Crawling to the Top: An Empirical Evaluation of Top List Use

Qinge Xie[0009−0007−8481−7649] and Frank Li[0000−0003−2242−048X]

Georgia Institute of Technology, Atlanta, GA, USA
{qxie47,frankli}@gatech.edu

Abstract. Domain top lists, such as Alexa, Umbrella, and Majestic, are key datasets widely used by the networking and security research communities. Industry and attackers have also been documented as using top lists for various purposes. However, beyond these scattered documented cases, who actually uses these top lists and how are they used? Currently, the Internet measurement community lacks a deep understanding of real-world top list use and the dependencies on these datasets (especially in light of Alexa's retirement).
In this study, we seek to fill in this gap by conducting controlled experiments with test domains in different ranking ranges of popular top lists, monitoring how network traffic differs for test domains in the top lists compared to baseline control domains. By analyzing the DNS resolutions made to domain authoritative name servers, HTTP requests to websites hosted on the domains, and messages sent to email addresses associated with the websites, we evaluate how domain traffic changes once placed in top lists, the characteristics of those visiting the domain, and the behavioral patterns of these visitors. Ultimately, our analysis sheds light on how these top lists are used in practice and their value to the networking and security community.

1 Introduction

For well over a decade, domain top lists, such as those by Alexa [40], Cisco Umbrella [42], and Majestic [55], have provided a set of purportedly popular or commonly used domains. Anecdotally, these top lists have been critical resources, widely used across both academia and industry. Dozens of prior academic studies [61,67] have used top lists as sources of interesting domains to crawl and evaluate. On the industry side, some security analysis tools and services (e.g., DNSthingy [15] and Quad9 [29]) have incorporated existing top lists into their security offerings. In addition, attackers have manipulated top lists for attracting traffic to their sites [30,68] and offer top list manipulation as a paid service [2,3], suggesting that top list placement provides non-trivial value. However, beyond these scattered documented cases, the research community currently lacks a deeper understanding of how top lists are used in practice and dependencies on these datasets. This understanding is crucial for offering insights to key stakeholders, such as illuminating top list design considerations for list providers,

improving top list usage by researchers and security tools, and informing domain owners on the impact of being ranked. This limitation has become a more pressing issue of late, as Alexa, one of the most popular top lists in academic research [61,67], has been retired, and it is unclear yet what the consequences of this change will be and what potential alternatives are suitable.

In this paper, we take a step in closing this gap by providing empirical grounding on who actually crawls the domains in top lists, and their behavior when visiting these domains. We measure the usage of top lists, which includes both automated (given the large number of domain names included in a top list, practical usage often involves automated, large-scale crawling) and manual/human-driven uses. We seek to answer three primary research questions.

1. How does top list ranking affect the volume of domain visitors?
2. Who visits a domain once placed in a top list?
3. What are the behaviors of visitors to top list domains?

To answer these questions, we conduct controlled top list experiments comparing test domains that are manipulated into different ranking ranges of top lists commonly used by prior work (Alexa before its retirement, Umbrella, and Majestic) with control baseline domains that remain unlisted. We monitor DNS resolutions for these domains as well as requests to websites hosted on these domains, and messages sent to email addresses associated with those sites (on the webpage, in WHOIS, and in `security.txt`). This data affords analysis of the impact of top list placement, and characterization of the domain visitors and their visit behavior. To evaluate whether different types of domain names may result in different behavior, we experiment with two categories of domains: those with realistic names and those that are long and randomly generated.

From our experiment, we find that placement in a top list does drive consistent DNS and web traffic to domains (including suspicious traffic), although primarily once a domain is in the top 100K. We also find that Alexa domains attract more traffic compared to the other top lists, highlighting the need for alternative top list options given that Alexa is now retired. Furthermore, the scale of traffic observed suggests that academic research accounts for a limited portion of top list use in practice. Once a domain falls out of the top list, traffic quickly returns to pre-listing levels, indicating that most uses of top lists rely on the most recent rankings. By analyzing the ASNs and IP geolocations of visitors to our test websites once ranked, we observe extensive use of cloud infrastructure by visitors, various organizations in both industry and academia, as well as visitors predominantly from the US, the Netherlands, China, and Russia. The HTTP user agent headers of these visitors identify various crawlers of web and security companies, and the resources requests hint at many of their purposes, including for RSS feed aggregation, advertising, potential censorship, and security evaluation.

Ultimately, our study provides a systematic characterization of how top lists are used in practice, shedding light on their value to the networking, web, and security community. Moving forward, our findings can inform the design and deployment of top lists to better support their uses.

2 Background

Here, we provide background on the domain top lists investigated in this study, and summarize the related work.

2.1 Domain Top Lists

In this work, we investigate the use of three public domain top lists that have been widely used in prior research [61,67,72]: Alexa, Umbrella, and Majestic.

Alexa. The Alexa's Top 1 Million Sites has been one of the most widely used domain top lists [61,67,72], and ranks domains based on web traffic telemetry collected from user installs of Alexa's browser extension, as well as participating websites that subscribe to Alexa's Certify service [72]. Alexa's ranking is constructed on daily telemetry snapshots, and ranks second-level domains (SLDs) only, rather than fully qualified domain names (FQDNs).

On Dec. 8, 2021, Alexa suddenly announced the retirement of its top list as of May 1, 2022 [39]. On May 1, it retired its web portal but the URL endpoint [40] for downloading its full CSV list remains active and updating until 2023. As of February 1, 2024, the URL endpoint has become unaccessible. Despite its retirement, we include Alexa in our study as its results can still shed light on how top lists are being used in practice.

Umbrella. The Cisco Umbrella Top Million [42] is another popular top list, whose ranking is constructed using passive DNS (PDNS) requests observed across Cisco's Umbrella global network (including OpenDNS [41], PhishTank [58], etc.). Umbrella ranks FQDNs by computing a score for each domain on two-day windows, considering the number of different IP addresses issuing DNS lookups for the domain compared to others [69,72]. Unlike Alexa, Umbrella's traffic data telemetry also accounts for non-web traffic.

Majestic. The Majestic Million [32] is a third domain top list that has been used in several prior studies. Majestic regularly crawls websites and uses the URLs visited within the last 120 days to produce a daily ranking based on a site's backlinks. Specifically, Majestic ranks a site based on the number of referring IPv4 /24 subnets hosting other sites that link to it [54,55]. Thus, unlike the prior two lists, Majestic does not consider web traffic to a domain, but rather the amount of backlinks to it. Majestic's list comprises mostly of SLDs, but ranks FQDNs for certain very popular sites [61].

Other Lists. In this study, we do not focus on other top lists that are paid or otherwise restrict usage. For example, the Quantcast Top Million [67] ranks the most visited domains in the United States, but has not been available since April 2020 [72]. Meanwhile, the Chrome User Experience Report (CrUX) [12] does not provide daily fine-grained rankings, instead providing domains in ranking buckets (e.g., top 10K) in monthly releases. There have also been several more recent top lists released, such as SecRank [72], Cloudflare's Radar Ranking [13] and the Farsight Ranking [22]. We did not include these lists in our study as they were released after our experiments and have not yet been widely used by prior work. We also do not explicitly investigate the Tranco top list [61], as at the time of

our experiments, it was constructed by aggregating the rankings of the three top lists we study and thus cannot be disentangled from the input lists (discussed further in Section 3.6). As of February 1, 2024, the Tranco list also includes data from Google CrUX, the Cloudflare Radar rankings, and the Farsight rankings, and excludes Alexa (which has been retired).

2.2 Related Work

Empirical Investigation of Top Lists. Although used for years in both academia and industry, top lists received little empirical evaluation until late 2018 when Scheitle et al. [67] analyzed the structure, stability, and significance of Alexa, Umbrella, and Majestic. Around the same time, Le Pochat et al. [61] identified various ways for an adversary to manipulate top lists, and created Tranco to account for existing top list shortcomings by aggregating those lists. Rweyemamu et al. [65,66] also identified ways to manipulate Alexa and Umbrella, and investigated the two lists' alphabetically ordering and weekend effects. In 2022, Xie et al. [72] proposed a new top list, SecRank, which serves as an open and transparent ranking method for the research community. Later, Ruth et al. [64] evaluated the relative accuracy of different top lists using popularity metrics derived from a Cloudflare dataset. However, to date, there has not been a systematic investigation of how top lists are actually used in practice, especially beyond surveys of prior academic works, which is the focus of our study.

Internet Service Measurements. Beyond top lists, prior work has evaluated various types of Internet services [37,50,53,71,73,74]. For example, Vallina et al. [70] compared popular domain classification services. Gharaibeh et al. [44] studied the accuracy and consistency of public and commercial IP geolocation databases. Similar to domain top lists, these services are often black-box operations, inhibiting public understanding of their data quality and limitations.

3 Method

In this study, we aim to measure the usage of popular top lists, focusing on Alexa, Umbrella, and Majestic. To answer our core research questions (from Section 1), we conduct controlled experiments with newly created domains under our control, manipulating top lists to include our domains at different ranking ranges. We monitor these domains over time to collect traffic telemetry, affording analysis of top list usage. As our study involves testing real-world artifacts, we discuss ethical considerations in Section 3.5.

3.1 Top List Manipulation

We aim to manipulate our experiment domains into different portions of top lists, to observe the impact on visitors to these domains. Specifically, we investigate how domain visit activity differs when the domain is placed in a top list's top 1M, top 100K, and top 10K. To manipulate the top lists, we rely on existing techniques, which we describe in detail in Appendix A. In short:

Table 1: The highest, lowest, and median rankings obtained when manipulating our test domains into the Alexa, Umbrella and Majestic lists.

