Ads versus Regular Contents: Dissecting the Web Hosting Ecosystem

Pradeep Bangera*
PriceFlier, India

Sergey Gorinsky
IMDEA Networks Institute, Spain

***Abstract*—Ads (advertisements) have become a common source of revenues for websites and transformed the web hosting ecosystem. This paper reports an extensive measurement-based study of web hosting with an explicit focus on differences in the hosting of ads versus regular contents. Using a VPN (Virtual Private Network) service and 22,040 open recursive DNS (Domain Name System) resolvers, we collect contents of top country-specific websites in 52 countries around the globe and characterize the hosting infrastructures. While we observe that ads employ more servers for broader load distribution, replication is local for ads and global for regular contents. Our results clearly show that transit ASes (Autonomous Systems), including the tier-1 networks, diversify their economic roles and prominently provide web hosting. A small number of ASes dominate heavily in the byte volume of hosted contents, with the top hosts being different for ads and regular contents. Compared to ASes and organizations, the distribution of hosting countries is even more heavily skewed, with the USA consistently taking the overwhelmingly dominant position. While ads have shorter response times, their download times are longer because websites are developed to serve requests for regular contents with a higher priority.**

***Index Terms*— Web; ad; regular content; hosting; infrastructure; autonomous system; transit hierarchy; byte volume; location; delivery performance.**

I. Introduction

Websites constitute a major source of Internet traffic and deliver their contents to users via a vast network of ASes (Autonomous Systems). Traditionally, while access ASes at the Internet edge hosted websites, a hierarchy of transit ASes in the Internet core specialized in traffic delivery [1]. A transit AS typically raised revenues by charging customers for the bidirectional traffic on the transit links with the customers [2].

The web hosting and delivery ecosystems keep evolving and incorporating new features [3]–[5]. Individual contents increase in their byte volume, and their delivery to the users requires higher throughput and lower latency. Major websites commonly employ CDNs (Content Delivery Networks) that reduce latency by serving the contents from caches near the users [6]. Following the same trend of content replication, some websites host contents in their own globally distributed server infrastructures. To reduce traffic costs, many players interconnect with peering links to exchange their local traffic directly, rather than through their transit providers.

* Part of the work was done when Pradeep Bangera was with IMDEA Networks Institute and Carlos III University of Madrid.

ISBN 978-3-901882-94-4

The massive emergence of online advertisements (ads) powerfully transforms web hosting. Compared to regular contents, ads have a more local scope of interest. By tracking the locations and specific interests of users, advertisers get an opportunity to host relevant ads more locally. By displaying ads, a website can raise substantial revenues. In 2015, the total online ad revenues in the USA reached about $60 billion dollars [7]. On the other hand, transit prices keep falling, and transit ASes struggle to upgrade their delivery infrastructures to accommodate the unrelenting growth of the Internet traffic.

In this paper, we report extensive measurements of web hosting worldwide and examine hosting differences for ads versus regular contents. We use a VPN (Virtual Private Network) service to establish vantage points in 52 countries and collect contents of most popular country-specific websites. We determine which IP (Internet Protocol) addresses, ASes, organizations, and countries host the contents. We also detect the position of the hosting infrastructures within the transit hierarchy, estimate the byte volumes of contents, and evaluate the performance of content delivery. Our quantitative characterization for hosting of ads and regular contents is valuable for realistic modeling of Internet traffic, planning of infrastructure deployments, and understanding of Internet resiliency. Our study also contributes a number of significant qualitative results: (1) Confirming the trend towards diversification of AS economic roles, we show that transit ASes are prominently involved in web hosting; (2) The hosting infrastructures for ads and regular contents are substantially different, e.g., replication is local for ads and global for regular contents; (3) While ads have shorter response times, their download times are longer.

The rest of the paper is organized as follows. Section II details our measurement methodology. Section III analyzes the collected data. Section IV discusses related work. Finally, section V concludes the paper with a summary of its contributions.

II. Methodology

A. Collection of contents

We provide a global perspective on web hosting by using a novel VPN-based approach to collect popular online contents. Specifically, we use the VPN service from *hma.com* [8] that operates servers in 52 countries around the world. For each of these 52 countries, we select the top-50 websites as ranked by Alexa [9]. From our central control node, we connect via the VPN to its server in each of the 52 countries and automatically

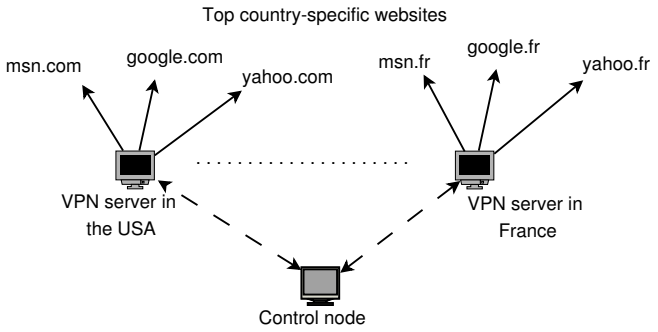

Fig. 1: Global collection of contents from top country-specific websites via the VPN service.

browse the top-50 country-specific websites to collect their ads and regular contents. Figure 1 shows our measurement setup. While simple, our VPN-based methodology for collection of online contents is new and effective in providing a worldwide coverage of top country-specific websites.

After connecting our control node to the remote VPN server, we retrieve all the hyperlinks on the landing page of the targeted website using Lynx (Linux-based text browser). Since content providers embed URLs (Uniform Resource Locators) into scripts on their websites for fetching ads, videos, and other contents, and because Lynx does not execute XML (Extensible Markup Language), JavaScript, and Flash objects, we employ the methodology proposed in [10] to collect the embedded URLs of each website. These hyperlinks are then opened in the Firefox browser configured with the Firebug, FireStarter, and NetExport add-ons. Using the above methodology, we browse the top-50 country-specific websites in 52 countries and collect around 300 GB of HTTP (HyperText Transfer Protocol) and HTTPS (HTTP Secure) header data from 2,165 unique websites in mid-January 2013. The compressed dataset is available online [11].

