The Web is Smaller than it Seems

Craig A. Shue, Andrew J. Kalafut, and Minaxi Gupta
Computer Science Department, Indiana University at Bloomington
{cshue, akalafut, minaxi}@cs.indiana.edu

ABSTRACT

The Web has grown beyond anybody's imagination. While significant research has been devoted to understanding aspects of the Web from the perspective of the documents that comprise it, we have little data on the relationship among servers that comprise the Web. In this paper, we explore the extent to which Web servers are co-located with other Web servers in the Internet. In terms of the location of servers, we find that the Web is surprisingly smaller than it seems. Our work has important implications for the availability of Web servers in case of DoS attacks and blocklisting.

Categories and Subject Descriptors

C.2.5 [Local and Wide-Area Networks]: Internet

General Terms

Measurement, Experimentation

Keywords

World Wide Web, server co-location, block lists, DoS attacks

1. INTRODUCTION

The Web is vast. According to some estimates there are at least 10 billion documents comprising the Web. While many aspects of the Web have been explored from the perspective of the documents that make up the Web, much less is known about the relationships between the servers that host these documents. At one extreme, well provisioned Web sites distribute content using replicated servers or content distribution networks (CDNs). At the other extreme, hosting services host thousands of domains on a handful of machines, causing Web servers to be co-located. The phenomenon of sharing extends beyond the Web servers, for even the DNS servers which direct clients to these Web servers are often co-located.

Understanding co-location of Web and DNS servers is important. It can help judge the extent to which targeted denial-of-service (DoS) attacks can hurt the availability of services on the Web. Another reason to study this phenomenon pertains to the use of block lists in the Internet. These lists serve to block communication with sites that host malicious or unwanted content and programs such as spyware and adware. Many of these block lists are IP-based, implying that if they intend to block a certain Web site, they block the IP address corresponding to the machine hosting that Web site. This hurts the availability of all the co-hosted Web sites, many of which may be useful (or at least harmless).

In this paper, we undertake the first study to quantify *co-location* of Web servers and their corresponding authoritative DNS servers. We use extensive data sets regarding Web server names and addresses throughout the Internet toward this goal. Using popular block lists, we also examine the collateral damage due to co-hosting of Web servers.

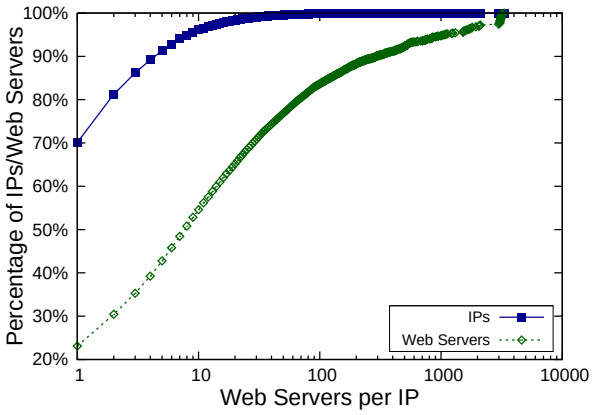
We find that as much as 60% of the Web servers are co-hosted with $10,000$ or more other Web servers, indicating that the Internet contains many small co-hosted Web servers. Likewise, more than 95% of Web servers share their AS with 1000 or more other Web servers. We additionally find that heavily co-hosted Web servers contribute much less traffic than Web servers that are not co-hosted, confirming that popular servers are not co-located, while less popular servers co-locate more frequently. When considering block lists, we find the vast majority of blocked Web servers are hosted on IPs hosting 100 Web servers or more. This indicates there may be a great deal of collateral damage with IP blocking. Finally, when looking at authoritative DNS servers, we see a high degree of co-location on a very small number of DNS servers, which may result in the Web being fragile from a DNS perspective.

The rest of this paper is organized as follows. In Section 2, we describe the primary data sets used in this paper. Section 3 examines the extent of Web server co-location. In Section 4, we examine how block lists affect the co-located Web servers. Section 5 looks at the extent of DNS server co-location. Finally, in Section 6, we survey related work and conclude with discussion in Section 7.