	Alexa			Umbrella			Majestic
Range	1M	100K	10K	1M	100K	10K	1M
Highest	642,006	44,418	2,968	119,330	20,066	8,325	148,983
Lowest	983,356	62,485	6,393	490,015	54,925	9,683	623,773
Median	879,894	52,289	5,913	197,957	35,432.5	9,127	152,756

– **Alexa:** We manipulate Alexa using the same method from Xie et al. [72], which forges fake visits to a domain by generating requests for that domain to the Alexa Certify service's data collection endpoint [38].
– **Umbrella:** As Umbrella is a PDNS-based ranking, we manipulate it by generating DNS requests for a domain to Umbrella's DNS resolvers [61,66,72].
– **Majestic:** We manipulate Majestic using the method from Le Pochat et al. [61]. The method leverages certain "reflecting" sites, particularly MediaWiki sites, that accept user-provided URLs (i.e., our target domains) in the site's URL query parameters and reflect these user-provided URLs as anchor elements in their web pages. As a result, Majestic's crawler would observe a backlink to the target domain when visiting such reflecting sites. Using the Fofa search engine [17] and the set of reflecting wiki sites used in [61], we found and used 1,642 valid reflecting links for manipulation. We also subscribe to Majestic's service to better trigger Majestic's crawler to visit those links.

All manipulation experiments were conducted from a single server on one IP address within a large academic network, and we rate limited requests to minimize load at receiving endpoints (discussed further in Section 3.5).

Table 1 lists the rankings obtained for our Alexa, Umbrella and Majestic experiment for different ranking ranges. We note that manipulating Majestic is more challenging than Alexa and Umbrella though, as its ranking is not based on user traffic/visit, but rather the amount of backlinks to it. The highest ranking Le Pochat et al. obtained was ∼500K [61] and we obtain rankings as high as 150K. Thus, we only consider measuring Majestic's top 1M in this study.

3.2 Experimental Design

Our experiment is a controlled study of the three top lists (Alexa, Umbrella, and Majestic) across three ranking ranges (top 1M, top 100K, and top 10K), started performing from January, 2022 (just after Alexa announced the retirement). Note, as discussed in Section 3.1, we only evaluate Majestic on the top 1M. For each top list and ranking range pair, we created and monitored eight domains, split equally between a test set of four domains and a control set of four domains. For each set of four domains, we evaluated two realistic-looking domains and two randomly-generated (RG) domains, to investigate if domain name characteristics may affect its use in top lists. We chose to replicate our experiments across

two domains of matching characteristics to identify whether observations are consistent. Here we discuss how those domains are created, grouped, and tested.

Realistic versus RG Domains. As our study focuses on the usage of top lists, we are interested in understanding whether usage may vary based on the domain name. While there are various methods for constructing domain names, as a first exploration, we experiment with two classes of domain names, those that have a realistic domain name and those that are long and generated randomly (akin to those generated by domain generation algorithms).

To generate a realistic domain, we manually created a 9-letter domain using complete English words, although we avoided using words that may imply the site's function/content (e.g., "shopping") to avoid biasing visitors seeking certain classes of sites. For RG domains, we randomly generated a 50-letter domain name composed of randomly selected words (using the `nltk` corpus [26]), similarly avoiding functionality/content-related words.

Control versus Test Group. We manipulated our four test group domains into top lists and observed the traffic they received. All four domains in our control group were configured identically to the manipulated domains and similarly monitored, but were never manipulated into top lists, serving as baselines.

Experiment Phases. For each top list and ranking range, our experiment evaluated eight domains (2 realistic test domains, 2 RG test domains, 2 realistic control domains, and 2 RG control domains) over three phases spanning at least six weeks total. Phase I and II are two weeks each. Phase III is over two weeks, starting from the end of Phase II until the conclusion of our study.

(1) **Preparation Phase (Phase I)**: At least one week before Phase I, we configured valid DNS resolutions for all domains, but performed no further activities to avoid characterizing any initial traffic due to DNS registration. No activities (including top list manipulation) were performed during Phase I.
(2) **Active-manipulation Phase (Phase II)**: We manipulated test domains into the target top list and ranking range, and performed continued manipulation throughout this phase to maintain the domain's ranking. We remained inactive with control domains.
(3) **Idle Phase (Phase III)**: We halted top list manipulation of our test domains (and remained inactive with control domains).

3.3 Experiment Domain Setup

Here, we describe the setup for the experiment domains, which had valid DNS configurations and hosted websites.

Domain Names. All domains are newly registered with no past registration history. We registered our domains with NameSilo [23], using the "`.xyz`" top-level domain (TLD)[1]. We provided a unique email under our control in the WHOIS registration for each domain, listed as the registrant, admin, and technical contacts (the abuse contact was automatically set to the domain registrar).

[1] The "`.xyz`" TLD is a generic alternative to "`.com`", and has been used in prior web traffic measurement studies [49,60].

For each domain, we configured our own authoritative name servers using BIND9. For each domain, we set up two name servers with NS and A records, at the *ns1* and *ns2* subdomains. We set up A records for the SLD and the FQDN with a "*www*" prefix. We do not configure other DNS record types. We set the Time-to-live (TTL) of each domain's A record to 0 to limit DNS caching, aiming for our name server to observe as many DNS resolutions as possible.

Site Content. Our domains hosted Nginx-based websites that consisted of a simple HTML page that includes one line of text indicating that the site was a research experiment site, and listing an email contact for further information. We also provide a `security.txt` file [43] listing a different email contact. We do not configure a `robots.txt` file [48], allowing crawlers to visit our site. To monitor top list users visiting domains over HTTPS, our web server supports both HTTPS and HTTP (which redirects to HTTPS), using a valid TLS certificate from Let's Encrypt [20]. We note that while enabling TLS may lead to domain names appearing in certificate transparency logs, the potential impact of resulting traffic on the observed differences between the test and control groups should be negligible, as we use the same TLS configurations for all domains.

Web Hosting. We hosted our experiment websites at Vultr Cloud Hosting [34] on static IP addresses, allocating each site to a unique address, allowing us to reliably associate traffic with sites.

Justification of Domain Setup. We intentionally use fresh domains and content-free websites without existing visitors/traffic, to allow us to confidently associate domain traffic with top list placement and avoid confounding factors during analysis. If we used existing domains already receiving traffic, we would lack clean signals to distinguish the traffic driven by top list placement from other traffic sources. Similarly, providing realistic content on our site could attract traffic driven by the content rather than top list placement (e.g., site appearance on social media or search engines due to its content), rendering it infeasible to isolate the impact of top list ranking. Our method should capture both automated and manual/human-driven uses of top lists, although realistically, we expect that top lists are crawled at scale in an automated fashion. As a consequence, our study's results should largely identify and characterize the automated uses of top lists such as by various researchers and organizations.

3.4 Data Collection

We collected three types of telemetry for all experiment sites.

DNS Telemetry. For each DNS requests received at our authoritative name server, we recorded the following telemetry: Timestamp, Source IP address, and Requested DNS record type (e.g., A/AAAA/TXT).

Web Telemetry. We recorded the web traffic logs generated by our web servers. Specifically, for each web request, we recorded the following telemetry: Timestamp, Source IP address, Protocol (HTTP vs. HTTPS), HTTP method (e.g., GET, POST), Resource path URL (e.g., `/index.html`), Host HTTP header,

and User-Agent HTTP header. We further used the Maxmind database [21] to geolocate and identify the ASN mapping for source IP addresses.

Email Telemetry. As described in Section 3.3, each test site is associated with three unique contact emails (i.e., from the main web page, in `security.txt`, and in WHOIS records, we will call them the main email, `security.txt` email, and WHOIS email, respectively). Every site has a distinct set of emails, which we registered at Microsoft Outlook. We monitored emails received at these inboxes.

3.5 Ethics

As our study involved experimenting with real-world top lists, we must account for ethical considerations throughout the experiment design. Here, we discuss these ethical considerations for each component of this study.

Test Domains. All domains used in our experiment were new domains under our own control, specifically set up for this study. We notified our DNS registrar (Namesilo [23]) and hosting provider (Vultr [34]) about our study and received their consent. Our experiment does not affect any other domains. Beyond placing them in top lists, we do not distribute these domains elsewhere. We also signal the research nature of these domains through both the simple website hosted on the domains as well as the domain's WHOIS records. While our domain is associated with multiple emails (as described in Section 3.3), we did not receive any organic emails and did not respond to any messages. We did not interact with any human subjects in this study.

Top List Manipulation. Multiple prior studies [61,65,66,72] have conducted top list manipulation on the same set of top lists that we investigated. Our manipulation techniques are adopted from these prior works and the extent of manipulation (i.e., rankings achieved) are commensurate with these previous efforts. As we staggered our study over multiple months, at any given time during our study, we manipulated up to only four domains for a top list, which should have a negligible impact on the list's overall rankings.

For Alexa and Umbrella, we required generating requests to Alexa's data collection endpoint and Umbrella's DNS resolvers, respectively. While these endpoints by nature should be capable of handling heavy traffic load, as they collect the vast amounts of telemetry that feed into the top lists, we heavily rate limited our requests to these endpoints to avoid potentially burdening them and any transit networks. Even for our largest-scale manipulation, we did not generate more than 5 packets per second (with all traffic generated from a single server).

Our manipulation of Umbrella required spoofing the source IP address of DNS requests sent to the Umbrella DNS resolvers. We only spoofed IP addresses within our own local network, which was permitted. Furthermore, all DNS requests were only for test domains involved in our experiment, rather than any real-world online services. Thus, there should be negligible risk or harm to any hosts/individuals residing on a spoofed IP address. For Majestic's manipulation, the reflecting URLs we used were only crawled by Majestic, and did not have any impact on the sites themselves. We only submitted the reflecting URLs to

Majestic once, and the number of submitted URLs was within Majestic's quota for our subscription, and should not have overburdened Majestic.