B. Detection of ad URLs

A typical top website is a complex collection of contents including third-party contents. The third parties usually provide the website with ads, APIs (Application Program Interfaces), widgets, etc. To detect the ad URLs in our dataset, we use the filtering rules from the widely used filtering plugins, such as Adblock Plus and Ghostery. Using these filtering rules, we detect around 7,380 ad URLs in our collected data.

C. Detection of regular-content URLs

It is more complicated to detect regular-content URLs among third-party URLs. Browser plugins do not filter out third-party URLs for regular contents, e.g., URLs for widgets and external websites. An intuitive approach is to detect the regular-content URLs of a website by matching their STLD (Second Top-Level Domain) names with the STLD of the website's primary URL, e.g., URLs *mail.yahoo.com* and *www.yahoo.com* share primary STLD *yahoo.com*. On the other hand, URLs that are native to the website can also have STLD names that cannot be straightforwardly matched with the primary STLD, e.g., URL *us.yimg.com* versus primary STLD *yahoo.com*. We refer to such URLs as alternate URLs. The methodology in [10] relies on authoritative nameservers to detect the alternate URLs of websites. If any URL on a website shares its authoritative nameserver with the website's primary URL, then such URL is flagged as a native regular-content URL.

While the method in [10] works for websites that manage their own authoritative nameservers, this technique fails to distinguish the URLs of websites hosted by third-party hosting providers. For example, third-party hosting provider Amazon hosts websites *www.9gag.com* and *www.rockmelt.com* and also supplies an authoritative name service for them. The detection method based on the authoritative nameservers flags all the URLs of *www.9gag.com* and *www.rockmelt.com* as native to both websites. Besides, *www.9gag.com* uses alternate URLs with the *d*.cloudfront.net* pattern to host static contents from Amazon's Cloudfront CDN service. This makes the task of distinguishing between alternate and third-party URLs on *www.9gag.com* difficult because the cloudfront.net STLD is registered and administered by Amazon. Therefore, even a *whois* query for this domain name does not reveal the host website of such URLs.

We use a combination of techniques to detect regular-content and alternate URLs of each website and separate the third-party URLs:

Step 1: Matching of the STLD. First, we flag a URL as regular-content if the STLD of the URL matches the STLD of the website's primary URL.

Step 2: Verification of the referrer. The URLs remaining after the previous step are filtered out if they are not referred by the website's primary URL according to the Referrer field in the HTTP request headers. This verification removes most of the third-party URLs referred by third-party domains. For example, website *www.dropbox.com* has external links to several third-party websites such as *nytimes.com*; when third-party URLs such as *css.nyt.com* belonging to *nytimes.com* appear in the HTTP data collected for *www.dropbox.com*, the Referrer field of such third-party URLs is *nytimes.com*, and URL *css.nyt.com* is filtered out for website *www.dropbox.com*.

Step 3: Check of the request frequency. While the previous step does not handle the case where the website refers URLs to multiple third-party websites, the native STLDs of the website get high request counts compared to third-party STLDs. The following condition separates all the alternate URLs from the website-referred extraneous URLs. Each STLD i is selected if:

$$y_i \cdot n_i > \sigma \tag{1}$$

where y_i is the total request count of STLD i, n_i is the total number of URLs with STLD i, and σ is the standard deviation for the request counts of all the STLDs on the website. The URLs of the selected STLD i are considered to be regular-content URLs of the website.

Applying the above 3-step methodology, we detect 19,140 regular-content URLs. This completes our collection of ads and regular contents of the websites.

D. From contents to their hosting infrastructures

To determine the IP addresses, ASes, organizations, and countries that host the collected ads and regular contents, we utilize a network of open recursive DNS (Domain Name System) servers to resolve each of the gathered URLs [12]. From the available pool of about 130,000 open recursive DNS servers, we eliminate around 36,500 DNS servers for such reasons as recursion unavailable (23,000), unreachable (12,500) and invalid DNS response (1,080) to get a globally distributed platform of reliable vantage points across the Internet.

E. Elimination of misleading DNS servers

While small in number, the servers with invalid responses are particularly important to filter out because of their potentially large negative impact on the measurement accuracy. We detect DNS servers that inject fake IP addresses during the URL resolution. Two types of invalid responses are observed. The first type, which is also observed in [13] and [14], consists of invalid replies for only particular URLs, such as *www.facebook.com* and *www.youtube.com*. Invalid responses of the second type come from DNS servers that fail to resolve a URL or are unwilling to perform DNS recursions. Among the latter, the invalid IP addresses point to ad-displaying search pages managed by the entities that operate the misleading DNS servers.

To detect misleading servers of the first type, we send a single DNS query to each of the 130,000 DNS servers to resolve the primary *www.facebook.com* URL of Facebook. The IP address returned by a DNS server is then used to launch a reverse DNS query. The reverse DNS request to a valid IP address of the *www.facebook.com* URL should return a name record containing *facebook.com* as the STLD. On the other hand, an invalid IP address yields a name record with an irrelevant STLD. Employing the above method, we detect around 630 misleading DNS servers. Surprisingly, the largest number of such DNS servers are hosted in the USA with 417 servers in 15 different ASes. China stands second with a total of 180 such misleading DNS servers hosted by 32 ASes.

To detect misleading DNS servers of the second type, we send a single DNS query to resolve a non-existent URL, e.g., *pppqqppqqp.com*. An open recursive DNS server is expected to respond to such query with an NXDOMAIN error code. However, a misleading DNS server replies with an IP address pointing to an ad-displaying search page. Using this method, we detect 450 misleading DNS servers hosted across 48 countries and 210 ASes, bringing the total of eliminated misleading DNS servers to 1,080.

