Getting Under Alexa's Umbrella: Infiltration Attacks Against Internet Top Domain Lists

Walter Rweyemamu, Tobias Lauinger,
Christo Wilson, William Robertson, and Engin Kirda

Northeastern University, Boston, MA
walter@iseclab.org

Abstract Top domain rankings such as Alexa are frequently used in security research. Typical uses include selecting popular websites for measurement studies, and obtaining a sample of presumably "benign" domains for model training or whitelisting purposes in security systems. Consequently, an inappropriate use of these rankings can result in unwanted biases or vulnerabilities. This paper demonstrates that it is feasible to infiltrate two domain rankings with very little effort. For a domain with no real visitors, an attacker can maintain a rank in Alexa's top 100 k domains, for instance, with seven fake users and a total of 217 fake visits per day. To remove malicious domains, multiple research studies retained only domains that had been ranked for at least one year. We find that even those domains contain entries labelled as malicious. Our results suggest that researchers should refrain from using these domain rankings to model benign behaviour.

1 Introduction

Many security researchers rely on "top site" rankings [26] such as the lists compiled by Alexa [2] and Umbrella [4]. For example, researchers use domains from these lists to train or evaluate proposed security systems, or they whitelist ranked domains to improve classifier performance [7, 9, 13, 17, 18, 21]. In doing so, they assume that the "most popular" domains are benign.

This assumption is problematic because prior research has shown evidence for malicious domains in Alexa's ranking [19,20,23]. Some researchers have taken additional precautions to address this concern, such as checking whether domains are blacklisted [9,13], or retaining only domains that have been ranked for long time periods [7,18,24]. To date there is no consensus on which method should be used, and we are unaware of any study that has investigated whether the latter method effectively removes malicious domains.

In the first part of this paper, we survey how often domains have been ranked in Alexa and Umbrella over the course of one year, and contrast their presence with their blacklist status. We find that even Alexa's top 10 k contains a domain labelled as malicious, but consistently ranked year-round. The full Alexa list contains 27 malicious domains (Umbrella: 292) during the entire year. These results indicate that the duration of a domain's presence in a top ranking alone is not a reliable indicator for benignness.

While researchers might not trust the *domains* listed in the rankings because some of them are known to be malicious, a common assumption appears to be that the *ranking* itself is reliable. Under this assumption, rankings contain malicious domains because

these domains receive visits from real users, not because the ranking was manipulated. For example, Nadji et al. justify their whitelisting of the Alexa top 10 k by arguing that *"If an attacker is aware of our whitelisting strategy there is little room for abuse. For an attacker to abuse our whitelisting strategy to evade our analysis, they would have to commandeer and point a whitelisted domain to their malicious infrastructure"* [21]. However, in anecdotal reports users claim to successfully bolster their own website rank by faking visits [8, 10, 27], and a cursory exploration of list infiltration attacks in prior work [16, 26] cast doubt on how resilient these lists are to manipulation.

In the second part of this paper, we conduct a systematic study of list infiltration attacks for both Alexa and Umbrella, and demonstrate that such attacks can be carried out with negligible resources. We find that maintaining a rank in the top 100 k domains requires approximately 217 requests per day from seven fake users for Alexa, and 24 k requests from spoofed source IP addresses for Umbrella, for domains that do not receive any real visitors. An Alexa rank of around 500 k could even be obtained *manually*, by a single user, by installing Alexa's toolbar and visiting 15–30 pages per day. As an illustration of the research impact of such attacks, our experimental domains with fake ranks have begun attracting crawler traffic, including various research crawlers from university networks.

Since we control all (fake) traffic to our experimental domains, we can quantify the extent of the *weekend effect*. This phenomenon, first mentioned by Scheitle et al. [26], is a temporary change in the rankings of Alexa and Umbrella that reoccurs every weekend, presumably due to different Internet traffic patterns compared to the workweek. To date, it is unknown how much Alexa domains with constant traffic change their ranks over time. We find that domains with constant (fake) traffic considerably improve their ranks during the weekend, such as from 448 k to 299 k, or from 88 k to 61 k in Alexa, and in Umbrella from 379 k to 230 k, or from 160 k to 72 k. Conversely, an Alexa rank of 84 k during the weekend requires roughly two fake users (62 fake URL visits) fewer than a similar rank during the workweek. This result implies that weekend ranks are based on less traffic, thus less reliable and more susceptible to fluctuation.