3.6 Limitations

Experimenting on top lists is challenging, as they are live, complex, and opaque. As a result, our study does bear several limitations:

- We lack direct visibility into all top list use, and instead infer use through visits to our test domains. As our experiments are controlled, we can more confidently attribute observed differences between the test and control groups to top list presence. However, we ultimately may not fully capture all top list use (e.g., use of the top list where our test domains may have been filtered), and some differences may be partially driven by external factors.
- For our experiments, we manipulated four domains into a top list. We did not use a more diverse set of domains (e.g., different TLDs), as we aimed to limit our experiment's impact on top lists during this initial study, avoiding manipulating many domains concurrently into each list.
- To limit our experiment's impact on top lists, we chose a two-week duration for Phase II to capture top list usage behaviors within the two-week observation window. Thus, one-off usage occurring outside this window or periodic top-list usage exceeding two weeks may not be captured in this study. Users also may not visit a domain immediately after finding it; our study only captures usage behaviors where the domain visits occur within our observation window.
- For each experiment, our manipulated domains had similar rankings within each range, but the exact rankings varied. Results may differ slightly for sites at different rankings within a given range.
- Our experiment sites provide no meaningful content, and as fresh sites, have no traffic history or existing user base. This experiment decision is **necessary** to control for confounding variables in measuring the impact of top list placement. However, we will not capture top list uses that are restricted to types/sets of sites which exclude our domains. Rather our study will primarily capture automated crawling of the top lists, such as by various researchers and organizations.
- The Tranco top list [61] is now often used, and its ranking aggregates the three top lists we investigate. Domains appearing on one input list may also appear in Tranco, and we lack visibility into which list domain visitors are using. Our experiment results include the potential side effect of being listed in Tranco as well (although the Tranco ranking will be much lower).

4 Findings

In this section, we analyze our datasets that comprise of web, DNS, and email telemetry to evaluate how domain traffic changes once placed in top lists (RQ1), the characteristics of those visiting top list domains (RQ2), and the behavioral patterns of these visitors (RQ3). As we expect that top lists are primarily crawled

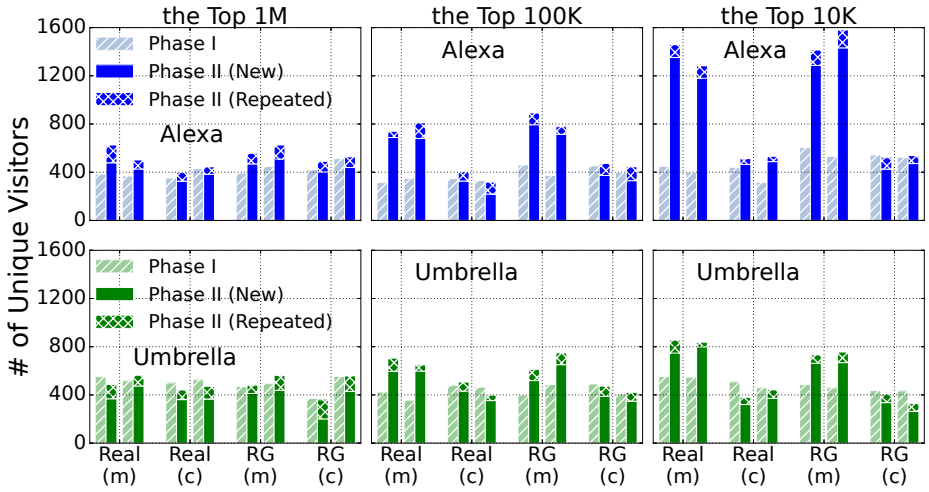


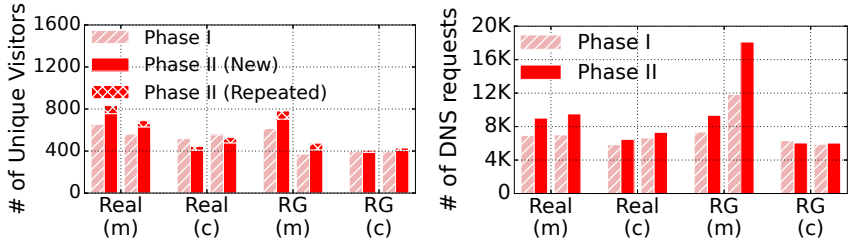
Fig. 1: The number of visitors (unique IP addresses) to each test domain's website observed in HTTP server logs during Phase I and II, across different ranking ranges for Alexa and Umbrella. For Phase II, we distinguish between "New" visitors not previously seen in Phase I, and "Repeated" visitors observed in both phases. ("Real" = realistic domain, "RG" = randomly-generated domain; "c" = control domain, "m" = manipulated domain)

at scale in an automated fashion, our findings will largely characterize the automated use of top lists, such as by various researchers and organizations.

4.1 RQ1: Impact of Top List Placement

We first evaluate the impact of top list placement by comparing the incoming traffic of manipulated domains before and after entering a top list. We compare the DNS and HTTP requests for manipulated domains as well as control group domains between the preparation phase (Phase I) and active-manipulation phase (Phase II). We investigate the long-term effects when a domain falls out of a top list by monitoring traffic during the idle phase (Phase III). We also examine the differences between the three top lists' academic use versus broader use (by observed visitors).

Web Traffic. For each experiment domain, we consider the number of visitors observed in HTTP server logs, counting the number of unique IP addresses issuing HTTP requests for the domain and its sub-domains. (We note that individual IP addresses do not necessarily represent unique visitors. Here we utilize unique IP addresses to analyze visiting traffic as we lack the visibility behind IP addresses. We will further analyze IP characteristics in Sections 4.2 and 4.3.) We analyze how the number of visitors varies in the two phases and across ranking ranges. (We do not consider request volume as visitors can generate



(a) (b)

Fig. 2: Majestic top 1M results: (a) the number of unique visitors (unique IP addresses) to each test domain's website; (b) the number of DNS requests for each experiment domain. ("Real" = realistic domain, "RG" = randomly-generated domain; "c" = control domain, "m" = manipulated domain)

varying numbers of requests, although our findings remain consistent for request volume.)

Figure 1 depicts the number of distinct visitors to each experiment domain observed in Phase I and II, across the three ranking ranges of Alexa and Umbrella. The Majestic top 1M result is shown in Figure 2a. Overall, for all three top lists, we observe more visitors accessed manipulated domains in Phase II compared to I, while control domains observed little to no increases in visitors. Thus, we conclude that top list placement does result in a notable increase in domain visitors, even without meaningful online services provided. (By manually checking the HTTP server logs, we identified that the increased traffic to the control group domains mainly stems from Internet-wide scans. Other cases include traffic resulting from the side effects of enabling HTTPS, as discussed in Section 3.3.) We also observe that across lists and ranking ranges, the number of daily visitors is stable throughout Phase II, indicating that the increase in visitors is consistent rather than ephemeral.

Higher rankings resulted in more web traffic, for both realistic and RG domains. In particular, placement in the top 1M resulted in significantly less visitor increases compared to the top 100K and top 10K. For example, placement in the Alexa top 1M produced a 45.1% (179) increase in visitors, averaged across the four manipulated domains. Meanwhile, the number of visitors can double or triple once a domain is in the top 100K or top 10K. Umbrella exhibits a similar pattern, although the increases are less extreme (50.5% increase for the top 100K and 55.6% for the top 10K).

We also observe that RG domains typically received more visitors once ranked compared to realistic ones, across lists and ranking ranges, hinting at some focus on odd/suspicious domains (e.g., randomly generated ones).

DNS Traffic. Here, we consider the number of DNS requests to each experiment domain. Unlike with web traffic, we do not consider the number of unique addresses as most observed DNS requests are from recursive resolvers rather

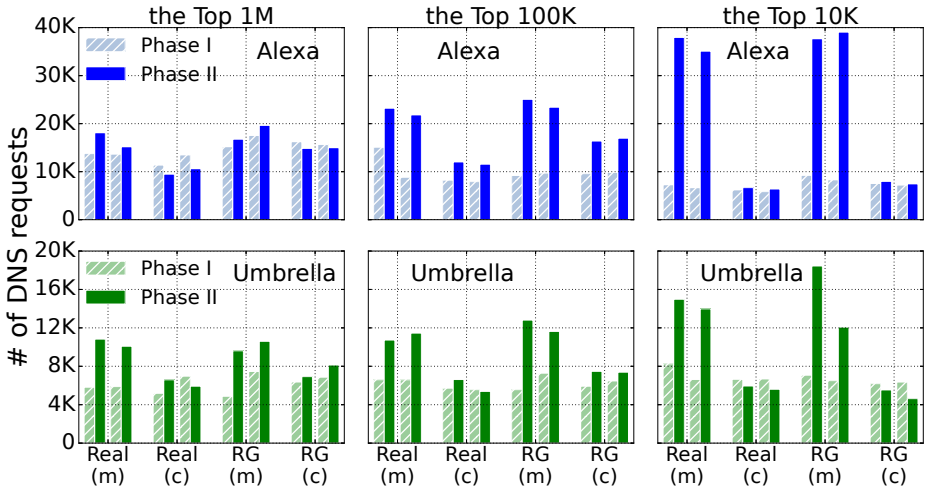


Fig. 3: The number of DNS requests for each experiment domain observed in DNS logs during Phase I and II, across different ranking ranges for Alexa and Umbrella. ("Real" = realistic domain, "RG" = randomly-generated domain; "c" = control domain, "m" = manipulated domain)

than clients. We briefly note that we observed two short massive bursts of DNS traffic for our manipulated domains while listed in the Alexa top 100K (and not for control domains). Based on the distinct request features in these bursts, we identified and filtered them out from our DNS telemetry (discussed more in Section 4.3). While we are not certain of this traffic's purpose, its anomalous nature highlights that top list ranking may render a domain as an attack target.

Figure 3 depicts the total number of DNS requests observed in Phase I and II for each experiment domain, across the ranking ranges of Alexa and Umbrella. The Majestic top 1M result is shown in Figure 2b. The patterns found in our web traffic analysis hold for DNS traffic too, across lists and ranking ranges. Again, DNS telemetry shows that placement in top lists consistently drives more domain traffic, with higher DNS lookup volumes for domains once ranked and minimal increases for control domains. Higher-ranked domains also observe higher DNS lookup volumes, and RG domains on average receive more DNS traffic compared to realistic ones. Thus, both web and DNS telemetry demonstrate that entering top lists positively affects multiple types of incoming traffic.

Long-term Effects. We investigate the long-term effects when a domain falls out of the top list. Here we focus on characterizing data from our top 100K experiments, which we conducted first and hence had the longest period for Phase III, although we observe similar outcomes for other ranges. We observe that over the four month period after de-listing, web visitors and DNS lookup volumes quickly decreased to pre-listing levels within two weeks, for both lists, indicating that most uses of top lists rely on more recent daily snapshots.