F. Resolution of URLs using open recursive DNS servers

The remaining set of 93,500 reliable DNS servers covers 172 countries, 8,500 ASes, and 22,040 prefixes. By selecting a single IP address in each prefix, we converge to 22,040 DNS servers as vantage points. Each ad and regular-content URL is resolved from these open recursive DNS servers to obtain the CNAMEs (Canonical Names) and IP addresses associated with the URL. The DNS resolutions of 7,380 ad and 19,140 regular-content URLs yield 90,000 and 102,600 IP addresses with the average of 12 and 5 per-URL IP addresses respectively.

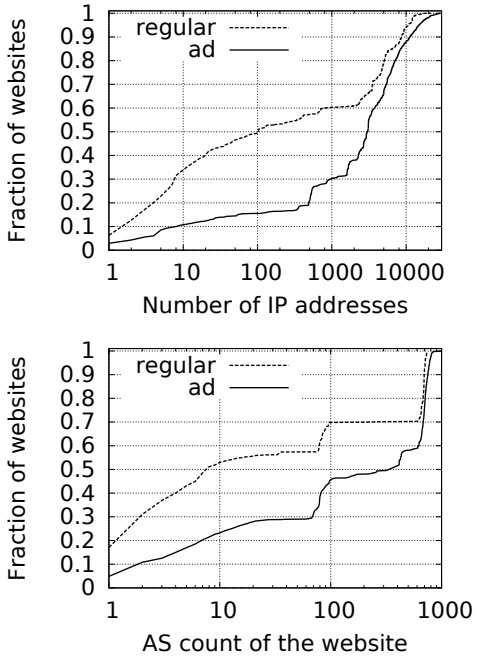

Fig. 2: Per-website numbers of IP addresses and ASes hosting the websites.

G. Mapping of IP addresses to hosting infrastructures

Finally, we map the obtained IP addresses to their respective ASes, organizations, and geographic locations. We utilize the IP-to-AS mapping service [15] to assign each IP address to its hosting AS, prefix, registry, and organization. To map the IP address to its hosting country, we use the GeoIP tool from MaxMind [16] which is largely accurate on the country level compared to its city-level and finer resolutions. When feasible, we also map each URL of the website to its CNAME, hosting AS, IP address, prefix, country, registry, and organization to form a network-level record of the URL. An URL with multiple IP addresses is mapped to multiple records. At the end, the 7,380 ad and 19,140 regular-content URLs are mapped to 2,272 and 2,177 ASes in 134 and 115 countries respectively. The dataset with the resolved URLs and their mapping is available online [17].

III. Measurement results

A. Hosting infrastructures

IP addresses and ASes. Figure 2 presents the numbers of IP addresses and ASes hosting the websites. The median of the websites use around 3,000 IP addresses per website to serve ads versus only 100 IP addresses per website for regular

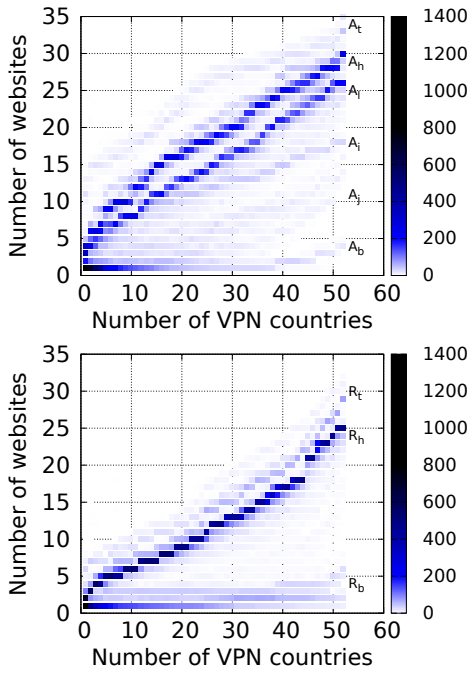

Fig. 3: Superimposition of the individual AS profiles: (top) ads and (bottom) regular contents.

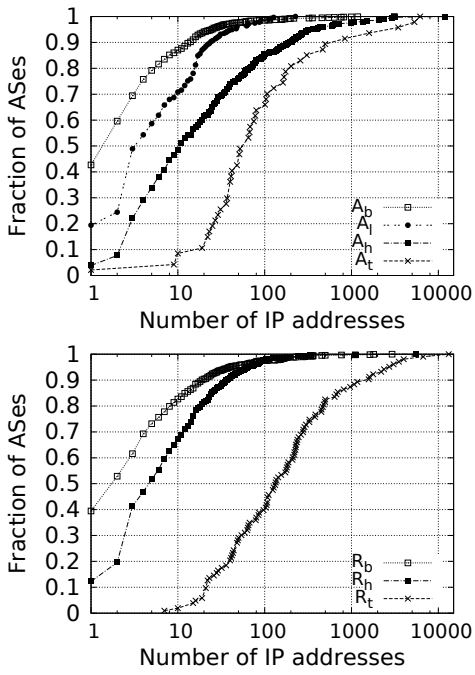

Fig. 4: IP addresses for: (top) ad-hosting AS clusters and (bottom) regular-hosting AS clusters.

contents. Similarly, 70% of the websites employ around 700 ASes to host ads versus 100 ASes for regular contents. Thus, ads use more IP addresses and ASes per website than regular contents, suggesting that *ads employ more servers for broader load distribution.*

The distributions of the hosting ASes in figure 2 reveal series of step increases. Both ads and regular contents exhibit step increases in ranges 70-100 ASes and 600-800 ASes. Ads have another step increase around 400 ASes. Each of these increases represents websites hosted by CDNs, with different steps corresponding to different geographic coverage by the CDNs.