Overall, this paper sheds light on several aspects of top domain lists that researchers should account for when using these lists in their work. Specifically, we make the following contributions:

- We demonstrate infiltration attacks for Alexa and Umbrella where attackers add new domains to the rankings even though these domains do not receive any real visitors.
- Through controlled experiments, we measure the impact of (fake) traffic characteristics, and notably quantify the differences between weekday and weekend ranks.
- We analyse (real) web traffic to our experimental domains, and show that once ranked, domains start receiving regular visits from crawlers in various university networks.
- We are the first to assess a mitigation strategy against malicious domains in the rankings used in prior work, and find that it fails to fully eliminate malicious domains.

2 Background & Related Work

In this paper, we often refer to entries of rankings or lists, which can lead to confusion as to a "high" rank being good or bad. As a convention, a *higher* rank is a *better*, *numerically lower* rank, towards the top of the list with the most popular entries.

2.1 Use of Top Lists in Security Research

Top domain lists are frequently used in research as observed by Le Pochat et al. [16], who found 102 papers using the Alexa ranking at the four highest tier security conferences between 2015 and 2018. Furthermore, Scheitle et al. [26] reference 68 studies using the Alexa Top Sites published at the top measurement, security, and systems conferences in 2017 alone. Researchers typically use these rankings in one of two ways.

Designating the "largest" websites. Especially for measurement studies, researchers often seek to cover a representative set of websites so that their findings can be considered relevant with respect to the browsing habits of typical users [11, 15, 22, 28]. When researchers select domains for their popularity, it is less of a concern whether the domains are compromised or malicious. Similarly, while attackers might manipulate the ranks of their domains to make them appear more popular, this is likely not a major concern for measurement results aggregated over a large number of domains, as long as the extent of rank manipulation remains moderate relative to the frequency of the measured property.

Designating "benign" websites. Many security papers need labelled training and evaluation data for detection mechanisms. Some researchers resort to domain rankings and use popular domains as an approximation of "benign" websites. For example, Lever et al. [18] obtain the malicious domains contacted by malware samples by filtering out domains that have been present in the Alexa top 10 k for at least one year (except for several commonly abused dynamic DNS domains). Similarly, Rahbarinia et al. [24] detect malware control domains after labeling domains as benign when they have appeared in the Alexa top 1 M for one year. Alrwais et al. [7] study bulletproof hosting in AS sub-allocations and create a "noisy" set of benign allocations from domains that have been present in the Alexa top 50 k for two years. While these papers aim to reduce the likelihood of ranked domains being malicious by requiring them to be ranked for a long time period, we are not aware of any study showing that this is indeed a sound approach. Other papers such as EXPOSURE [9] or IceShield [13] vet ranked domains through blacklists. Unfortunately, many authors do not make such an effort. WarningBird, for example, whitelists the Alexa top 1 k "to reduce false-positive rates" of a URL classifier [17].

Several prior studies have reported evidence that malicious domains exist in the Alexa ranking. Li et al. [20] mention a fake antivirus campaign on a website ranked 2,404 on Alexa. Pitsillidis et al. [23] detect a 1–2 % overlap between blacklists and the Alexa top 1 M (even though they consider them false positives of the blacklists). Lever et al. report that "more than 100 ... domains were ranked in the top 10,000 by Alexa on the day they were added to the blacklist" [19].

2.2 List Compilation Methodology

In this paper, we consider two measurement-based top site lists: Amazon Alexa Top Sites [2], which is the most popular list in research [26], and Cisco Umbrella Top 1 Million [4], a more recent list ranking arbitrary (sub)domains instead of websites. Table 1 summarises the data sources and popularity model of these lists.

Table 1: Data Sources of Common Top Site Lists.