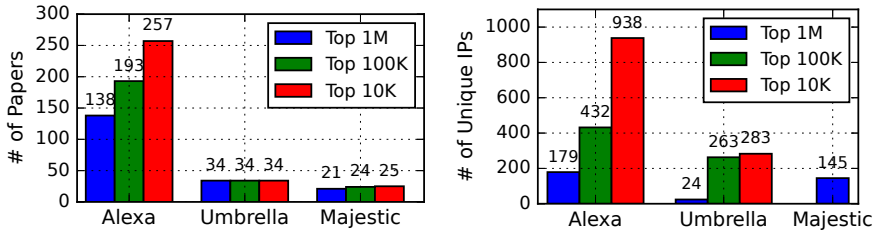
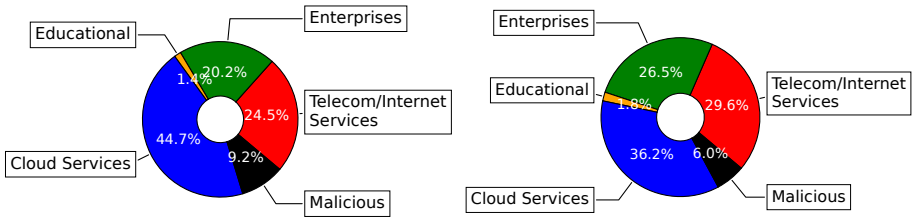


Fig. 4: Academic vs broader top list use in practice: (a) shows the number of papers using each top list and ranking range from 2017 to 2022, across 10 networking and security venues, while (b) depicts the number of distinct new visitors in Phase II across different lists and ranking ranges, averaged across the two types of manipulated domains (note that we did not experiment with the Majestic top 100K and top 10K.)

Academic vs Broader Use. Scheitle et al. [67] evaluated top list use by academic studies published in 2017 across 10 networking/security venues, finding that 69 papers relied on top lists: 59 used Alexa, 3 used Umbrella, and none used Majestic. We extended this survey across the same 10 venues from 2017 to 2022 (we list the 10 venues and describe our survey details in Appendix B). Figure 4a shows the number of studies using each list across each ranking range (note, if a study uses a top 1M list, we also include it as using the top 100K and top 10K.) We observe, similar to the previous study, that academic studies have skewed heavily towards using Alexa (particularly higher ranking ranges), although Umbrella and Majestic are both used (primarily the top 1M).

When considering our top list domain visitors, as shown in Figure 4b, we found significantly more top list use over the two-week observation period compared to use by the academic studies published in the prior 6 years. While our academic literature survey was limited in the venues considered, the scale of the discrepancy suggests that academic research accounts for a minority of broader top list use.

We do see that, similar to academic studies, Alexa was most heavily used, especially the higher ranking ranges. However, we see similar numbers of Majestic top 1M visitors as Alexa top 1M visitors (we were unable to experiment with Majestic top 100K and top 10K). We hypothesize that this similarity arises because broader top list use skews towards investigating websites, as ranked by Alexa and Majestic, even if ranking methods vary significantly. In comparison, we observed minimal use of the Umbrella top 1M (whereas academic studies primarily used the Umbrella top 1M), although we do see elevated usage of the Umbrella top 100K.



(a) Alexa Top 10K (b) Umbrella Top 10K

Fig. 5: Distributions of new visitors in Phase II, averaged across the four manipulated domains, for the Alexa top 10K and Umbrella top 10K.

4.2 RQ2: Characteristics of Top Domain Visitors

In Section 4.1, we answered our first research question, finding that top list placement can result in significant amounts of traffic to a domain. However, top 1M placement produces relatively limited traffic increases. Thus, moving forward, we focus on analyzing the traffic that results from placement in the higher ranking ranges (also removing Majestic from consideration, as we only experimented with its top 1M).

We now analyze the web traffic logs to evaluate top domain visitors. We do not consider DNS traffic here, as we lack visibility into the DNS lookup clients (as discussed in Section 4.1). Specifically, we analyze new IP addresses observed in the HTTP server logs once a domain is ranked (Phase II) but not previously in Phase I, characterizing their ASes, countries, and user agents. We use the Maxmind databases [21] for IP geolocation and ASN mapping.

IP Reputation. Before characterizing visitors, we investigate the reputation of IP addresses newly observed during Phase II. We query the addresses using VirusTotal [33], which aggregates classification results across many third-party security vendors. As some vendors may produce false positive labels, we consider an address as malicious only if classified by at least three vendors (as suggested by prior work [59]).

As shown in Figure 5, we found that 9.2% of addresses were classified as malicious for the Alexa top 10K and 6.0% for the Umbrella top 10K, averaged across the four manipulated domains. For the top 100K range, we observed that 7.4% and 5.2% of addresses were labeled as malicious for Alexa and Umbrella, respectively. (We omit the top 100K from Figure 5, as it exhibits similar patterns.) By manually inspecting the malicious addresses, we do find suspicious behavior. We observe addresses with at least 10 vendors classifying as malicious, which issue suspicious requests to our domains (e.g., "\x03\x00\x00/*\xE0\x00\x00\x00\x00 \x00Cookie: mstshash=Administr"), potentially searching for web vulnerabilities to exploit. Thus, top list domains do attract malicious visitors, especially for the Alexa list and higher ranking ranges.

ASN. Here we investigate the top ASNs that contribute the most new visitors once a test domain is manipulated into a top list. A new visitor represents an address observed in Phase II that was not observed in Phase I. Through manually

inspecting AS names, we categorize visitor ASes into different categories, and depict the distribution of visitors over different AS categories for both the Alexa and Umbrella top 10K in Figure 5. (We elide a figure for the top 100K, which exhibits a similar distribution.)

Across top lists, rankings, and domain types, we observe that the top ASNs are primarily cloud service and hosting providers, such as Google Cloud (396982, 15169), DigitalOcean (14061), Amazon (16509), and HostRoyale (203020). While these organizations may use top lists themselves (which we do observe during user-agent analysis, discussed shortly), the volume of visitors suggests that many top list users access ranked domains through cloud platforms (even for the malicious users), presumably using automated methods.

Beyond cloud-related ASes, we identify visitors from various enterprise networks (around 20-27% of visitors, as shown in Figure 5, particularly for security organizations such as Zscaler (22616), Eonscope (208417), and Censys (398324, 398722). Other types of enterprises, such as Hangzhou Alibaba Advertising (37963, advertising), hint at different purposes behind top list uses.

We also find ASes providing telecom and Internet services, potentially serving both residential and enterprise networks (accounting for 24-30% of visitors, as shown in Figure 5), such as Comcast (7922) in the US, PJSC VimpelCom (3216) in Russia, and TATA Communications (4755) in India.

We do see visitors located in the ASes for educational/research institutions, aligning with known uses of top lists in academic research studies [67]. As examples, we observe MIT (3) and Boston University (111) from the US, Seoul National University (9488) from South Korea, Technische Universitaet Muenchen (209335) from Germany, and China Education and Research Network Center (4538). Such visitors only accounts for 1-2% of all visitors, aligning from our prior analysis showing that academic research likely accounts for only a minority of top list use.

Notably, we find that for Alexa, there is a 40% increase in the number of distinct visitor ASNs (averaged across experiment domains) for the top 10K compared to the top 100K (169 vs 120), indicating more diversity in visitor ASNs for higher-ranked Alexa sites. However, for Umbrella, we do not find a consistent difference in the number of visitor ASNs between the top 100K and the top 10K (107 vs 108). The number of visitor ASNs for the Alexa top 10K is also significantly higher than for Umbrella (169 vs 107), suggesting that Alexa domains not only receive more visitors than Umbrella (as observed in Section 4.1), but more diverse visitors.

Country. Table 2 lists the top 5 counties/regions by the number of new visitors in Phase II, using the same definition for a new visitor as with ASNs. We observe that across lists, rankings, and domain types, the top countries overall include the US, the Netherlands, and China (Russia and Germany also frequently appeared).

When inspecting the ASNs of visitors geolocated to the US, the majority are from cloud providers such as Amazon (16%–47% of visitors, across all manipulated domains), Google Cloud (6%–30%), and DigitalOcean (7%–18%). Simi-

Table 2: The top 5 countries/regions by their number of new visitors in Phase II, for the four manipulated domains, across Alexa and Umbrella's top 100K and top 10K.

	Alexa Top 100K				**Alexa Top 10K**			
	Realistic-#1	Realistic-#2	RG-#1	RG-#2	Realistic-#1	Realistic-#2	RG-#1	RG-#2
1	US (224)	US (181)	US (287)	US (248)	US (427)	RU (422)	RU (397)	US (397)
2	NL (211)	NL (196)	NL (209)	NL (206)	RU (391)	US (298)	US (387)	RU (393)
3	CN (76)	CN (63)	BR (83)	BR (80)	CN (316)	CN (228)	CN (227)	CN (383)
4	RU (50)	RU (51)	RU (57)	CN (61)	NL (91)	NL (101)	NL(103)	NL(106)
5	DE (23)	DE (22)	CN (32)	RU (55)	DE (35)	DE (37)	DE (31)	HK (37)
#Countries	33	38	35	40	42	37	42	38
#Cities	114	111	101	110	149	141	158	154
	Umbrella Top 100K				**Umbrella Top 10K**			
1	NL (190)	NL (193)	US (186)	US (197)	US (289)	CN (329)	US (275)	US (270)
2	CN (164)	CN (172)	NL (166)	NL (191)	CN (182)	US (224)	NL (115)	NL (122)
3	US (159)	US (171)	RU (48)	CN (73)	NL (128)	NL (87)	CN (89)	CN (88)
4	RU (48)	RU (70)	CN (45)	RU (67)	RU (32)	RU (24)	RU (32)	RU (30)
5	DE (20)	DE (27)	DE (17)	FR (31)	DE (28)	DE (16)	DE (21)	DE (24)
#Countries	34	37	35	38	34	28	31	35
#Cities	99	103	81	99	98	108	98	102

larly, most visitors from the Netherlands are from Google Cloud (80%–95%) and DigitalOcean (1%–5%) data centers in the Netherlands. Note that as many visitors geolocated to the US and the Netherlands used cloud platforms to host their clients, we could not confidently attribute these visitors to those two countries. Meanwhile, visitors geolocated to other countries are more strongly geographically correlated. Chinese visitors are primarily from two Chinese ASes: ChinaNet (36%–49% of Chinese visitors) and ChinaUnicom (21%–46%).