Clusters of hosting ASes. To examine the hosting ASes closer, we profile every AS with the number of websites hosted by the AS in each of the 52 VPN-reached countries and arrange the 52-country sequence in the increasing order of the hosted websites. Figure 3 superimposes the individual AS profiles, separately for the 2,272 ad-hosting ASes and 2,177 regular-hosting ASes, in a graph with tiles of different shade intensity. When individual AS profiles overlap, i.e., have the same number of hosted websites in the same position in their country sequences, the shade intensity of the respective tile denotes how many individual AS profiles overlap in this tile. The shade intensity rises from white to black as the number of the profile overlaps increases.

Figure 3 reveals 2 distinct clusters and 4 faint clusters A of ad-hosting ASes and only 3 clusters R of regular-hosting ASes. Subscripts t, h, l, and b refer to the top, high, low, and bottom clusters respectively. An AS in cluster A_t or R_t typically hosts many websites. Figure 4 shows that 30% of the ASes in cluster A_t and 60% of the ASes in cluster R_t employ more than 100 IP addresses each. On the other side of the spectrum, an AS in cluster A_b or R_b usually hosts few websites. Around 40% of the ASes in clusters A_b and R_b use a single IP address. The 4 significant ad-hosting AS clusters A_t, A_h, A_l, A_b and 3 regular-hosting AS clusters R_t, R_h, R_b comprise of 47, 362, 278, 1,456 ASes and 103, 759, 1,252 ASes with the average per-AS per-country number of hosted websites equal to 24, 18, 14, 4 and 15, 11, 2 respectively. Around 86% and 96% of the ASes in clusters A_h and A_l are also present in cluster R_h. Hence, while ad-hosting ASes form more clusters, the clusters of regular-hosting ASes are larger in size, indicating that *replication is local for ads and global for regular contents.*

AS clusters versus the transit hierarchy. To quantify the involvement of transit ASes in hosting, we use the CAIDA (Center for Applied Internet Data Analysis) dataset that ranks each AS according to its *customer-cone size*, i.e., the number of its direct and indirect transit customers [18]. In our classification, *core ASes* are the tier-1 networks, which usually have huge customer cones exceeding 20,000 ASes [1]. *Edge ASes* are those, mostly access, networks with the customer-cone size smaller than 5 ASes [19]. All the other ASes are called *intermediate ASes* and largely comprise tier-2 and tier-3 transit networks.

Figure 5 plots the IP addresses in the AS clusters arranged according to the AS customer-cone size. The fraction of IP addresses used by the edge ASes is the highest in clusters A_b, R_b and lowest in clusters A_t, R_t. In contrast, the fractions of IP addresses used by the core and intermediate ASes are

AS number	Organization	Regular IP addresses	Ad IP addresses	Customer-cone size	Service type (hierarchy position)
20940	Akamai	13,182	11,832	5	CDN (edge)
4436	GTT (nLayer)	3,513	3,196	975	Transit (intermediate)
209	CenturyLink	3,054	2,922	20,294	Transit (core)
3257	GTT (Tinet)	2,725	2,340	24,729	Transit (core)
7922	Comcast	2,368	2,202	2,164	Access (intermediate)
1299	TeliaSonera	2,196	2,054	25,753	Transit (core)
7843	Time Warner	1,603	1,500	932	Access (intermediate)
1273	Vodafone	1,577	1,346	13,780	Transit (intermediate)
1239	Sprint	1,167	1,070	22,042	Transit (core)
5511	Orange	1,073	989	4,431	Transit (intermediate)

TABLE I: Top-10 ASes with the highest numbers of IP addresses in clusters A_h and R_t.

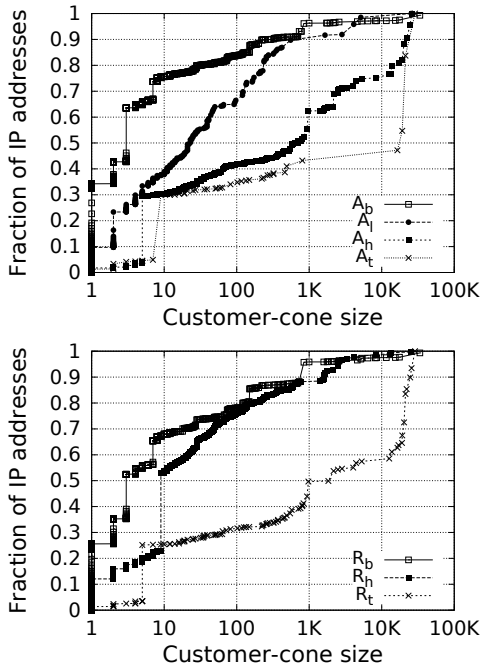

Fig. 5: IP addresses in AS clusters arranged according to the AS customer-cone size: (top) ads and (bottom) regular contents.

