

# I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis

Brad Miller<sup>1</sup>, Ling Huang<sup>2</sup>, A.D. Joseph<sup>1</sup>, and J.D. Tygar<sup>1</sup>

<sup>1</sup> UC Berkeley, USA

<sup>2</sup> Intel Labs, USA

**Abstract.** Revelations of large scale electronic surveillance and data mining by governments and corporations have fueled increased adoption of HTTPS. We present a traffic analysis attack against over 6000 webpages spanning the HTTPS deployments of 10 widely used, industry-leading websites in areas such as healthcare, finance, legal services and streaming video. Our attack identifies individual pages in the same website with 90% accuracy, exposing personal details including medical conditions, financial and legal affairs and sexual orientation. We examine evaluation methodology and reveal accuracy variations as large as 17% caused by assumptions affecting caching and cookies. We present a novel defense reducing attack accuracy to 25% with a 9% traffic increase, and demonstrate significantly increased effectiveness of prior defenses in our evaluation context, inclusive of enabled caching, user-specific cookies and pages within the same website.

## 1 Introduction

HTTPS is far more vulnerable to traffic analysis than has been previously discussed by researchers. In a series of important papers, a variety of researchers have shown a number of traffic analysis attacks on SSL proxies [1,2], SSH tunnels [3,4,5,6,7], Tor [3,4,8,9], and in unpublished work, HTTPS [10,11]. Together, these results suggest that HTTPS may be vulnerable to traffic analysis. This paper confirms the vulnerability of HTTPS, but more importantly, gives new and much sharper attacks on HTTPS, presenting algorithms that decrease errors 3.9x from the best previous techniques. We show the following novel results:

- Novel attack technique capable of achieving 90% accuracy over 500 pages hosted at the same website, as compared to 60% with previous techniques
- Impact of caching and cookies on traffic characteristics and attack performance, affecting accuracy as much as 17%
- Novel defense reducing accuracy to 25% with 9% traffic increase; significantly increased effectiveness of packet level defenses in the HTTPS context

We evaluate attack, defense and measurement techniques on websites for healthcare (Mayo Clinic, Planned Parenthood, Kaiser Permanente), finance (Wells Fargo, Bank of America, Vanguard), legal services (ACLU, Legal Zoom) and streaming video (Netflix, YouTube).

We design our attack to distinguish minor variations in HTTPS traffic from significant variations which indicate distinct webpages. Minor traffic variations may be caused by caching, dynamically generated content, or user-specific content including cookies. To distinguish minor variations, our attack employs clustering and Gaussian similarity techniques to transform variable length traffic into a fixed width representation. Due to similarity with the Bag-of-Words approach to text analysis, we refer to our technique as Bag-of-Gaussians (BoG). We augment our technique with a hidden Markov model (HMM) leveraging the link structure of the website and further increasing accuracy. Our approach achieves substantially greater accuracy than attacks developed by Panchenko *et al.* (Pan) [8], Liberatore and Levine (LL) [6], and Wang *et al.* [9].<sup>1</sup>

We also present a novel defense technique and evaluate several previously proposed defenses. In the interest of deployability, all defenses we evaluate have been selected or designed to require minimal state. Our evaluation demonstrates that some techniques which are ineffective in other traffic analysis contexts have significantly increased impact in the HTTPS context. For example, although Dyer *et al.* report exponential padding as decreasing accuracy of the Panchenko classifier from 97.2% to 96.6% on SSH tunnels with website homepages [5], we observe a decrease from 60% to 22% in the HTTPS context. Our novel defense reduces the accuracy of the BoG attack from 90% to 25% while generating only 9% traffic overhead.

We conduct our evaluations using a dataset of 463,125 page loads collected from 10 websites during December 2013 and January 2014. Our collection infrastructure includes virtual machines (VMs) which operate in four separate collection modes, varying properties such as caching and cookie retention across the collection modes. By training a model using data from a specific collection mode and evaluating the model using a different collection mode, we are able to isolate the impact of factors such as caching and user-specific cookies on analysis results. We present these results along with insights into the fundamental properties of the traffic itself.

Our evaluation spans four website categories where the specific pages accessed by a user reveal private information. The increased importance of contents over existence of communication is present in traditional privacy concepts such as patient confidentiality or attorney-client privilege. We examine three websites related to healthcare, since the page views of these websites have the potential to reveal whether a pending procedure is an appendectomy or an abortion, or whether a chronic medication is for diabetes or HIV/AIDS. We also examine legal websites, offering services spanning divorce, bankruptcy and wills and legal information regarding LGBT rights, human reproduction and immigration. As documented by Chen *et al.*, specific pages accessed within financial websites may reveal income levels, investment and family details; hence we examine three financial websites [12]. Lastly, we examine two streaming video sites, as the Netflix privacy breach demonstrates the importance of streaming video privacy.

---

<sup>1</sup> To facilitate further research, code and data from this work are available for download at <http://secml.cs.berkeley.edu/pets2014>.