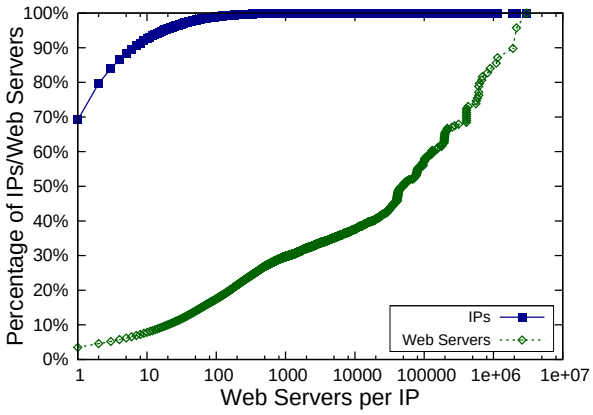
2. PRIMARY DATA SETS

We use two primary data sets throughout this paper. The first is from the DMOZ Open Directory Project [1]. The project contains user submitted links and is the largest and most comprehensive directory of Web URLs. A typical URL from DMOZ data contains several pieces of information. For example, in the URL `www.example.com/content.html`,

IMC'07, October 23–26, 2007, San Diego, CA, USA.



(a) DMOZ Data set.



(b) Zone files.

Figure 1: CDFs showing Web servers per IP address as a percentage of IP addresses and Web servers.

`www.example.com` is the Web server name, which belongs to domain `example.com` and top level domain (TLD) `.com`. The actual file being accessed is `content.html`.

The DMOZ data set covers over 234 different TLDs, making it an international data set covering over 90% of the TLDs. We use a snapshot of DMOZ data from October 28th, 2006. From the DMOZ URLs, we extract the names of unique Web servers offering content. We conduct DNS lookups on each of these names to get their corresponding IP addresses, which are returned in the form of type `A` resource records. The unique IP addresses from these DNS responses are used to infer the relationship between Web servers and IP addresses. If a Web server name resolves to multiple IP addresses, we select the lowest IP address returned. This helps avoid counting a cluster of Web hosting servers multiple times.

The second data set contains DNS zone files [2] from the `.net` and `.com` TLDs. These zone files list each of the domains in the respective TLD zones. The data presented here is from the zone files we obtained on March 7th, 2007. To obtain the Web server name for each domain listed in the zone file, we simply prefix each domain name with "www.", since most Web servers are named in this fashion. We then resolve each Web server name into an IP address using DNS queries, as we do for the DMOZ data set.

The DMOZ data set contains URLs corresponding to `.com` and `.net` TLDs, in addition to other TLDs from around the world. In fact, about half of the DMOZ Web servers correspond to these TLDs. Since these domains are exhaustively listed in the zone files, we eliminate them from the DMOZ data. Henceforth, when we refer to the DMOZ data set, we mean its curated version which excludes entries from the `.com` and `.net` TLDs. Together, these two data sets represent a sizeable chunk of the Web today, since they contain 75.7 million of the 128 million domains registered worldwide in June 2007 [3].

Table 1 shows the number of unique URLs, Web servers, and IP addresses contained in both the data sets. Two things are noteworthy about this table. First, the `.com` and `.net` TLDs contained in the zone files by themselves contain an order of magnitude more domain names than the rest of the TLDs represented in the DMOZ data. Second, for each data set, the number of unique IP addresses belonging to the Web servers is also an order of magnitude less than the number of Web servers themselves. This is an initial indication that many Web servers are *co-located*. We explore this in detail in Section 3.

	DMOZ Data (curated)	Zone Files
Number of URLs	4,667,792	-
Unique Web Servers	1,487,481	74,326,215
`A` Records Received	1,396,998	71,855,113
Unique IPs	487,797	3,641,329
TLDs Represented	232	2
Unique ASes Represented	12,374	18,356

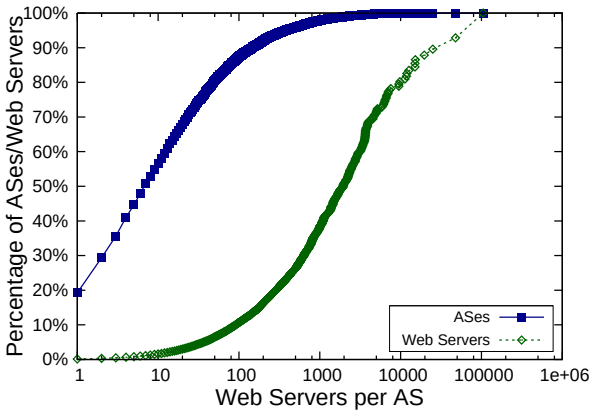
Table 1: Overview of DMOZ and zone files data.