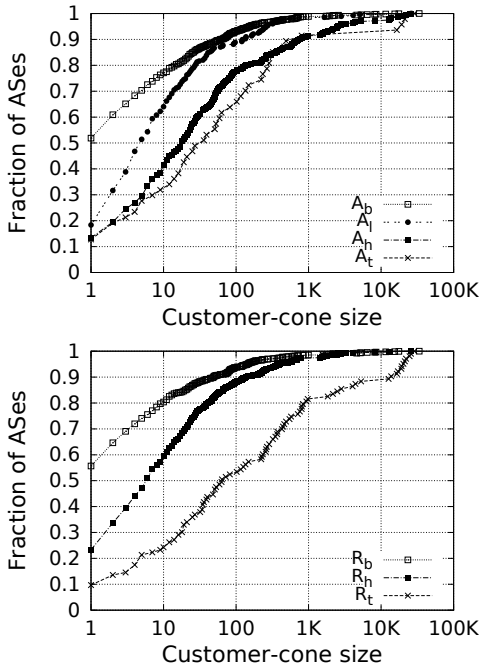

Fig. 6: Customer-cone sizes in AS clusters: (top) ads and (bottom) regular contents.

the highest in clusters A_t, R_t and lowest in clusters A_b, R_b. Thus, *the ASes that typically host more websites and employ more IP addresses tend to lie closer to the core of the transit hierarchy.*

With respect to the total number of IP addresses, the champions are cluster A_h for ads and cluster R_t for regular contents. Their numbers of IP addresses are around 47,000 and 61,000 respectively. The 2 clusters share 50 prominent ASes that collectively account for around 40,000 IP addresses. Among these 50 heavyweights, the fractions of core, intermediate, and edge ASes are 10%, 80%, and 10% respectively. Zooming in further on these ASes, table I presents the top 10 ASes with the highest numbers of IP addresses and reports the customer-cone size, service type, and hierarchy position of the ASes. The top-10 set consists of 7 transit networks (including 4 core and 3 intermediate ASes), 2 access networks (which are both intermediate ASes), and 1 CDN (which is an edge AS).

Looking at other clusters, we observe that ad-hosting cluster A_t has about 21,000 IP addresses and that its 5 most prominent ASes (owned by Google and 4 large transit providers NTT, Level 3, Deutsche Telekom, and Telecom Italia) account for 81% of its total IP addresses. Clusters A_b and R_b, formed predominantly by edge ASes, have 18,000 and 21,000 IP addresses respectively. In both clusters, the top-2 ASes with the largest numbers of IP addresses are AS 16509 of Amazon and AS 8075 of Microsoft.

Figure 6 plots the customer-cone sizes in the ad-hosting and regular-hosting AS clusters. The ASes host ads and regular contents across the transit hierarchy. While the fraction of edge ASes within a cluster grows from 18% through 44% to 72% in clusters R_t, R_h, R_b respectively, this fraction

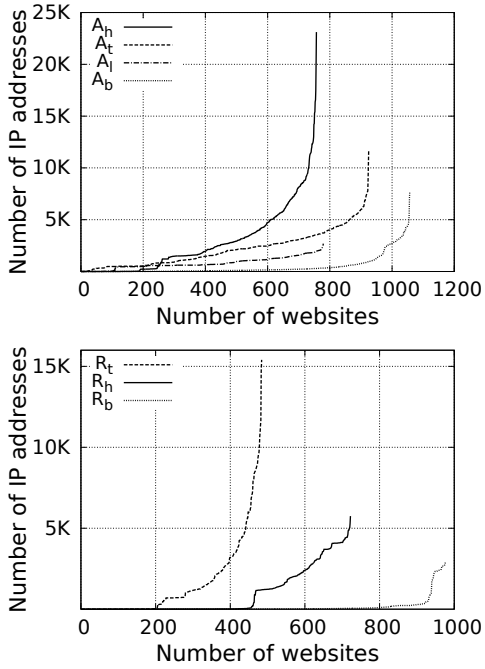

Fig. 7: Websites and their IP addresses: (top) ads and (bottom) regular contents.

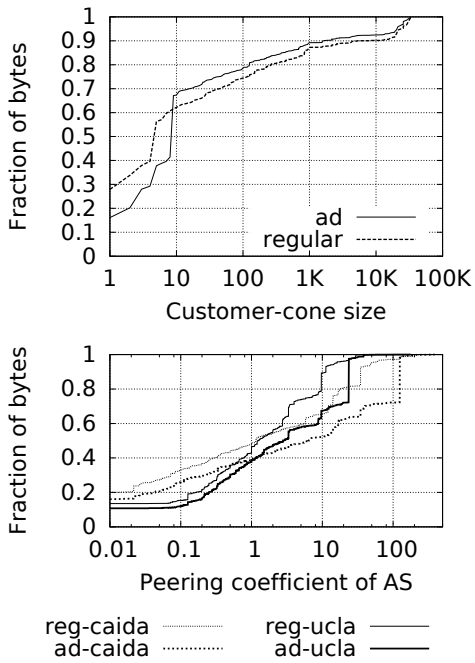

Fig. 8: Byte volumes for ads and regular contents hosted by ASes arranged according to: (top) customer-cone size and (bottom) peering coefficient.

has a qualitatively similar rise across clusters A_h, A_l, A_b. Complementing the edge ASes, the core and intermediate ASes follow the opposite declining trend: while core and intermediate ASes respectively comprise 7% and 75% of the ASes in cluster R_t, the corresponding numbers decline to 0.3% and 27.7% for cluster R_b. Hence, *the AS clusters inherit their positions in the transit hierarchy from their individual ASes.*

Websites. Figure 7 depicts the number of IP addresses of the hosted websites arranged in the increasing order of their IP addresses. In regard to this metric, *www.yahoo.com* and *www.msn.com* are the top-2 websites for each of the 4 ad-hosting clusters. While *www.weather.com* and *www.msn.com* are the top-2 websites for clusters R_t and R_h, *www.bing.com* and *www.lemonde.fr* take the top-2 positions for cluster R_b. The clusters with larger fractions of core and intermediate ASes, such as clusters A_h and R_t, use more IP addresses per website. In contrast, the dependence on the cluster's transit-hierarchy position is reverse for the number of hosted websites: while clusters A_h, A_l and A_b host 757, 780, and 1,057 websites respectively, clusters R_t, R_h, and R_b host 484, 722, and 977 websites respectively. Cluster A_t, which hosts 925 websites, is the only deviation from this trend. Therefore, *while the ASes that lie closer to the core of the transit hierarchy tend to host websites with more IP addresses, the total number of websites hosted by such ASes is lower.*

Overall, the results of this section reveal that transit ASes, including the tier-1 networks, have become heavily involved in web hosting. This confirms the trend towards diversification of economic roles played by an AS.