Ranking	Data Source	List Contents
Alexa	browser toolbar	typed-in website domains
Umbrella	DNS resolver	resolved (sub)domains

Alexa The data for the Alexa ranking originates primarily from "millions of users" [3] who have installed the Alexa browser toolbar and share their browsing history with Alexa. Its website documents Alexa's methodology as follows: The installed browser toolbar records all URLs that are visited from the address bar of the browser window, meaning that third-party resources such as advertisements or tracking code are ignored. Only blogs and personal homepage subdomains are ranked separately from the main domain. Domains are ranked according to a combination of the number of users visiting the site, and the unique URLs on that site visited by each user. While the ranking is updated daily, the (API) data is smoothed over a 3-month time window. Ranks below 100 k are not statistically meaningful because the data collected about those domains is too scarce [3, 5]. The ranking is available through an API, on the website and as a CSV download [1], with noticeable differences (Section 4.3).

Umbrella The Umbrella rankings are derived from incoming DNS lookups observed in Cisco's Umbrella Global Network and the OpenDNS service, which amount to over 100 B daily requests from 65 M users in 165 countries [4]. Consequently, the list reflects the popularity of domains used in any Internet protocol, not only web traffic. Umbrella states that ranks are based on unique client IPs looking up each domain [14]. However, our findings in Section 4.4 differ.

2.3 Related Work

Recently, Scheitle et al. [26] studied the contents and stability of Internet top domain lists such as Alexa and Umbrella. Additionally, they demonstrated rank manipulation with a successful attack against Umbrella, obtaining ranks of up to 30 k on a Friday, and 17 k on a Sunday using the same traffic characteristics. The authors attributed this rank difference to the *weekend effect*, that is, a decrease in traffic to other ranked domains during the weekend.

In prior work, we studied potential consequences of the weekend effect in Alexa and Umbrella, such as different country and website category distributions of the ranked domains on weekdays and the weekend [25]. Furthermore, we observed the presence of clusters of alphabetically ordered domains in Alexa and Umbrella, which we speculated to be due to these domains being considered equivalent in terms of observed traffic.

Le Pochat et al. [16] described multiple list infiltration attacks against various top domain lists. Regarding Alexa, Le Pochat et al. proposed two different attacks. Their first attack variant involved installing the Alexa browser toolbar in real browser instances, where they were able to obtain a rank of 345 k with only nine fake requests. The second attack variant targeted Alexa Certify, a paid service to directly measure website visits

using a tracking script provided by Alexa. Le Pochat et al. also studied whether domains ranked by Alexa and Umbrella are malicious according to Google Safe Browsing.

This paper contrasts and extends on prior work in the following ways:

- We extend Scheitle et al. by introducing an attack against Alexa.
- We are the first to analyse the crawler traffic received by domains with fake ranks.
- We improve the attack technique against Alexa by Le Pochat et al., where instead of installing the Alexa extension in real browser sessions, we submitted fake browsing traffic to Alexa's internal API, a more scalable approach. In contrast to Le Pochat et al., who do not mention some parameters of their attack, we document in detail the parameters involved in the attack such as the number of distinct URLs, and distinct fake users (AIDs), to explore their effect, and allow for comparisons with later work.
- We measure the magnitude of the weekend effect by comparing weekday and weekend ranks of experimental domains with identical amounts of (fake) traffic.
- We experimentally confirm the hypothesis that the alphabetically ordered domains are equivalent in terms of observed traffic, as speculated by our earlier work.
- We extend the malicious domain analysis by Le Pochat et al., who considered only a single snapshot of the rankings, by investigating how long malicious domains remain ranked, and whether all domains ranked for one year are benign.

3 Domain Longevity & Maliciousness

To obtain a set of *benign* domains, several researchers have selected domains ranked by Alexa for one or more years [7, 18, 24]. The rationale behind this approach is that malicious domains are often active for only a few days before they are blacklisted [12].