3. WEB SERVER CO-LOCATION

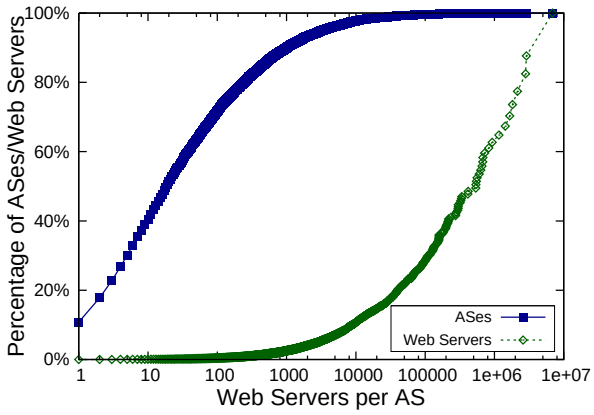
We begin by investigating where the Web servers are located in terms of the IP addresses of machines that these servers are hosted on. Notice that our analysis focuses on the actual Web servers and does not include the servers belonging to CDNs, which many well-provisioned Web sites tend to use.

Figure 1(a) shows the number of Web servers per unique machine as a percentage of IP addresses and Web servers for the DMOZ data set. Figure 1(b) shows similar information for the zone files. Note that the X-axis is a log scale in both figures. From these figures, we draw several key observations. First, they show that most machines host only a handful of Web servers. As many as $69-71\%$ of the IPs in both our data sets host just one Web server. While this may lead one to conclude that there is a one-to-one correspondence between the Web servers and the IP addresses, the story changes completely when one looks at the Web servers per IP address as a percentage of Web servers. We find that only between $4-24\%$ of Web servers in our two data sets are hosted on a machine by themselves. The rest are *co-hosted* on the same machines with other Web servers. This implies that while a rather small percentage of well-provisioned Web servers employ dedicated machines to host, the rest are co-hosted.

Figures 1(a) and 1(b) also illustrate the differences between the DMOZ and zone files data sets. First, the X-axis



(a) DMOZ Data set.



(b) Zone files.

Figure 2: CDFs showing Web servers per AS as a percentage of ASes and Web servers.

differs in that the zone files have Web servers that have orders of magnitude more Web servers per IP address than those in DMOZ data. Since zone files exhaustively represent the `.com` and `.net` TLDs, this implies that more Web servers in these TLDs are co-located. Second, Figure 1(a) also shows that a much larger percentage of Web servers represented in the DMOZ data are hosted either by themselves or are co-hosted with a small number of other Web servers. Specifically, as much as 84% of the DMOZ Web servers are co-hosted with 100 or fewer other Web servers while only 15% of the Web servers contained in the zone files are co-hosted with 100 or fewer other Web servers. Further, under 6% of the DMOZ Web servers are co-hosted with 1,000 or more Web servers while as much as 65% of the Web servers contained in the zone files are co-hosted with 1,000 or more Web servers. In fact, as much as 60% of the Web servers in zone files are co-hosted with 10,000 or more Web servers! There could be two explanations for the differences in the two data sets. First, TLDs outside of `.com` and `.net` are co-located less often. Second, that the DMOZ data may be dominated by well-provisioned Web servers.

3.1 Co-location in Terms of ASes

Here, we analyze Web server co-location as seen from the perspective of ASes the Web servers are located in.

Additional Data Used: In order to infer co-location in terms of ASes advertised by these ASes, we gather a third data set: a BGP routing table from a router in the Route Views Project [4]. The table contains 237,819 prefixes advertised by BGP routers in the Internet, along with the ASes that originate these prefixes. We use an April 22, 2007 snapshot of the routing table, which is from around the same time as when we performed the DNS resolutions on Web server names. For each IP address, we perform a longest prefix match on this table to obtain the AS for the IP address.