B. Byte volume and location of contents

In this section, we go beyond the IP addresses and estimate the volume of contents hosted by individual ASes. To do the estimate, we measure the total byte volume of the contents retrieved from each URL and uniformly split this content volume among all the IP addresses of the URL. Then, we aggregate the content volumes associated with all the IP addresses of each AS to determine the total content volume hosted by this AS.

Content location in the transit hierarchy. The top graph in figure 8 presents the byte volumes for ads and regular contents hosted by ASes arranged according to the customer-cone size. While constituting 55% and 59% of the ad-hosting and regular-hosting ASes respectively, the edge ASes host about 33% and 40% of the ad and regular-content volumes. The core ASes host around 4% and 6% of the total ad and regular-content volumes respectively. These are significant shares because the 12 core ASes comprise only about 0.5% of the hosting ASes for both content types. The intermediate ASes host the remaining 63% and 54% of the ad and regular-content volumes respectively. Both for ads and regular contents, the per-AS content volume is the lowest for the edge ASes and highest for the core ASes. Hence, *transit ASes host substantial byte volumes of ads and regular contents*.

Content volumes and peering coefficients. The *peering coefficient* of an AS is the ratio of the number of its peering links to the number of its transit links. This metric reflects the role played by the AS in the Internet economy. Because

Organization	AS number	Content volume (%)
Google	15169	24.0
EdgeCast	15133	4.2
Akamai	20940	2.8
Amazon	16509	2.7
Level 3	3356	1.3
NTT	2914	1.2
NetVision	1680	1.2
UAB Hostex	47205	1.2
Telia	1299	1.2
fibre one	24961	1.0

Organization	AS number	Content volume (%)
EdgeCast	15133	10.5
Akamai	20940	5.2
Level 3	3356	3.3
Wikimedia	43821	3.3
Microsoft	8075	1.7
NTT	2914	1.6
China Telecom	4134	1.5
CDNetworks	36408	1.3
Amazon	16509	1.3
China Unicom	4837	1.2

TABLE II: Top-10 ASes for the volume of hosted contents: (top) ads and (bottom) regular contents.

CDN	Content volume (%)	ASes	Websites (%)
Google	29.0	448	61.5
Akamai	21.5	730	49.3
EdgeCast	4.2	1	16.3
Amazon	3.2	26	26.1
Level 3	1.4	11	15.6

CDN	Content volume (%)	ASes	Websites (%)
Akamai	27.5	730	34.2
EdgeCast	10.5	1	8.1
Level 3	3.4	15	10.8
Microsoft	2.4	11	4.2
Amazon	1.5	23	13.0

TABLE III: Top-5 CDNs for the volume of hosted contents: (top) ads and (bottom) regular contents.

a large transit AS typically has many transit customers and relatively few peers, the peering coefficients of large transit ASes are usually below 1. The peering coefficient of a small AS is typically above 1 because small ASes peer extensively to reduce their transit costs [20], [21]. To compute the peering coefficients of the hosting ASes in our measurements, we use the CAIDA and UCLA inter-AS relationship datasets which are similar in their numbers of transit links and greatly differ in their number of peering links. The bottom graph in figure 8 shows the byte volumes for ads and regular contents hosted by ASes arranged according to the peering coefficient. For both CAIDA and UCLA datasets, the ASes with peering coefficients below 1 host around 40% of the total content volumes. This alternative economic perspective confirms that *transit ASes host significant volumes of online contents.*

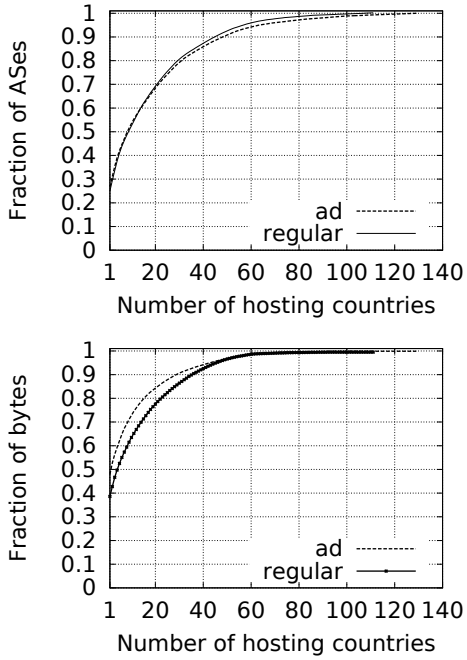

Fig. 9: Hosting ASes and hosted content volumes arranged according to the hosting country.

Top-volume ASes and CDNs. Table II presents the top-10 ASes in regard to the volume of hosted contents. These top-10 ASes host around 41% and 31% of the total byte volume for ads and regular contents respectively. AS 15169 operated by Google is the top host for ads, with its 24% share of the total ad volume in our measurements. Hosting 10.5% of the total regular-content volume, AS 15133 operated by EdgeCast is the top-volume AS for regular contents and predominantly hosts such websites as Pinterest, Twitter, and WordPress.