Table 2 also lists the number of countries and cities that new visitors in Phase II geolocate to. We do not see significant differences in the number of visitor counties across the ranking ranges for both lists. At the city granularity, we observe limited variation between the Umbrella top 100K and 10K. In contrast, visitors to Alexa top 10K domains exhibit significantly higher city diversity compared to Alexa top 100K (151 vs 109, 38.5% higher, averaged across domains), with the variation primarily arising through US and Chinese cities.

User Agent. Finally, we investigate the HTTP user-agent headers in web requests from IP addresses newly observed during Phase II. Note that as user-agent strings can be spoofed, we lack ground truth on the real user-agent used, and our analysis is limited to characterizing the user-agents as is.

Across all manipulated domains for both top lists and ranking ranges, we identify that over 75% of visitor have user-agent strings for various browsers and browser versions. Among browser user agents, the majority were for the Chrome browser (64% of visitors, across all manipulated domains in both top lists and ranking ranges), followed by Firefox (13%), Opera (10%) and IE (6%). (As discussed above, visitors can modify their HTTP header arbitrarily [56], thus the results may show a browser tendency to be set by the visitors, rather than

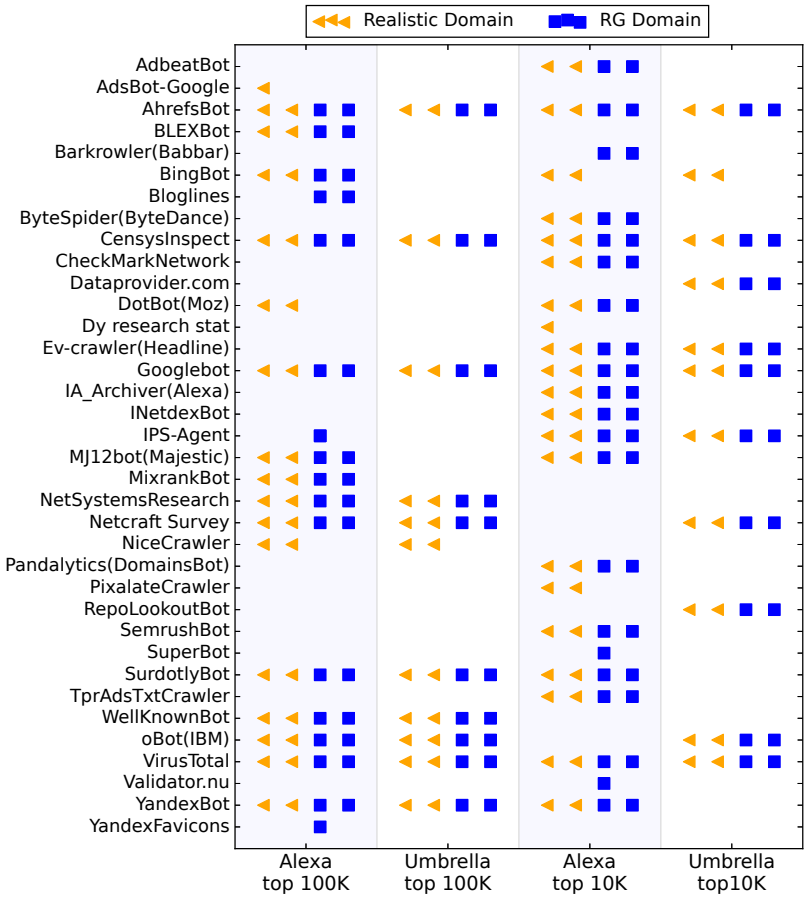


Fig. 6: Crawlers observed as new visitors during Phase II for the four manipulated domains. Each dot represents a crawler observed for a domain.

actual browser use.) We see several mobile browsers, where Mobile Safari and Chrome Mobile are typically in the top 10 software/browsers used by visitors. (We suspect these are spoofed user agents, as it seems less likely that we would observe organic mobile traffic to our experiment sites.) We also observe the use of different variants of browsers, e.g., Chromium and Waterfox. According to these user agent headers, we found Windows to be the most common OS among visitors, followed by Mac OS X, then Linux variants. For mobile systems, we observe visitors used Android more than iOS (with a long tail of other mobile OSes). By manually inspecting the remaining non-browser user agents, we find common networking tools and programming frameworks used by visitors (e.g., networking libraries for Python and Go, command-line tools such as `curl` and `wget`, scan tools such as `masscan` and `zgrab`).

We also uncover visitors self-identifying as web crawlers (around 9-15% of visitors, across all manipulated domains for both lists and ranking ranges) for various organizations, as shown in Figure 6. These include for search engines (e.g., Googlebot [18], BingBot [8], YandexBot [36]), web analytics services (e.g., Pandalytics [28], Netcraft Survey [24], Dataprovider.com [14]), advertising/marketing services (e.g., AdbeatBot [5], AdsBot-Google [6], AhrefsBot [7]) and security organizations (e.g., VirusTotal [33], CensysInspect [10], SurdotlyBot [31]), hinting at the purposes behind top list uses. We observe about twice as many crawlers for Alexa as for Umbrella, for both ranking ranges, aligning with our prior observations of Alexa's popularity over Umbrella (see Section 4.1). Interestingly, we observe Majestic's crawler on Alexa domains (across both ranking ranges) but not on Umbrella domains, likely due to the web-specific nature of both Alexa and Majestic. Only AhrefsBot, CensysInspect, GoogleBot and VirusTotal crawled all of our test domains across both top lists and ranking ranges. These crawlers are associated with search engine or security services.

Unsurprisingly, some crawlers (e.g., CheckMarkNetwork [11], Ev-crawler [16]) only visited domains within the top 10K. However, we also found crawlers only accessing domains in the top 100K range (e.g., Bloglines [9], NetSystemResearch [4], NiceCrawler [25], and WellKnownBot [35]). We hypothesize that, as our top 100K and top 10K experiments were conducted at different periods of time, this behavior may have arisen due to the crawlers executing infrequently, such that we did not observe them during our top 10K experiments. We believe it is unlikely that many crawlers visit the top 100K domains of a top list but intentionally avoid crawling the top 10K.

We also note minor differences between domain types. For example, NiceCrawler only accessed realistic domains for both Alexa and Umbrella's top 100K, while Bloglines only accessed RG domains in Alexa's top 100K. Experiment timing and ranking differences cannot account for this behavior, as the domains of both types are concurrently listed with similar rankings. We hypothesize that these top list users focus on only certain types of domain names.

Finally, we observe possible effects from Alexa's retirement, where oBot (from the IBM Security X-Force Threat Intelligence [19]) crawled test domains in all other three groups but was absent in Alexa's top 10K (as our top 10K experiment was close to Alexa's retirement).

4.3 RQ3: Behaviors of Top Domain Visitors

Here, we tackle our final research question on the behavioral patterns of top domain visitors. Again, we specifically investigate new visitors (IP addresses) during Phase II, when a test domain is placed into a top list, who had not previously appeared during Phase I. We again focus on web traffic as observed DNS queries primarily originate from recursive resolvers rather than clients. In this analysis, we focus on visitor access frequency, use of TLS, popular resources requested, and messages sent to domain-associated emails.

Access Frequency. We first study how often visitors access a domain once placed in a top list. Figure 7a depicts the distribution of the number of days a

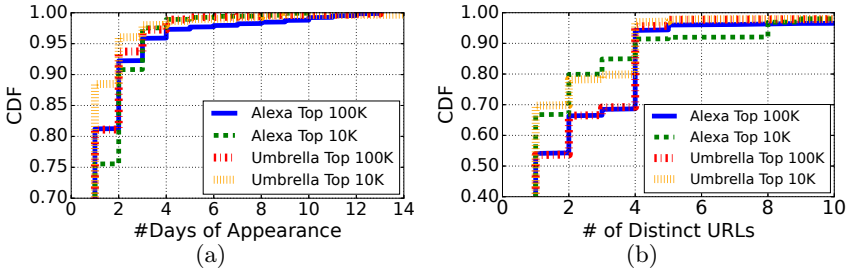


Fig. 7: CDF of (a) # of days that a new visitor accessed a ranked domain in Phase II; (b) # of distinct URLs/resources accessed by new visitors in Phase II.

new visitor in Phase II accessed our experiment domains (aggregated across the manipulated domains for each list and ranking range). We observe a dominant pattern across the top lists and ranking ranges, where the majority of visitors (over 75%) only accessed the experiment domains on a single day. This access pattern likely reflects two classes of top list use: 1) one-off/adhoc top list uses, such as single snapshot measurements as often done in prior academic studies (e.g., [46,52]), or 2) repeated periodic top list use, where the crawling periodicity exceeds two weeks (our Phase II period), such as monthly crawls of a top list.

Looking at the outlier visitors that crawled our domains on more than 7 days, we identify that those visitors are either associated with security-related organizations (Forcepoint and VirusTotal), or cloud providers (Google, Azure, and Digital Ocean). Security-related organizations may need to crawl domains more frequently for detecting malicious domains. Meanwhile, cloud provider visitors may include companies hosting their crawling infrastructure on cloud platforms, as well as researchers conducting longitudinal top list measurements. For example, we observe one Digital Ocean visitor repeatedly crawling for the `dnt-policy.txt` of domains, and one visitor from Amazon EC2 crawling `security.txt` files.

HTTP Protocol/Request Method Prevalence. We now investigate the HTTP protocol (HTTP vs HTTPS) and HTTP request methods used by new visitors to top list domains during Phase II. Table 3 lists the percentages of visitors that were seen accessing our domains over HTTP, HTTPS, or both, for each top list and ranking (as the realistic and RG domains exhibit similar patterns, we list averages across manipulated domains). Domains in the Alexa and Umbrella top 100K exhibit similar distributions, with ∼40% of visitors only using HTTP and ∼25% only using HTTPS. The remaining third accessed the domain using both protocols. For domains in the top 10K, we observe notable differences. Over half of visitors to Alexa top 10K domains used both protocols, almost 20% more than in the other three experiment groups. For Umbrella top 10K domains, the percentage of visitors using both protocols remained similar to the top 100K, but 10% more visitors used only HTTPS. This difference may be driven by the broader TLS adoption observed for higher-ranked domains [67].