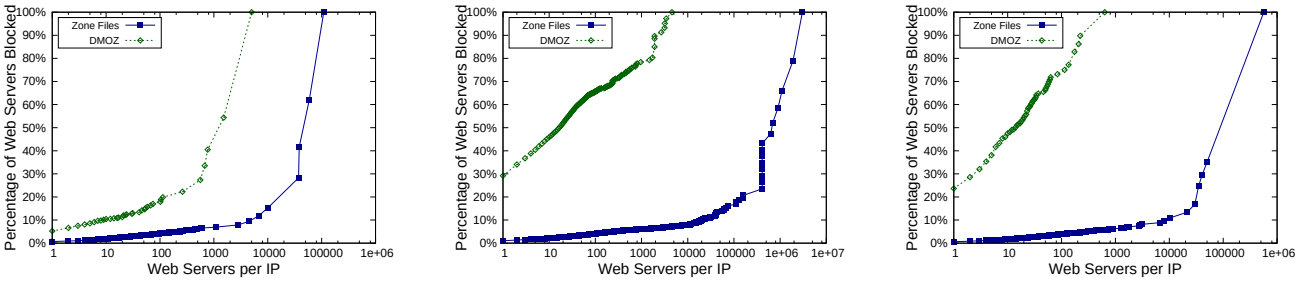
Analysis: Figures 2(a) and 2(b) show the co-location in terms of Ases for the DMOZ data and zone files respectively. These figures show that 19.27% of ASes in the DMOZ data set and 10.78% in the zone file data have only one Web server. However, only a very small percentage of Web servers are hosted in an AS by themselves. The case is more pronounced for the zone files, where even fewer Web servers exist by themselves. Specifically, more than 60% of the DMOZ Web servers share their ASes with 1,000 or more other Web servers. Correspondingly, more than 95% of the Web servers in the zone files share their AS with 1,000 or more other servers. These findings indicate that the Web is even smaller when seen from the perspective of ASes the Web servers belong to.

3.2 Are Popular Web Servers Co-located?

Given the extent of co-location, it is obvious to wonder if popular Web servers are also co-located. To determine the extent of impact due to the unavailability of co-located servers, we now examine the amount of traffic associated with the IP addresses corresponding to the servers in our DMOZ and zone files data sets. While more exhaustive traffic measurement techniques exist, such as those in [5], they do not analyze the traffic for individual Web sites. Hence, we perform our own simple analysis.

Additional Data Used: Toward this goal, we collect Netflow [6] logs from the Indiana University campus. In this capture, we analyze two days, February 4 and February 15, 2007, to get both weekend and weekday sample points. The logs contain the total number of flows, packets, and bytes transfered from our campus to the remote IP addresses. We cross-reference the IP addresses contained in the Netflow logs with the IP addresses corresponding to the Web servers contained in DMOZ and zone file data sets to determine the distribution of traffic destined to IPs hosting a varying numbers of Web servers. Our data offers a limited view because it comes from a single academic site. However, it gives a sense of the relationship between traffic and Web server co-location.

Analysis: Table 2 shows the traffic for Web servers contained in DMOZ data and zone files. We find that relatively little traffic is destined to IPs hosting 50 Web servers or more, both for the weekday and weekend. In contrast, a significant amount of traffic is destined to IP addresses associated with only one Web server. The effect is more pronounced for the DMOZ data sets, which represents TLDs outside of `.com` and `.net`. Irrespective, both data sets confirm the intuition that popular servers are less likely to be co-located, while less popular servers co-locate more frequently.



(a) Ad-block. (b) Copy-block. (c) Spy-block.

Figure 3: CDF of the percent of Web servers blocked with the indicated block lists.

	February 4	February 15
Distinct IPs in Netflow data	2,948,641	3,338,043
DMOZ Data		
IPs in Common	26,704 (0.01%)	43,316 (1.30%)
IPs with only 1 Server		
Flows	48.20%	46.76%
Bytes	49.31%	67.33%
Packets	50.96%	69.04%
IPs with 50+ Servers		
Flows	12.97%	12.02%
Bytes	2.73%	4.43%
Packets	8.06%	6.06%
Zone Files Data		
IPs in Common	49,508 (1.7%)	118,348 (3.55%)
IPs with only 1 Server		
Flows	39.79%	41.64%
Bytes	50.73%	55.92%
Packets	42.57%	49.51%
IPs with 50+ Servers		
Flows	14.02%	12.03%
Bytes	4.30%	4.96%
Packets	11.20%	9.94%

Table 2: Netflow traffic from our department to Web servers in both data sets.

4. CO-LOCATION AND BLOCK LISTS

Co-location of Web servers has important implications on the availability of Web content. This is because a strategic denial-of-service (DoS) attack on a handful of machines or routers can make a large number of Web servers unavailable. When looking at the zone files data set, we note that just 645 IPs, representing only 0.02% of the IPs in the data set, host 64.00% of the Web servers. With the growing spread of botnets, targeted attacks can easily overwhelm these servers, making a significant portion of the Internet unavailable.

