The Tangled Web We Weave: Analyzing the Web's Dependency Graph

Sunjay Cauligi† Brian Johannesmeyer† Ariana Mirian† Gary Soeller†
† University of California, San Diego
{scauligi, bjohanne, amirian, gsoeller}@cs.ucsd.edu

ABSTRACT

Abstract here

Keywords

Keywords

1. INTRODUCTION

Introduction

2. RELATED WORK

The importance of the world web in our daily lives is a well worn justification for its study. Of particular importance to this work is the comprehensively studied area of web caching in which researchers observe that popular websites exist and that storing local copies of these sites can provide performance gains and bandwidth reductions [3]. There are a myriad of techniques for performing caching that rely on locality of reference for pages and objects [16]. While exploring the long history of web caching is beyond the scope of the paper, we note several pieces of work that tease out specific properties of web pages. For example, researchers have noted that caching can be performed at finer-grained levels than pervious discussed [20] and others have pointed out that pages consist of extraneous content that might be suitable for filtering [8]. Further, recent studies in this area explore interesting properties of web pages and objects for use in specific areas such as the broadband domain [18] as well as for use by mobile devices [15].

Several measurement papers have explored the recent trends in page-level characteristics as well as in redundancy and caching [7]. A striking finding that shares motivation with this work is the increased complexity of web pages. Ihm et al. [4, 7], for example, explore metrics for measuring website complexity and characterize object types as well as their locations, servers, origins. Much of this complexity and third-party content has been attributed to online advertising. Several papers explore this space, but most notably Barford et al. [1] and Guha et al. [6].

One implication of this complexity is an increase in the opaque nature of the interactions on the web. Several pieces of work shine light on this problem at a coarse-grained level, for example, by using DNS to understand transactions [2] or by examining interactions and dependancies between websites [14]. An additional consequence of complexity is performance. A new body of work explores how to understand and improve web page load performance [19]. A particularly relevant piece of work in this domain is [11], which creates fine-grained dependency graphs in order to help prioritize object loads.

The potential security and privacy consequences of third-party content are well explored. Initial work in the drive by download arena [5, 12, 13] note that in addtion to webpage compromise, inclusion of object from other domains create risk. Malicious advertising, as a particular form of third party content on a website, has received considerable attention [9, 21]. Another form of third-party content that has been an active area of measurement is exploring the privacy implications of online tracking [10, 17].

3. FUTURE WORK

Future work

4. CONCLUSION

Conclusion

5. REFERENCES

[1] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 597–608, New York, NY, USA, 2014. ACM.

[2] I. N. Bermudez, M. Mellia, M. M. Munafo, R. Keralapura, and A. Nucci. Dns to the rescue: Discerning content and services in a tangled web. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, pages 413–426, New York, NY, USA, 2012. ACM.

[3] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, pages 126–134. IEEE, 1999.

[4] M. Butkiewicz, H. V. Madhyastha, and V. Sekar. Understanding website complexity: measurements, metrics, and implications. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 313–328. ACM, 2011.

[5] M. Cova, C. Kruegel, and G. Vigna. Detection and analysis of drive-by-download attacks and malicious javascript code. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 281–290, New York, NY, USA, 2010. ACM.

[6] S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 81–87, New York, NY, USA, 2010. ACM.

[7] S. Ihm and V. S. Pai. Towards understanding modern web traffic. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 295–312. ACM, 2011.

[8] B. Krishnamurthy and C. E. Wills. Cat and mouse: Content delivery tradeoffs in web access. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 337–346, New York, NY, USA, 2006. ACM.

[9] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: Understanding and detecting malicious web advertising. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 674–686, New York, NY, USA, 2012. ACM.

[10] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 413–427. IEEE, 2012.

[11] R. Netravali, J. Mickens, and H. Balakrishnan. Polaris: Faster page loads using fine-grained dependency tracking. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, Santa Clara, CA, Mar. 2016. USENIX Association.

[12] N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose. All your iframes point to us. In *Proceedings of the 17th Conference on Security Symposium*, SS'08, pages 1–15, Berkeley, CA, USA, 2008. USENIX Association.

[13] N. Provos, D. McNamee, P. Mavrommatis, K. Wang, and N. Modadugu. The ghost in the browser analysis of web-based malware. In *Proceedings of the First Conference on First Workshop on Hot Topics in Understanding Botnets*, HotBots'07, pages 4–4, Berkeley, CA, USA, 2007. USENIX Association.

[14] E. Pujol, P. Richter, B. Chandrasekaran, G. Smaragdakis, A. Feldmann, B. M. Maggs, and K.-C. Ng. Back-office web traffic on the internet. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 257–270, New York, NY, USA, 2014. ACM.

[15] F. Qian, K. S. Quah, J. Huang, J. Erman, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck. Web caching on smartphones: Ideal vs. reality. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 127–140, New York, NY, USA, 2012. ACM.

[16] M. Rabinovich and O. Spatscheck. *Web caching and replication*. Addison-Wesley Boston, USA, 2002.

[17] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 12–12. USENIX Association, 2012.

[18] S. Sundaresan, N. Feamster, R. Teixeira, and N. Magharei. Measuring and mitigating web performance bottlenecks in broadband access networks. In *ACM Internet Measurement Conference*, 2013.

[19] X. S. Wang, A. Balasubramanian, A. Krishnamurthy, and D. Wetherall. Demystifying page load performance with wprof. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 473–485, Lombard, IL, 2013. USENIX.

[20] X. S. Wang, A. Krishnamurthy, and D. Wetherall. How much can we micro-cache web pages? In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 249–256, New York, NY, USA, 2014. ACM.

[21] A. Zarras, A. Kapravelos, G. Stringhini, T. Holz, C. Kruegel, and G. Vigna. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 373–380, New York, NY, USA, 2014. ACM.