Table 3: The averaged percentages of new site visitors during Phase II using only HTTP/only HTTPS/both.

	HTTP	HTTPS	Both
Alexa Top 100K	41.4%	25.6%	33.0%
Umbrella Top 100K	41.2%	26.0%	32.8%
Alexa Top 10K	**19.6%**	**28.8%**	**51.6%**
Umbrella Top 10K	**29.9%**	**36.1%**	**34.0%**

For HTTP request methods, we observe a wide array of methods across domains in the Alexa and Umbrella top 100K and top 10K, including `GET`, `HEAD`, `POST`, `CONNECT`, `OPTIONS`, and `PRI` (HTTP/2.0) methods. The `GET` method is used by more than 90% of visitors across the two ranking ranges, for both Alexa and Umbrella. The `HEAD` method is the second most popular, particularly for Alexa, which we observe used by 4.6% and 6.0% of visitors to Alexa top 100K and top 10K domains, respectively (averaged across the manipulated domains). For Umbrella, ∼3% of visitors used the `HEAD` method across the two ranking ranges. Beyond the top HTTP methods, there is a long tail of other methods used, including `PUT` and `REQMOD` (ICAP mode), illustrating diverse purposes behind top domain visits.

Resource Prevalence. We next study what web resources are commonly requested by new visitors in Phase II. We observe similar resources frequently requested from manipulated domains in each top list and ranking range group. In Table 4, we select one of the realistic domain's results as a representative example, and list the top 10 most accessed URL paths by the percentage of new visitors in Phase II requesting it. We observe several interesting cases:

- *Root.* The majority of visitors (58–80% across different lists and ranking ranges) requested the root "/" home page of a domain. This does indicate that a notable fraction of visitors do not crawl the domain root at all, and rather focus on accessing other resources.
- *Favicons.* Across top lists and ranking ranges, another one of the top 3 requested resource is the "/favicon.ico" path, for a domain's favicon. Modern browsers generate favicon requests when visiting a domain, and we hypothesize that many of these requests arise from visitors using real browsers (whether manually or programmatically). Between 10% and 31% of visitors requested the favicon, suggesting only a minority of visitors use real browsers.
- *RSS Feeds.* We observe a large fraction of visitors (up to 16%) requesting "/feed", "/feeds", and "/rss" URL paths, commonly associated with RSS feeds [62]. These visitors cluster within two Google Cloud ASes (396982 and 15169). We note that Google's Feedfetcher [47] similarly crawls these RSS or Atom feeds, but presents a distinct user agent (i.e., "Feedfetcher-Google"). Thus, these visitors are likely other RSS feed aggregators operating on Google Cloud, rather than Google's Feedfetcher.

Table 4: Top URL paths/resources by the percentage of new visitors in Phase II accessing it.

	the Alexa Top 100K		the Alexa Top 10K		the Umbrella Top 100K		the Umbrella Top 10K	
	URL Path	%	URL Path	%	URL Path	%	URL Path	%
1	"/"	80.0	"/"	57.7	"/"	75.8	"/"	75.0
2	"/feed", "/feeds", "/rss"	13.8	"/favicon.ico"	30.8	"/feed", "/feeds", "/rss"	16.0	"/favicon.ico"	9.7
3	"/favicon.ico"	10.3	Russia-Related URLs[3]	8.6	"/favicon.ico"	15.8	"/feed", "/feeds", "/rss"	7.2
4	"/ads.txt"	4.5	"/feed", "/feeds", "/rss"	5.1	"/1jlbdmb"	6.5	"/robots.txt"	6.1
5	"/robots.txt"	4.2	"/gaocc/g445g"	3.7	"/ads.txt "	3.8	"/1jlbdmb"	4.5
6	""	3.9	"/robots.txt"	3.5	""	3.1	"*"	3.0
7	WordPress Scans[1]	3.1	"/1jlbdmb"	2.3	"-"	2.5	""	2.3
8	"/1jlbdmb"	2.2	""	1.6	"/robots.txt"	2.3	"/ads.txt"	2.3
9	security.txt[2]	1.1	"/.git/config"	1.4	WordPress Scans[1]	2.0	"MGLNDD_[IP]_443"[4]	1.6
10	"/.env"	0.9	"*"	1.3	security.txt[2]	1.0	WordPress Scans[1]	1.6

[1] We observe a group of IP addresses accessing several WordPress vulnerability URLs, such as "//blog/wpincludes/wlwmanifest.xml", "//wordpress/wp-includes/wlwmanifest.xml", and "//2019/wp-includes/wlwmanifest.xml".
[2] Requests for `security.txt` were to the "/.well-known/security.txt" path.
[3] Russian-related URL paths as discussed under *Nation-State Related Activity.*
[4] "[IP]" represents the IP address of our sites.

– *`.txt` Files.* A non-trivial fraction of visitors requested different standard `.txt` resources, with `robots.txt` and `ads.txt` being most frequent overall across top lists and ranking ranges. `robots.txt` informs crawlers on which site resources are permitted to be crawled, but only a small fraction of visitors (less than 7% for all domains) requested this information, indicating limited adherence to this standard. Meanwhile, `ads.txt` provides information on advertising relationships, and such visitors crawling this resource (less than 5% for each domains) are likely involved in online advertising. We observe up to ∼1% of visitors accessing `security.txt` files on some domains, suggesting harvesting of security contacts or signals about security postures. Other `.txt` files were accessed but by less than 10 visitors per domain, including `app-ads.txt`, `humans.txt` and `dnt-policy.txt`.
– *Nation-State Related Activity.* We observe several interesting resources accessed by a notable fraction of visitors that seem related to nation-state activities, specifically for domains in the Alexa top 10K. Approximately 4% of visitors accessing our domains placed in the Alexa top 10K requested the "/gaocc/g445g" URL path, all from Chinese IP addresses. We identify anecdotal evidence that this URL path is related to configuring the V2Ray censorship circumvention tool[2], and that crawls for this resource may indicate Chinese censors attempting to detect servers supporting censorship circumvention[3]. (We also observe "/1jlbdmb" frequently requested. It is unclear what this resource is associated with, although upon Google searches, we do note a large number of query results about odd traffic to this URL path in Chinese, so this resource may also be related to Chinese censorship.) Similarly, we observe

[2] https://github.com/v2fly/v2ray-core/issues/304
[3] https://twitter.com/germanyorthoped/status/1405413138468536322

Table 5: The number of distinct contacts that messaged email addresses on our test domains' main pages or `security.txt` files.

#Contacts	Alexa			Umbrella		
	1M	100K	10K	1M	100K	10K
Main (control)	3	2	3	1	1	0
Main (manipulated)	**7**	**6**	**12**	**1**	**4**	**6**
`Security.txt` (control)	0	0	0	0	0	0
`Security.txt` **(manipulated)**	**0**	**3**	**5**	**1**	**4**	**2**

that ∼9% of visitors requested the "/russianfederation", "/russia-w1", "/lenta", "/aeroflot" URL paths (as well as "/kfc", "/kfccorporationx" and "/blablacar-w", which appear related to US and French corporations), all from Russian IP addresses. We are uncertain of the reason behind requesting these resources, although we suspect that they may be related to either Russian censorship or the ongoing war in Ukraine.

– *WordPress Scans.* Many visitors launching WordPress scans on our ranked domains, specifically requesting for the `wlwmanifest.xml` file under various URL paths. We find anecdotal evidence [1,63] that such scans are often associated with vulnerability scans of WordPress installations, suggesting that security researchers or attackers leverage top lists to find vulnerable sites.

In Figure 7b, we also depict the distribution of the number of distinct resources requested by new visitors during Phase II (again using a realistic domain for each top list and ranking range as a representative example). Overall, the number of resources requested is low, with at least 55% of visitors only requesting a single resource and over 90% of visitors requesting four resources or less, across both top lists and ranking ranges. Thus, most visitors do not extensively crawl a site (although we note that as our site was simple, it is possible that some visitors would have more extensively crawled had our site contained more links). Looking at the outliers, we observe a small number of visitors crawling hundreds of resources (up to 1.5K). The most extensive crawler scanned for 1,545 Polycom VOIP configuration files (e.g., "/dms/Polycom_VVX_201_000000000000.cfg"). Another visitor accessed over 200 URLs seemingly for vulnerability scanning (e.g., "/error3?msg=30&data=';alert('nuclei');", likely related to the Nuclei vulnerability scanner [27]).

Email Traffic. Finally, we look at how visitors use contact information on top list domains. Table 5 lists the number of unique email addresses that contacted an email associated with our domain once placed in a top list (Phase II). We received few emails throughout our measurement. Interestingly, we did not receive any emails for our Majestic domains nor to our WHOIS contacts.

Overall, we observe that manipulated domains received more emails than control domains, for both emails associated with domain landing pages and `security.txt`. We note that all email addresses that contacted our control domain also contacted our manipulated domain, indicating that these contacts

identified our domains through means other than top lists (potentially domain registration information or certificate transparency logs). We also observe that higher ranked domains generally received more messages as well, and that more contacts used the main page emails compared to `security.txt` ones (as discussed earlier, we did observe a small fraction of visitors crawling `security.txt` files, but significantly fewer than the domain root). By manually inspecting all messages received, we classify the emails as either advertising/spam or phishing/scams. Interestingly, all emails to `security.txt` addresses were phishing/scams, suggesting that some visitors crawl `security.txt` to identify valuable security contacts for a site to target. Thus, top list placement does result in additional unwanted/malicious emails to email contacts associated with the page. This is particularly relevant for `security.txt`, as a common concern is that listing security contacts in such files will result in high volumes of spam content [51], although we note that the email volumes we directly observed is small.