Table III reports the 5 top-volume CDNs as organizations. Again, Google takes the top spot for ads by hosting 29% of the total ad volume and 61% of the websites. Google hosts 24% of the total ad volume in its primary AS 15169 and its remaining 5% in more than 400 third-party ASes. Akamai is the top-volume CDN for regular contents. It hosts 27% of the total regular-content volume and 34% of the websites. Akamai hosts 5% of the total regular-content volume in its flagship AS 20940 and its other 22% in more than 700 third-party ASes. EdgeCast is the second top-volume CDN for regular contents. Unlike Akamai, EdgeCast hosts all its contents in its single flagship AS 15133 which is broadly present throughout the Internet. Thus, *either on the AS or organization level, a small number of big players dominate heavily in the byte volume of hosted contents.* Our results also show that *the top-volume hosts are different for ads and regular contents.*

Geographic location of contents. Figure 9 plots the hosting ASes and hosted content volumes arranged according to the hosting country. For both ads and regular contents, the top-10 countries ranked by the AS count amass about 54% of all hosting ASes. The top-10 hosting countries ranked by the

Ranking by the AS count		Ranking by the content volume	
Country	AS count	Country	Content volume (%)
USA	558	USA	47.8
Germany	110	Ireland	4.8
UK	108	Germany	3.6
Russia	93	UK	2.9
Japan	76	Denmark	2.8
France	66	Russia	2.7
Netherlands	61	Israel	1.9
Canada	61	Lithuania	1.9
Poland	52	Japan	1.7
Australia	51	Austria	1.7

Ranking by the AS count		Ranking by the content volume	
Country	AS count	Country	Content volume (%)
USA	511	USA	38.6
UK	109	Netherlands	4.1
Germany	102	China	3.9
Russia	96	Russia	3.0
Japan	84	France	2.8
France	71	Germany	2.2
Canada	59	Estonia	2.1
Australia	54	UK	2.0
Singapore	49	Sweden	1.9
Poland	49	Poland	1.7

TABLE IV: Top-10 hosting countries ranked by the number of hosting ASes and volume of hosted contents: (top) ads and (bottom) regular contents.

content volume account for around 72% and 62% of the total ad and regular-content volumes respectively. Table IV shows that the two ranking criteria result in substantially different top-10 lists: some countries with relatively few hosting ASes reach a top-10 rank for the content volume because of hosting popular content providers such as Google and Yahoo. Nevertheless, the USA easily remains the top hosting country regardless of the ranking criterion or content type: *compared to ASes and organizations, the distribution of hosting countries is even more heavily skewed, with the USA consistently taking the overwhelmingly dominant position.*

C. Performance of content delivery

For both user and content provider, it is relevant how quickly a requested content reaches the user. Hence, we measure how long it takes to retrieve the contents from the websites to our VPN servers. A webpage typically contains multiple contents. Its individual contents might arrive to the webpage requester at different times. The download time of a content is the time between requesting its webpage and arrival of the content. We also measure the response times of individual contents. The response time of a content is the time between requesting this particular content and its arrival.

Response times. The top graph in figure 10 shows that the response times for ads and regular contents remain somewhat similar up to around 100 ms and then diverge. The 90th percentile of the response times for ads and regular contents are 1.6 and 3.5 seconds respectively.

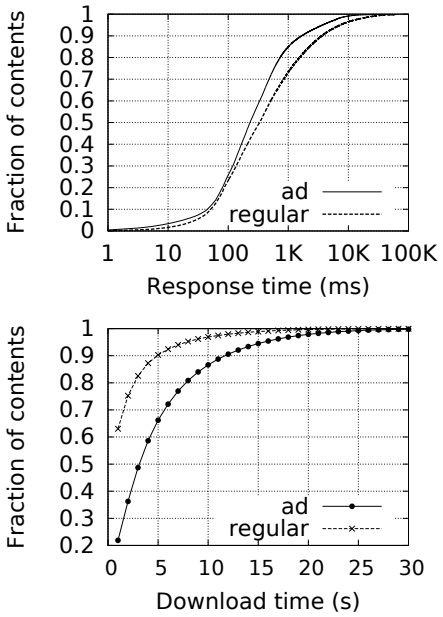

Fig. 10: Delivery performance for ads and regular contents: (top) response times and (bottom) download times.

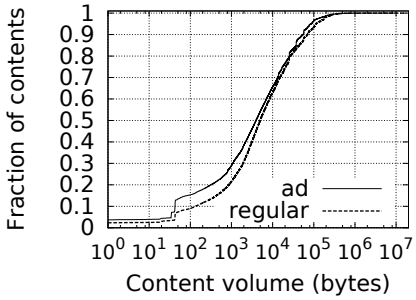

Fig. 11: Byte volumes of ads and regular contents.

Download times. On the other hand, the bottom graph in figure 10 reveals that the 90th percentile of the download times for ads and regular contents are 12 and 4 seconds respectively, suggesting that a majority of the ads arrive long after the regular contents of the webpage. Hence, *ads have shorter response times and longer download times than regular contents.*

While the above result might seem surprising, we study it deeper. By plotting the byte volumes of ads and regular contents in figure 11, we observe that ads have smaller volumes. Besides, we earlier observed that ads employ more servers for broader load distribution. These two observations are consistent with ads having shorter response times. To understand why ads have longer download times, we examine website designs and determine that *websites are developed to serve requests for regular contents with a higher priority.*

IV. Related work

Online contents were extensively researched before. [10], [22]–[24] examined web-based contents and services along with their features such as content type, byte volume, number of requests and servers. [12], [25], [26] carried out measurement-based studies of real operational CDNs, with a common focus on few selected CDNs. [27], [28] explored the footprints of hosting infrastructures across the Internet and geography via real measurements relying on either volunteers [27] or traffic traces [28]. Recently, few works explored content delivery performance for ad traffic. [29] analyzed delivery of ads with a focus on effectiveness of ad blocking. [30] charted 3 prominent ad networks and evaluated their latency and effectiveness of user targeting. [31] characterized mobile ad traffic using data collected in an operational network to study the traffic frequency, content type, and energy implication for mobile devices. That work also briefly glimpsed into the hosting infrastructures of ads.

Our work distinguishes itself from the prior efforts by using the novel VPN-based approach to collect popular website contents worldwide and characterize their hosting infrastructures from a large number of geographically distributed DNS servers. Another distinguishing trait of our study is its explicit comparison for the hosting of ads versus regular contents.