Co-location also impacts another practice widely used in the Internet today: that of blocking communication with known malicious machines and Web sites. A variety of block lists are available today to avoid communication with known spammers, malware[1] serving Web sites and machines, and known bots. The block lists are either *IP-based* or *Web site-based*. The IP-based block lists identify malicious entities by IP addresses and seek to block communication with machines whose IP addresses are listed in the block lists. Similarly, Web site-based block lists seek to block all communication with Web sites listed in the block lists. While Web site-based block lists are unlikely to cause much collateral damage on co-hosted Web sites, an IP-based block list may be quite damaging. We explore the extent of this damage in this section.

[1]The term malware is often used for *mal*icious soft*ware*.

4.1 Additional Data Used

To test the impact of IP-based block lists on the Web servers contained in DMOZ data and zone files, we used a set of block lists from the Bluetack Internet Security Solutions Web site [7], which we obtained on May 5, 2007.

We used three different flavors of block lists designed to block certain kinds of Web sites. Each does so by specifying an IP address corresponding to the Web site. The first aims to block IP addresses belonging to sites hosting advertisements and/or those providing advertisement tracking services (referred to as *Ad-block* subsequently). The second blocks sites engaging in copyright enforcement (referred to as *Copy-block* subsequently). These sites may employ techniques to detect copyright violators. The third block list aims to block Web sites serving spyware[2] (referred to as *Spy-block* subsequently).

Surprisingly, instead of specifying individual IP addresses to be blocked, these block lists specify a range of IP addresses to be filtered through software installed on the client. This raises the suspicion that these lists may end up blocking good machines sharing the same prefix ranges as known bad machines. For our purpose, we extracted individual IP addresses by expanding these ranges. Table 3 presents the number of IP addresses contained in each of these block lists. As the table shows, while Ad-block and Spy-block contain a modest number of IPs to block, the Copy-block list blocks entire organizational prefixes, resulting in a number of blocked IPs that is potentially larger than the number of total known hosts in the Internet today!

	Ad-block	Copy-block	Spy-block
Number of IPs in block list	266,019	781,942,220	444,451
Number found in:			
zone files	3,181	208,081	8,681
DMOZ data	195	26,227	578

Table 3: Overview of block lists used.

4.2 Implications

Many IP addresses corresponding to the Web servers contained in the DMOZ data and zone files showed up in each of the block lists. Table 3 shows the number of IPs contained

[2]Spyware is software that surreptitiously monitors user activities and reports them to the attackers.

in Ad-block, Copy-block, and Spy-block block lists that also appeared in the DMOZ data and zone files. A disproportionately high number of blocked IPs were present in the zone files, which represents just `.com` and `.net` TLDs.

Figures 3(a), 3(b), and 3(c) show the distribution of the percent of Web servers blocked on machines hosting the indicated number of Web servers for the ad-block, copy-block, and spy-block block lists respectively. We find that for the zone files data set, about $95.9 - 96.1\%$ of the blocked IPs in various block lists were hosting 100 Web servers or more. The corresponding numbers for the DMOZ data set were $26.81 - 83.01\%$. Clearly, blocking the IPs of these machines will impact the availability of the co-located Web servers as well. Further, for each of our three block lists, the number of Web servers hosted on an IP by themselves in the zone files ranged from 0.61% to 1.04% (5.25% to 29.21% for the DMOZ data set).

5. DNS SERVER CO-LOCATION

Besides being co-hosted themselves, Web servers may be considered to be co-located if the authoritative DNS servers that lead clients around the world to their respective Web servers are co-located. Thus, the co-location of DNS servers has important implications on the availability of Web servers. Here, we look at the extent to which authoritative DNS servers are co-located, both at the IP address and AS granularity.

5.1 Additional Data Used

To infer DNS server co-location, we needed to collect information about authoritative DNS servers for the Web servers contained in our two primary data sets. Fortunately, the zone files already contain information on the authoritative DNS servers for each domain listed. The process was not so straight-forward for the DMOZ data, however. We had to conduct DNS lookups for `NS` records to determine the list of authoritative DNS servers for each of the Web servers contained in the DMOZ data. Further, we resolved the output of each `NS` lookup, which is generally a host name, into IP address using the DNS `A` record lookups.