Anomalous DNS Traffic. During our Alexa top 100K experiment, we observed two massive floods of DNS requests on two different days. Each manipulated domain received more than 1M requests within a 20-minute window on both days, whereas the domains only received hundreds to thousands of requests per day otherwise. Thus during these bursts, nearly all DNS lookups were due to the flood. By inspecting the DNS requests during the floods, we identified that these queries were for subdomains of our experiment domains, and originated from only three ASes, Google (15169), WoodyNet (42) and Cisco OpenDNS (33692), which all host public DNS services. To prevent these floods from skewing our DNS telemetry, we filtered out all such queries (subdomain queries from the three ASes during the 20-minute windows). We briefly note that Google DNS forwards the original DNS client's network through EDNS Client Subnet, and we observed that client IP addresses were all within a DigitalOcean AS (14061). While we are not certain of the purpose behind this traffic, its anomalous and suspicious nature highlights that top list ranking may render a domain a ready target for attacks.

5 Concluding Discussion

In this paper, we empirically investigated real-world top list use by conducting controlled experiments with test domains in different ranking ranges of popular top lists. Here, we synthesize lessons from our study and future directions.

Lessons for Top List Design. Our findings demonstrate ongoing dependencies on top list datasets. While simple, this observation is especially salient in light of Alexa's retirement [39]. While the consequences of Alexa's retirement remain to be seen, there is clearly a need for alternative options. We observed that of the three top lists considered, despite facing impending retirement, domains placed in Alexa received the highest levels of traffic, and prior research studies have depended primarily on Alexa over the other options (see Section 4.1).

While the Tranco top list [61] has become more popular of late, at least in academic studies, it is ultimately an aggregator of existing lists rather than its

own distinct top list, and the loss of Alexa reduces Tranco's input data. After Alexa's end, Tranco has since replaced it with a new PDNS-based ranking by DomainTools [22]. However, we note that DomainTools' ranking itself combines Umbrella and Majestic, in addition to Netcraft top 100 sites [57] (only containing 100 sites ranked) and Farsight Security's PDNS data [22]. It is unclear currently what the implications are of the new Tranco's heavy dependence on PDNS data, as well as its double dependency on Umbrella and Majestic (both direct dependency as well as indirect dependency via DomainTools).

Given the community's dependence on top lists, there is a need to investigate new top list designs. Recently, Xie et al. [72] proposed a new PDNS-based top list design, SecRank, that achieves desirable top list properties. Such developments may serve as promising alternatives to Alexa in the future, especially as SecRank's design is transparent, although SecRank's current implementation inputs Asia-centric DNS data, and thus exhibits regional skew in its ranking.

Our study's findings can help inform top list design considerations. For example, we observed that traffic to top list domains primarily increases once in the top 100K, suggesting that most uses of top lists are within that ranking range. Thus, top lists should aim for higher-quality rankings at such scales, rather than prioritizing larger ranking quantities (i.e., there may be less value in having millions of domains ranked compared to a 100K). In addition, our findings hint at a preference for website-based top lists (e.g., Alexa, Majestic) regardless of the ranking methods, whereas SecRank and Umbrella (as well as Cloudflare's new Radar ranking [13]) are both DNS-based and contain non-website domains. New top lists may be more broadly used if focusing on collecting web traffic telemetry for ranking websites.

We also identified how geographically diverse top list use is (in Sections 4.2 and 4.3), with visitors from over 40 countries. However, existing top lists exhibit bias towards certain geographic regions. For example, SecRank is built on PDNS data from a Chinese DNS provider and thus skews towards Asia-centric domains, whereas the other top lists skew towards popular domains in Western countries, particularly due to their US and European-centric data sources. Constructing top lists that focus on different geographic regions could support more geographic diversity in network and security measurements.

Furthermore, our experiments identified various organizations relying on top lists for multiple purposes, including for search engine indexing and security evaluations (as discussed in Section 4.2). Given these sensitive use cases, top lists must be designed with robustness against manipulation. The threat of domains manipulated into top lists is not purely hypothetical though, as online websites have been identified offering top list manipulation as a paid service [72] (often with high prices, such as $40/month for entering the Alexa top 100K, and $500/month for its top 10K [3]).

Lessons for Top List Usage. While many security analysis tools/services allowlist domains on existing lists, our results highlight how readily this allowlisting can be abused. We identified in Section 4.3 that the majority of list users, including various security organizations, either assessed a site only once after

top list placement, or recrawled the site only with a long periodicity (although we did observe a few outliers who recrawled frequently). Thus, an attacker could first manipulate a benign domain into the rankings, resulting in allowlisting by security tools and services, and then subsequently modify that site to a malicious one. Instead, sensitive uses of top lists (such as for security purposes) should regularly revisit sites on top lists, to avoid relying on stale information.

Lessons for Ranked Domains. We observed that ranked domains receive various types of traffic. This traffic, as discussed in Section 4.3, includes advertising, spam, scam, and phishing messages to domain-associated emails, potentially by malicious actors looking to exploit sites. Of particular note is that deploying `security.txt` resulted in a small number of malicious emails, aligning with concerns of spam or low-quality reports to `security.txt` contacts [51]. Thus, while `security.txt` remains a promising protocol for providing security contacts for a website, domain owners must be prepared to handle spam reports.

We identified that most visitors to our experiment domains did not crawl `robots.txt` (Section 4.3), much less adhere to it, similar to prior findings [45]. Only a small fraction of visitors self-identify as web crawlers, despite the vast majority of visitors arriving from cloud platforms (and thus are likely crawlers). For ranked domains seeking to limit crawling, anti-bot techniques should leverage the AS classifications of visitors, beyond relying on `robots.txt` and signals from the user-agent. We also observed that top list ranking results in a non-trivial number of suspicious visitors and accessing patterns (Sections 4.2 and 4.3). Thus, ranked domain owners (especially higher-ranked domains) require defense measures, such as DDoS protection and appropriate DNS and web cache configurations.

Lessons for Future Research. Future work can more extensively study top list use in practice, as our work is ultimately a first step. One direction is in better understanding the purposes behind top list use, as our study's experiment did not afford detailed visibility into how top lists domains may be used after crawling. Such investigations may involve user studies, and could investigate interesting use cases such as allow/block-listing and domain classification. Our case studies in Section 4.3 also shed light on how top lists could be used as a gateway for measuring censorship (or other scanning behaviors) in a blackbox fashion, as top lists may serve as a source of sites evaluated for potential censorship/scanning. Related, top list usage could be studied over longer periods of time, as our experiments monitored sites only on the order of weeks, rather than months or years. Both ephemeral (e.g., holidays, special events) and long-term effects could be identified through such longitudinal studies, providing deeper insights into top list use in practice. Ultimately, the importance of top lists across various measurement and evaluation use cases motivates deeper future investigation into top list characteristics.

References

1. Weird GET and POST requests node. https://www.digitalocean.com/community/questions/weird-get-and-post-requests-node (2021)

2. RankStore. https://web.archive.org/web/20220314142408/http://www.rankstore.com/ (2022)
3. Alexa Specialist. https://web.archive.org/web/20230607092454/http://improvealexaranking.com/ (2023)
4. Net Systems Research. https://web.archive.org/web/20230219081618/http://netsystemsresearch.com/ (2023)
5. AdbeatBot. https://www.adbeat.com/policy (2024)
6. AdsBot. https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers#adsbot (2024)
7. AhrefsBot. http://ahrefs.com/robot/ (2024)
8. BingBot. https://www.bing.com/webmasters/help/which-crawlers-does-bing-use-8c184ec0 (2024)
9. Bloglines. http://www.bloglines.com (2024)
10. CensysInspect. https://about.censys.io/ (2024)
11. CheckMarkNetwork. http://www.checkmarknetwork.com/spider.html (2024)
12. Chrome UX Report. https://developer.chrome.com/docs/crux/ (2024)
13. Cloudflare Radar Domain Rankings. https://radar.cloudflare.com/domains (2024)
14. Dataprovider.com. https://www.dataprovider.com/ (2024)
15. Dnsthingy. https://www.dnsthingy.com/ (2024)
16. ev-crawler. https://headline.com/legal/crawler (2024)
17. Fofa. https://fofa.info/ (2024)
18. GoogleBot. http://www.google.com/bot.html (2024)
19. IBM X-Force Exchange. https://exchange.xforce.ibmcloud.com/ (2024)
20. Let's Encrypt. https://letsencrypt.org/ (2024)
21. MaxMind. https://www.maxmind.com/ (2024)
22. Mirror, Mirror, on the Wall, Who's the Fairest (website) of Them all? https://web.archive.org/web/20240124161803/https://www.domaintools.com/resources/blog/mirror-mirror-on-the-wall-whos-the-fairest-website-of-them-all/ (2024)
23. NameSilo. https://www.namesilo.com/ (2024)
24. Netcraft Web Server Survey. https://www.netcraft.com/ (2024)
25. Nicecrawler. http://www.nicecrawler.com/ (2024)
26. nltk.corpus Package. https://www.nltk.org/api/nltk.corpus.html (2024)
27. Nuclei. https://nuclei.projectdiscovery.io/ (2024)
28. Pandalytics. https://domainsbot.com/pandalytics/ (2024)
29. Quad9. https://www.quad9.net/ (2024)
30. Rankboostup. https://rankboostup.com/ (2024)
31. SurdotlyBot. http://sur.ly/bot.html (2024)
32. The Majestic Million. https://majestic.com/reports/majestic-million (2024)
33. VirusTotal. https://www.virustotal.com/ (2024)
34. Vultr Cloud Hosting. https://www.vultr.com/ (2024)
35. WellKnownBot. https://well-known.dev/about/#bot (2024)
36. YandexBot. https://yandex.com/support/webmaster/robot-workings/robot.html (2024)
37. Ahmad, S.S., Dar, M.D., Zaffar, M.F., Vallina-Rodriguez, N., Nithyanand, R.: Apophanies or Epiphanies? How Crawlers Impact our Understanding of the Web. In: The World Wide Web Conference (WWW) (2020)
38. Alexa: How do I get my site's metrics Certified? https://web.archive.org/web/20211127215835/https://support.alexa.com/hc/en-us/articles/200450354-How-do-I-get-my-site-s-metrics-Certified- (2021)