V. Conclusion

This paper presented a global perspective on web hosting. We used a VPN service to collect contents from the top 2,165 websites in 52 countries and characterized the content-hosting infrastructures. Our results revealed not only some similarities but also striking differences in the hosting of ads versus regular contents. Whereas ads employ more servers for broader load distribution, replication is local for ads and global for regular contents. In general, transit ASes – including the tier-1 networks – are prominently involved in web hosting. This confirms the trend towards increasing the number of roles an AS plays in the Internet ecosystem. Either on the AS or organization level, a small number of big players dominate heavily in the byte volume of hosted contents, with the top hosts being different for ads and regular contents. Compared to ASes and organizations, the distribution of hosting countries is even more heavily skewed. While ads have shorter response times, their download times are longer because websites are developed to serve requests for regular contents with a higher priority.

VI. Acknowledgments

This research was financially supported in part by the Regional Government of Madrid (S2013/ICE-2894, Cloud4BigData) and Spanish Ministry of Science and Innovation (TEC2014-55713-R, HyperAdapt).

References

[1] S. Hasan and S. Gorinsky, "Obscure Giants: Detecting the Provider-Free ASes," *Networking 2012.*

[2] P. Bangera and S. Gorinsky, "Traffic Attraction by Internet Transit Providers: An Economic Perspective," *Networking 2014*.

[3] I. Castro, R. Stanojevic, and S. Gorinsky, "Using Tuangou to Reduce IP Transit Costs," *IEEE/ACM Transactions on Networking*, October 2014.

[4] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. V. Vazirani, "How Many Tiers? Pricing in the Internet Transit Market," *SIGCOMM 2011*.

[5] I. Castro, J. C. Cardona, S. Gorinsky, and P. Francois, "Remote Peering: More Peering without Internet Flattening," *CoNext 2014*.

[6] S. Hasan, S. Gorinsky, C. Dovrolis, and R. K. Sitaraman, "Trade-offs in Optimizing the Cache Deployments of CDNs," *INFOCOM 2014.*

[7] PricewaterhouseCoopers, "IAB Internet Advertising Revenue Report, 2015 Full Year Results," http://www.iab.com/wp-content/uploads/2016/04/IAB-Internet-Advertising-Revenue-Report-FY-2015.pdf, April 2016.

[8] Privax Ltd., "HMA! Pro VPN," http://www.hidemyass.com/vpn/.

[9] Alexa Internet, Inc., "The Top Sites on the Web by Country," http://www.alexa.com/topsites/countries.

[10] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding Website Complexity: Measurements, Metrics, and Implications," *IMC 2011.*

[11] P. Bangera, "HTTP(S) Traffic Data," https://doi.org/10.5281/zenodo.556234, April 2017.

[12] C. Huang, A. Wang, J. Li, and K. W. Ross, "Measuring and Evaluating Large-scale CDNs," *IMC 2008.*

[13] L. Yuan, C.-C. Chen, P. Mohapatra, C.-N. Chuah, and K. Kant, "A Proxy View of Quality of Domain Name Service, Poisoning Attacks and Survival Strategies," *ACM Transaction on Internet Technology*, May 2013.

[14] Anonymous, "The Collateral Damage of Internet Censorship by DNS Injection," *ACM SIGCOMM Computer Communication Review*, July 2012.

[15] Team Cymru, "IP to ASN Mapping," http://www.team-cymru.org/Services/ip-to-asn.html.

[16] MaxMind, Inc., "GeoIP2 Precision: Country," https://www.maxmind.com/en/geoip2-precision-country-service.

[17] P. Bangera, "DNS Resolved Hostnames with Organization and Geographical Mapping," https://doi.org/10.5281/zenodo.555948, April 2017.

[18] X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, k. claffy, and G. Riley, "AS Relationships: Inference and Validation," *ACM SIGCOMM Computer Communication Review*, January 2007.

[19] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, "The (in)Completeness of the Observed Internet AS-level Structure," *IEEE/ACM Transactions on Networking*, February 2010.

[20] A. Dhamdhere and C. Dovrolis, "The Internet is Flat: Modeling the Transition from a Transit Hierarchy to a Peering Mesh," *CoNext 2010*.

[21] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles or Contrived Collapse?" *PAM 2008.*

[22] P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson, "Characterizing Organizational Use of Web-Based Services: Methodology, Challenges, Observations, and Insights," *ACM Transaction on the Web*, October 2011.

[23] H. Khandelwal, F. Hao, S. Mukherjee, R. Kompella, and T. Lakshman, "CobWeb: In-network Cobbling of Web Traffic," *Networking 2012.*

[24] S. Ihm and V. S. Pai, "Towards Understanding Modern Web Traffic," *IMC 2011.*

[25] A.-J. Su, D. R. Choffnes, A. Kuzmanovic, and F. Bustamante, "Drafting Behind Akamai: Inferring Network Conditions Based on CDN Redirections," *IEEE/ACM Transactions on Networking*, December 2009.

[26] S. Triukose, Z. Wen, and M. Rabinovich, "Measuring a Commercial Content Delivery Network," *WWW 2011.*

[27] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig, "Web Content Cartography," *IMC 2011.*

[28] 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," *IMC 2012.*

[29] B. Krishnamurthy and C. E. Wills, "Cat and Mouse: Content Delivery Tradeoffs in Web Access," *WWW 2006*.

[30] Y. Wang, D. Burgener, A. Kuzmanovic, and G. Macia-Fernandez, "Understanding the Network and User-Targeting Properties of Web Advertising Networks," *ICDCS 2011.*

[31] N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft, "Breaking for Commercials: Characterizing Mobile Advertising," *IMC 2012.*