Towards Classifying Third-Party Web Services at Scale

David Gugelmann
ETH Zurich, Switzerland
gugelmann@tik.ee.ethz.ch

Bernhard Ager
ETH Zurich, Switzerland
bager@tik.ee.ethz.ch

Vincent Lenders
armasuisse, Switzerland
vincent.lenders@armasuisse.ch

ABSTRACT

Many people are concerned about privacy in the Web. This has resulted in the emergence of various tools and suggestions for blocking tracking services. However, in order to block or analyze tracking services, these services have to be identified first. In this work we develop and evaluate a set of features for identifying tracking services in HTTP traffic. Our approach enables us to verify the effectiveness of existing blocking approaches as well as to find more candidates for blacklisting.

Categories and Subject Descriptors

C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*

Keywords

Privacy; tracking; HTTP; network measurement

1. INTRODUCTION

Today's Web sites include a multitude of elements from third-party services, such as images from content distribution networks (CDN), advertising and analytics (AA) scripts, and like buttons from social networks. When an object is loaded from a third-party, the third-party can collect information on the activity of the user. A third-party service included by many Web sites, such as major AA services [11, 16], can thus create exhaustive user profiles. The more detailed a user profile, the higher the profits [7].

Several anti-AA browser plugins try to block user tracking, e.g., Adblock Plus [14], Ghostery [15], or ShareMeNot [16]. However, these plugins rely on manually maintained black lists. AA services must therefore first be identified by the maintainers of the black lists.

We present a way towards characterizing Web services and identifying services that can collect exhaustive user profiles. We analyze HTTP traffic at the border gateway of a network and introduce several key features describing information flows to Web services. Our analysis in a large university network shows that these features seem to be suitable to classify third-party services without requiring detailed manual investigation. Our approach can be used to suggest candidates for black lists or to periodically check if deployed blocking policies are still effective against new privacy threats.

Related work in the area of HTTP traffic characterization [2–6, 9, 13] focused on *download* traffic, i.e., information flow *from* Web services. In contrast, our work characterizes Web services by looking mainly at the *upload* traffic. Previous studies on AA services relied on *active* measurements to estimate their prevalence on major Web sites [11, 16] and to describe the information leaking to third-parties [12]. In contrast, we use *passive* measurements to characterize Web services from real-world traces. Others have analyzed the associated money flow [7] and resulting privacy issues [10].

2. TRAFFIC FEATURES

The HTTP protocol is known to be highly redundant [1,9]. In order to estimate the actual amount of user-information being transmitted in HTTP requests to Web services, we use a method [8] that greatly reduces the redundancy in HTTP requests. Using this method, we approximate for each service the effective amount of information uploaded without HTTP's protocol redundancy. The volume of this upload data is referred to as *information bytes*. Based on this measure of information transfer, we derive the following features: (A) The amount of information received by a Web service acting as a third-party. (B) The percentage of third-party information bytes relative to the total number of information bytes going to a service. Domains that are not visited directly but only appear as third-parties, will have a high value. (C) The number of domains embedding an element that is hosted by the Web service. (D) The average amount of information transmitted to the service per HTTP request. (E) The percentage of clients accounting for 95 % of information reflects the upload volume client distribution. A low value indicates that the service only ranks high because few users upload a large amount of information (e.g. videos). A high value indicates the service collects information from many users. (F) Response / request byte ratio is a measure that reflects the ratio between bytes being received from a Web service and bytes being sent to it.

3. CASE STUDY

To demonstrate the utility of our features to classify third-party Web services, we collect real HTTP traffic of 15 k clients in a university network and analyze this traffic us-

CoNEXT Student Workshop'14, December 2, 2014, Sydney, Australia.
Copyright 2014 ACM 978-1-4503-3282-8/14/12 ...$15.00.
http://dx.doi.org/10.1145/2680821.2680838.