For both data sets, Table 4 illustrates the unique authoritative DNS servers by name and also the distinct IP addresses these correspond to. It also shows the distinct DNS servers by name and IP address for the combined data set. We combine the data sets before further analyzing them because 74.9% of the DNS servers from the DMOZ data are common to the DNS servers for the zone files. This indicates that Web servers from a variety of different TLDs are hosted on the same authoritative DNS servers.

	DMOZ Data	Zone Files	Combined
DNS Servers	278,169	1,611,145	1,710,847
Unique IPs	223,992	820,547	875,122

Table 4: Authoritative DNS servers for DMOZ data and zone files.

For DNS server co-location analysis based on ASes, we convert DNS server IP addresses to ASes by using a BGP routing table described in Section 3.1.

5.2 Analysis

As shown in Figure 4, most DNS servers are authoritative for only a small number of domains (and hence for the Web servers contained in those domains). Note the log scale on the X-axis. In particular, 30% of them are authoritative for only one domain. The median number of domains a DNS server is authoritative for is 4. However, there are several DNS servers that are authoritative for a very large number of domains. In particular, there are 11 DNS servers in our list which are each authoritative for over $1,000,000$ domains, with the highest being authoritative for $3,757,103$ domains! This raises questions about the availability of Web servers in the event of targeted DoS attacks. We show the results for AS-level analysis in Figure 5. The key results for the AS granularity are similar, with 63.61% of the ASes containing DNS servers that are authoritative for 100 or fewer domains. Also, we find 19 ASes that have authoritative DNS servers for over $1,000,000$ domains, with the highest one hosting $9,544,010$ domains.

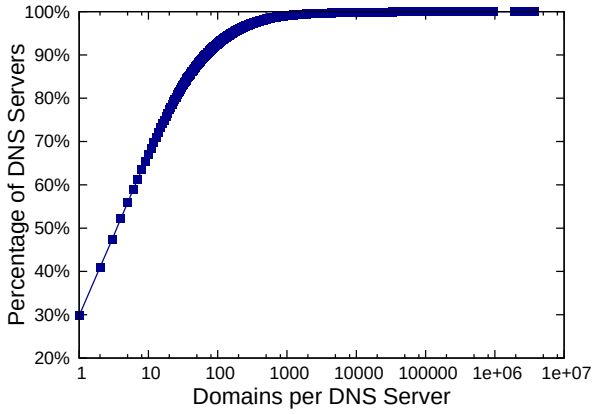


Figure 4: CDFs showing domains per DNS server as a percentage of DNS servers.

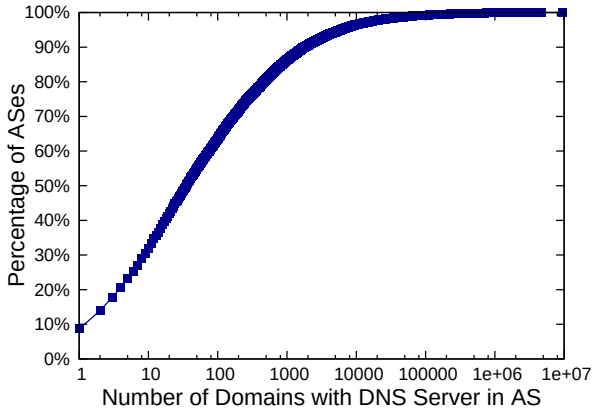


Figure 5: CDF of the ASes with DNS servers authoritative for the indicated number of domains.

While co-location threatens availability in the event of a DoS attack, and due to IP-based blocklisting, another phenomenon tries to balance it. That phenomenon relates to the *redundancy* of authoritative DNS servers, as recommended by [8]. Indeed, when looking at the number of DNS servers corresponding to each domain in the zone files, we find that almost all the domains have at least two DNS servers asso-

ciated with them. Some have many more. In fact, we see a maximum of 13 DNS servers per domain, which incidentally is the maximum number of responses that fit in a DNS response packet. Figure 6 shows the percentage of domains that have a specified number of DNS servers.

