Gland, S. Singla, and F. Ciucu. Exact analysis of TTL cache networks. *Performance Evaluation*, 79(0):2–23, 2014. Special Issue: Performance 2014.

[3] D. S. Berger, F. Gringoli, N. Facchi, I. Martinovic, and J. Schmitt. Gaining insight on friendly jamming in a real-world ieee 802.11 network. In *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks*, WiSec '14, pages 105–116, New York, NY, USA, 2014. ACM.

[4] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, and B. Weihl. Globally distributed content delivery. *Internet Computing, IEEE*, 6(5):50–58, 2002.

[5] C. Huang, A. Wang, J. Li, and K. W. Ross. Measuring and evaluating large-scale cdns. In *ACM IMC*, volume 8, 2008.

[6] Q. Huang, K. Birman, R. van Renesse, W. Lloyd, S. Kumar, and H. C. Li. An analysis of facebook photo caching. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, pages 167–181. ACM, 2013.

[7] E. Nygren, R. K. Sitaraman, and J. Sun. The akamai network: a platform for high-performance internet applications. *ACM SIGOPS Operating Systems Review*, 44(3):2–19, 2010.

[8] M. Schäfer, V. Lenders, and J. B. Schmitt. Secure Track Verification. In *IEEE Symposium on Security and Privacy (S&P '15)*, San Jose, CA, USA, May 2015.