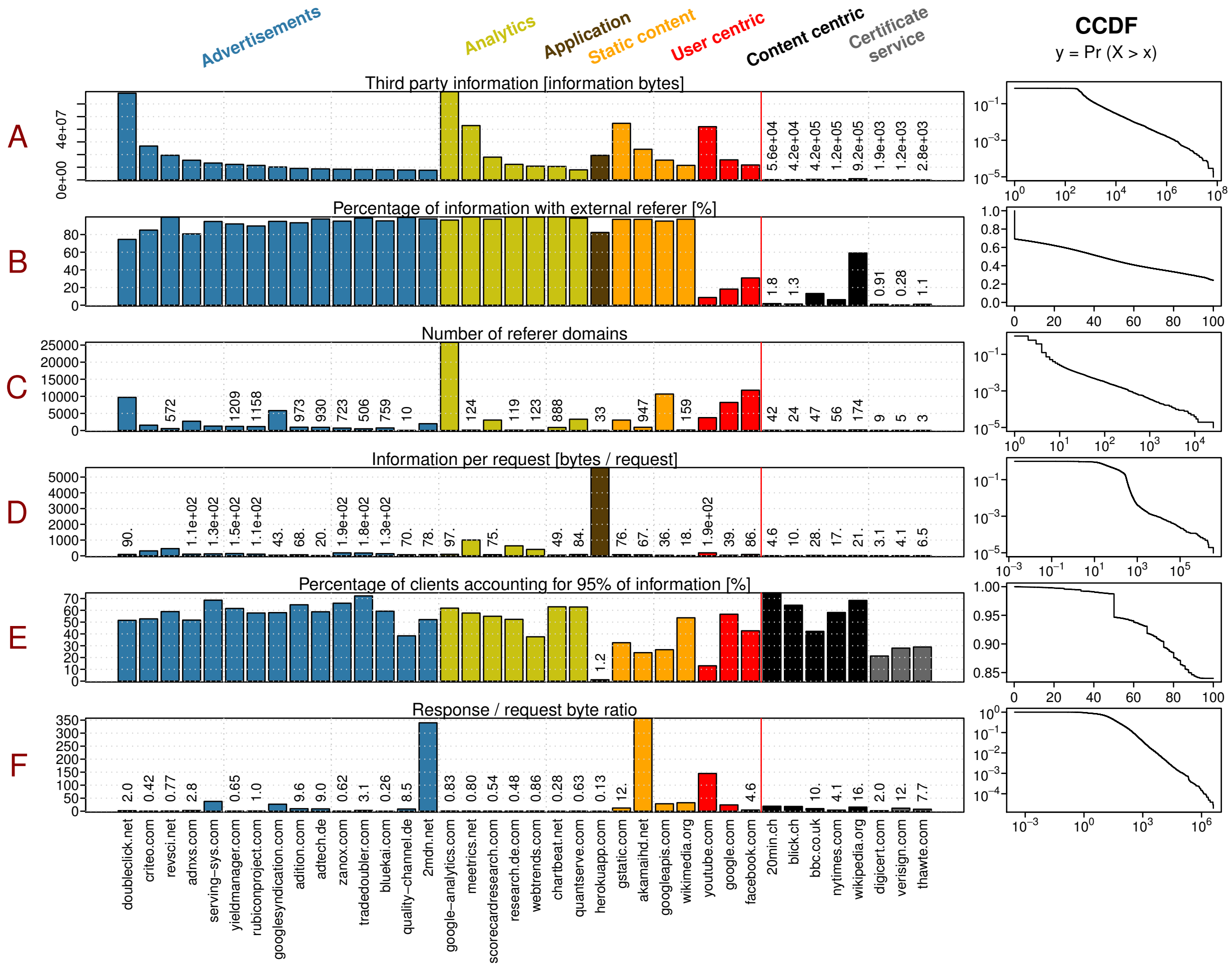

Figure 1: Comparison of service features of the 30 top third-party domains and selected domains on the right side of the red separator line for comparison. We manually classified the services as encoded by the colors. The CCDFs show the distributions of the respective features over all domains.

ing our features. The analyzed trace spans a duration of 24 h. It contains 75 GB of upload and 3.4 TB of download traffic to 103 k different domains.

Figure 1 shows the 30 third-party Web services that receive most information bytes and are visited by at least 100 clients. We additionally include content-centric Web sites for comparison (i.e., services with no or only little personalized content, such as news sites and Wikipedia). We see that AA services dominate the top third-party services followed by CDNs. Feature B shows that the AA services and CDNs receive most information in the role of a third-party. In contrast, user-centric services (i.e., services with personalized content) and news services receive most information as first-party since users often visit these services directly. Row D shows that the news services receive less than 30 information bytes per request, while most AA services receive more than twice as much information per request. The service *herokuapp.com* – a cloud application platform – has similar properties as AA services for features B and C but clearly stands out in features D and E. These features indicate that this service only ranks among the top 30 because of large requests by very few clients. Row F shows that analytics services have a really low response to request ratio. This is intuitive considering that these services usually do not provide any actual content. In contrast, services that provide content (the CDN Akamai, YouTube, and *2mdn.net*, which provides ads for Google) have a higher score.

To summarize, our results show that our proposed features are able to identify good candidates for anti-AA lists by looking for services with high values for metrics B to E and a low value for metric F.

4. SUMMARY AND FUTURE WORK

We present a new approach towards analyzing third-party services. We propose several features that allow to classify Web services in order to reduce the burden of digging into individual HTTP requests. Our case study, which is based on HTTP requests of 15 k clients, shows that our proposed features make AA services stand out and can therefore be used to identify possible candidates for AA black lists.

As future work, we plan to explore the characteristics of more Web service classes, including first-party services.

5. ACKNOWLEDGEMENT

This work was partially supported by the Zurich Information Security Center. It represents the views of the authors.

6. REFERENCES

[1] A. Anand, C. Muthukrishnan, A. Akella, and R. Ramjee. Redundancy in network traffic: findings and implications. In *SIGMETRICS '09*, 2009.

[2] M. F. Arlitt and C. L. Williamson. Web server workload characterization: the search for invariants. In *SIGMETRICS '96*, 1996.

[3] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web*, 2(1-2), 1999.

[4] M. Butkiewicz, H. V. Madhyastha, and V. Sekar. Understanding website complexity: Measurements, metrics, and implications. In *IMC '11*, 2011.

[5] M. E. Crovella and A. Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Trans. Netw.*, 5(6), Dec. 1997.

[6] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. C. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, Dec. 1997.

[7] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *IMC '13*, 2013.

[8] D. Gugelmann, B. Ager, and V. Lenders. Towards understanding http upload traffic. Under submission.

[9] S. Ihm and V. S. Pai. Towards understanding modern web traffic. In *IMC '11*, 2011.

[10] B. Krishnamurthy. I know what you will do next summer. *SIGCOMM Comput. Commun. Rev.*, 40(5), Oct. 2010.

[11] B. Krishnamurthy, D. Malandrino, and C. E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, 2007.

[12] B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Web 2.0 Security and Privacy Workshop*, 2011.

[13] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband internet traffic. In *IMC '09*, 2009.

[14] Adblock Plus. `https://adblockplus.org`.

[15] Ghostery. `https://www.ghostery.com`.

[16] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. In *NSDI '12*, 2012.