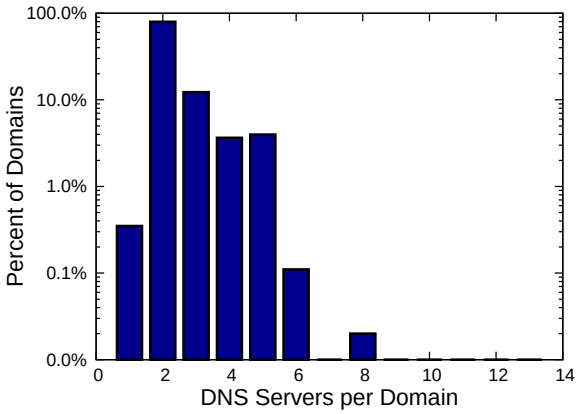


Figure 6: Percentage of domains with the indicated number of DNS servers.

6. RELATED WORK

A number of works have looked at the Web from the perspective of documents that comprise it. In [9], the authors use connectivity measurements to learn about the topology of the Web. Work in [10] examines how search engines should deal with the evolution of the Web. In [11], the authors demonstrate that Web traffic exhibits a high degree of self-similarity, much like wide-area and local area network traffic. In [12], the authors determine that while Web access does not exactly follow a Zipf distribution, simple Zipf-like models are sufficiently accurate for Web proxies. In [13], the authors examine methods for generating representative Web traffic. Work in [14] examines Web traffic using six data sets and suggests performance enhancements for Web servers. Our work differs from these in that we examine the Web from the perspective of the servers that make up the Web.

In [15], the authors perform an extensive analysis on the DNS infrastructure. Their work focuses on the availability of name servers, whereas ours examines the characteristics of the domains themselves.

Work in [16] comes closest to the portion of our work that deals with blocklisting. The author in this work examines the number of Web sites hosted on the same IP address. The motivation for this work was to determine the extent of collateral damage from IP-based filtering. However, because the work was focused on the societal impact of the practice, it does not provide a rigorous discussion of the technical details.

7. DISCUSSION

The analysis in this paper determined that a vast majority of Web servers are co-located with other Web servers. This co-location even extends to the DNS servers, which are used to guide the clients to these Web servers. While we looked at the `.com` and `.net` domains exhaustively, we had only limited information on other TLDs. It would be interesting to conduct a more detailed analysis using zone files for each of the country code TLDs.

Acknowledgments

We would like to thank Daniel Brazzell and Mark Bruhn for providing Netflow captures from the Indiana University campus.

8. REFERENCES

[1] DMOZ, "Open directory project."
[2] VeriSign, Inc., "COM NET Registry TLD Zone Access Program," http://www.verisign.com/information-services/naming-services/com-net-registry/page_001052.html.
[3] VeriSign, "Domain name industry brief," June 2007, http://www.verisign.com/static/042161.pdf.
[4] U. of Oregon Advanced Network Technology Center, "Route Views project," http://\\www.routeviews.org/.
[5] H. Chang, S. Jamin, Z. M. Mao, and W. Willinger, "An empirical approach to modeling inter-AS traffic matrices," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2005.
[6] B. Claise, "Cisco systems NetFlow services export version 9," IETF RFC 3954, Oct. 2004.
[7] Bluetack Internet Security Solutions, "B.I.S.S. Forums," http://www.bluetack.co.uk/forums/index.php.
[8] P. Mockapetris, "Domain names - concepts and facilities," IETF RFC 1034, Nov. 1987.
[9] R. Albert, H. Jeong, and A. Barabasi, "Diameter of the world wide web," *Nature*, vol. 401, pp. 130–131, 1999.
[10] A. Ntoulas, J. Cho, and C. Olston, "What's new on the web?: the evolution of the web from a search engine perspective," in *International World Wide Web Conference*, 2004.
[11] M. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *IEEE Transactions Of Networking*, vol. 5, no. 6, pp. 835–846, 1997.
[12] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *IEEE INFOCOM*, 1999.
[13] P. Barford and M. Crovella, "Generating representative web workloads for network and server performance evaluation," in *ACM SIGMETRICS*, 1998.
[14] M. Arlitt and C. Williamson, "Web server workload characterization: the search for invariants," in *ACM SIGMETRICS*, 1996.
[15] J. Pang, J. Hendricks, A. Akella, R. D. Prisco, B. Maggs, and S. Seshan, "Availability, usage, and deployment characteristics of the domain name system," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2004.
[16] B. Edelman, "Web sites sharing IP addresses: Prevalence and significance," Sept. 2003, http://cyber.law.harvard.edu/people/edelman/ip-sharing/.