39. Alexa: We will be retiring Alexa.com on May 1, 2022 (2021), https://web.archive.org/web/20221126132843/https://support.alexa.com/hc/en-us/articles/4410503838999
40. Alexa: Alexa Top 1 Million. https://web.archive.org/web/20220701000000*/https://s3.amazonaws.com/alexa-static/top-1m.csv.zip (2022)
41. Cisco: OpenDNS (2024), https://www.opendns.com/
42. Cisco: Umbrella Popularity List. http://s3-us-west-1.amazonaws.com/umbrella-static/index.html (2024)
43. EdOverflow, Shafranovich, Y.: security.txt. https://securitytxt.org/ (2024)
44. Gharaibeh, M., Shah, A., Huffaker, B., Zhang, H., Ensafi, R., Papadopoulos, C.: A Look at Router Geolocation in Public and Commercial Databases. In: Internet Measurement Conference (IMC) (2017)
45. Giles, C.L., Sun, Y., Councill, I.G.: Measuring the Web Crawler Ethics. In: The World Wide Web Conference (WWW) (2010)
46. Giotsas, V., Smaragdakis, G., Dietzel, C., Richter, P., Feldmann, A., Berger, A.: Inferring BGP Blackholing Activity in the Internet. In: Internet Measurement Conference (IMC) (2017)
47. Google: Feedfetcher. https://developers.google.com/search/docs/advanced/crawling/feedfetcher (2024)
48. Google: Introduction to robots.txt. https://developers.google.com/search/docs/advanced/robots/intro (2024)
49. Halvorson, T., Der, M.F., Foster, I., Savage, S., Saul, L.K., Voelker, G.M.: From. academy to. zone: An Analysis of the New TLD Land Rush. In: Internet Measurement Conference (IMC) (2015)
50. Khan, M.T., DeBlasio, J., Voelker, G.M., Snoeren, A.C., Kanich, C., Vallina-Rodriguez, N.: An empirical analysis of the commercial vpn ecosystem. In: Internet Measurement Conference (IMC) (2018)
51. Krebs, B.: Does your organization have a security.txt file? https://krebsonsecurity.com/2021/09/does-your-organization-have-a-security-txt-file/ (2021)
52. Kuchhal, D., Li, F.: Knock and Talk: Investigating Local Network Communications on Websites. In: Internet Measurement Conference (IMC) (2021)
53. Li, V.G., Dunn, M., Pearce, P., McCoy, D., Voelker, G.M., Savage, S.: Reading the tea leaves: A comparative analysis of threat intelligence. In: USENIX Security Symposium (2019)
54. Majestic: Majestic Million – Reloaded! (2011), https://blog.majestic.com/company/majestic-million-reloaded/
55. Majestic: Majestic Million now free for all (2012), https://blog.majestic.com/development/majestic-million-csv-daily
56. Mendoza, A., Chinprutthiwong, P., Gu, G.: Uncovering HTTP header inconsistencies and the impact on desktop/mobile websites. In: The World Wide Web Conference (WWW) (2018)
57. Netcraft: Most Visited Websites. https://trends.netcraft.com/topsites (2024)
58. OpenDNS: PhishTank. http://phishtank.org/ (2024)
59. Oprea, A., Li, Z., Norris, R., Bowers, K.: Made: Security analytics for enterprise threat detection. In: Annual Computer Security Applications Conference (ACSAC) (2018)
60. Peng, P., Yang, L., Song, L., Wang, G.: Opening the Blackbox of Virustotal: Analyzing Online Phishing Scan Engines. In: Internet Measurement Conference (IMC) (2019)

61. Pochat, V.L., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., Joosen, W.: Tranco: A Research-oriented Top Sites Ranking Hardened against Manipulation. In: Network and Distributed System Security Symposium (NDSS) (2019)
62. Pot, J.: How to Find the RSS Feed URL for Almost Any Site. https://zapier.com/blog/how-to-find-rss-feed-url/ (2019)
63. PumpkinSeed: I Wanted to Play and I Got Hacked. https://medium.com/swlh/i-wanted-to-play-and-i-got-hacked-e5314fd5b27f (2021)
64. Ruth, K., Kumar, D., Wang, B., Valenta, L., Durumeric, Z.: Toppling top lists: Evaluating the accuracy of popular website lists. In: Internet Measurement Conference (IMC) (2022)
65. Rweyemamu, W., Lauinger, T., Wilson, C., Robertson, W., Kirda, E.: Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research. In: International Conference on Passive and Active Network Measurement (PAM) (2019)
66. Rweyemamu, W., Lauinger, T., Wilson, C., Robertson, W., Kirda, E.: Getting Under Alexa's Umbrella: Infiltration Attacks Against Internet Top Domain Lists. In: International Conference on Information Security (ISC) (2019)
67. Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S.D., Vallina-Rodriguez, N.: A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In: Internet Measurement Conference (IMC) (2018)
68. Spaventa, L.: Behind the Scenes Part II: What Makes HARO So Successful in Getting You Media Coverage. https://www.cision.com/blogs/2012/10/behind-the-scenes-part-ii-what-makes-haro-so-successful-in-getting-you-media-coverage/ (2012)
69. Umbrella, C.: Cisco Umbrella 1 Million (2016), https://web.archive.org/web/20230218105309/https://umbrella.cisco.com/blog/cisco-umbrella-1-million
70. Vallina, P., Le Pochat, V., Feal, Á., Paraschiv, M., Gamba, J., Burke, T., Hohlfeld, O., Tapiador, J., Vallina-Rodriguez, N.: Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services. In: Internet Measurement Conference (IMC) (2020)
71. Wan, G., Izhikevich, L., Adrian, D., Yoshioka, K., Holz, R., Rossow, C., Durumeric, Z.: On the Origin of Scanning: The Impact of Location on Internet-wide Scans. In: Internet Measurement Conference (IMC) (2020)
72. Xie, Q., Tang, S., Zheng, X., Lin, Q., Liu, B., Duan, H., Li, F.: Building an Open, Robust, and Stable Voting-Based Domain Top List. In: USENIX Security Symposium (2022)
73. Zeber, D., Bird, S., Oliveira, C., Rudametkin, W., Segall, I., Wollsén, F., Lopatka, M.: The Representativeness of Automated Web Crawls as A Surrogate for Human Browsing. In: The World Wide Web Conference (WWW) (2020)
74. Zhao, B., Ji, S., Lee, W.H., Lin, C., Weng, H., Wu, J., Zhou, P., Fang, L., Beyah, R.: A large-scale empirical study on the vulnerability of deployed IoT devices. IEEE Transactions on Dependable and Secure Computing (TDSC) (2020)

A Top List Manipulation

Here we detail the top list manipulation methods we employed.

Alexa. We apply the Alexa manipulation method recently developed by Xie et al. [72], which leverages Alexa's Certify service. Xie et al. observed that beyond collecting web traffic telemetry from its browser extension, Alexa also collects

data from its paid Certify service [38]. For websites subscribing to the Certify service, they embed a JavaScript snippet[4] provided by Alexa on their own web pages, which uploads visitor information to an Alexa data collection endpoint. Xie et al. identified that the telemetry sent to Alexa differentiated users with a single ID field that could be modified arbitrarily to forge fake visits by distinct users to the site. Furthermore, Alexa did not apply rate limits to the collected telemetry. A single IP address could generate a large volume of visitor telemetry that appeared to represent distinct user visits. As a result, Alexa would more highly rank a domain as it receives more distinct visitors. Thus, to manipulate an experiment domain into the Alexa ranking, we subscribed that site to the Certify service, and then applied this same technique of generating data telemetry to Alexa's data collection endpoint with distinct visitor ID values.

Umbrella. We use IP spoofing for Umbrella manipulation, as its ranking is constructed with PDNS and heavily depends on the number of IP addresses issuing DNS lookups for a domain [72,61,66]. Spoofing only addresses within our institution's local network (ethical considerations are discussed in Section 3.5), we generate DNS A record requests to Umbrella's DNS resolvers for our manipulated domains from different source IP addresses, causing Umbrella to more highly rank our domains due to higher request volume and IP diversity.

Majestic. We apply the Majestic manipulation method of Le Pochat et al. [61], which uses reflecting sites (described shortly). Majestic ranks a domain based on the IP subnet diversity of other websites linking back to the domain. It collects data on these website backlink relationships through regular large-scale web crawls. The authors identified that certain *reflecting* sites, particularly MediaWiki sites, will accept user-provided URLs as values in the site's URL query parameters and embed (reflect) these user-provided values as anchor elements in their webpages. When Majestic's crawler analyzes such reflecting links with a target domain provided as the URL query value, it observes a backlink to the target domain.

To trigger Majestic's crawler to visit certain links, we subscribe to Majestic's online service that accepts submitted URLs for crawling. To find reflecting sites, we use the set of reflecting wiki sites used by Le Pochat et al. [61]. We also crawled 12,920 MediaWiki-related domains from the Fofa search engine [17], checking whether a URL provided as a URL query value is reflected in the returned webpage. In total, we found 1,642 wiki pages that reflected URLs. For manipulating our test domains into Majestic, we submitted wiki links that reflected URLs to the Majestic service, with our test domain provided as the links' URL query value.

B Survey of Top List Use in Academic Papers

We evaluate the use of our three investigated top lists in academic research by surveying research papers published at 10 networking and security-related

[4] https://web.archive.org/web/20220322000324/https://certify-js.alexametrics.com/atrk.js

venues from 2017 to 2022, choosing the same set of venues previously surveyed by Scheitle et al. [67]:

- Measurement (3): ACM IMC, PAM, and TMA.
- Security (4): USENIX Security, IEEE S&P, ACM CCS and NDSS.
- Systems (2): ACM CoNEXT and ACM SIGCOMM.
- Web Technology (1): WWW.

To do the survey, we searched all papers published in the 10 venues from 2017 to 2022 for keywords such as "top list", "toplist", "Alexa", "Umbrella", "Cisco", and "Majestic". We manually reviewed the matching papers and counted the studies that relied upon a subset of a top list as input for part of the study, such as for measurements, data analysis, or experiment deployments. If a study used multiple top lists, we count them separately for each list used.