3.1 Longevity of Ranked Domains

We begin our analysis of this strategy by studying how often domains appear in the ranking. This analysis is based on ranking CSVs downloaded from Alexa and Umbrella each day for a duration of one year, beginning with the ranking for 14 October 2017. While either ranking contains exactly 1 M domains each day, over the 365 days, Alexa included a total of 24 M unique domains (Figure 1a), and Umbrella over 7 M domains (Figure 1b). This implies that the rankings are very unstable. A large portion of the domains remain ranked for a short time only, before being replaced with new domains. For example, Figures 1c and 1d show that only 6.1 % and 20.3 % of Alexa and Umbrella domains, respectively, were listed on more than 50 (not necessarily consecutive) days. An implication of this instability is that fewer than 93 k domains in Alexa, and just over 303 k domains in Umbrella, were ranked consistently every day over the one-year period. Over 90 % of Alexa list entries, and almost 70 % of Umbrella entries on any given day will leave the ranking at least once within one year.

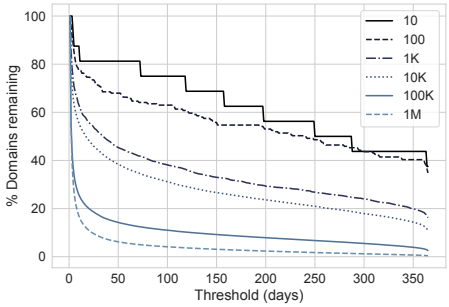
Note that several years before our study, Rahbarinia et al. [24] found a much larger number of 459 k domains had been present in the Alexa ranking during 365 consecutive days. We believe that this is due to a change in Alexa's ranking in January 2018, first reported by Scheitle et al. [26]. Before that date, presumably due to smoothing, Alexa's

(a) Alexa (absolute numbers)

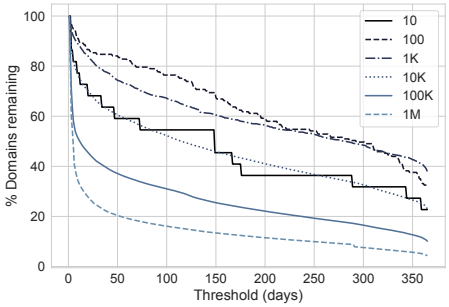
Prefix	All Domains		Malicious	
	total (∪)	**1 y** (∩)	**total**	**1 y**
10	16	6	0	0
100	181	63	0	0
1 k	2,972	483	36	0
10 k	36,679	3,935	493	1
100 k	1,005,275	25,708	15,907	13
1 M	24,161,278	92,832	65,755	27

(b) Umbrella (absolute numbers)

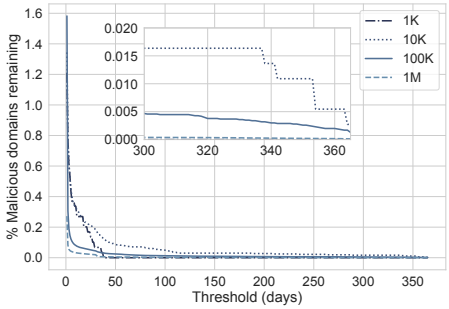
Prefix	All Domains		Malicious	
	total (∪)	**1 y** (∩)	**total**	**1 y**
10	22	5	0	0
100	157	51	0	0
1 k	1,670	634	0	0
10 k	22,207	5,089	17	0
100 k	386,493	39,074	858	43
1 M	7,065,560	303,057	34,974	292



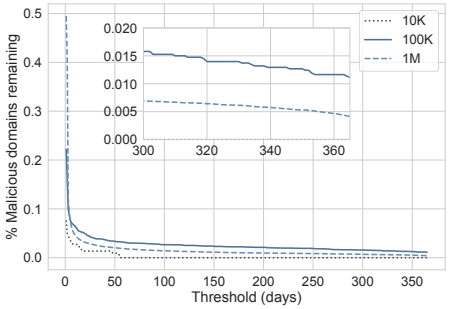
(c) Alexa (all domains, relative to ∪)



(d) Umbrella (all domains, relative to ∪)



(e) Alexa (malicious domains, relative to ∪ of all domains)



(f) Umbrella (malicious domains, relative to ∪ of all domains)

Figure 1: Number of days domains were ranked by Alexa and Umbrella in the year since 14 October 2017. Out of the 1 M domains that appeared in the Alexa top 100 k, only 25.7 k were ranked in this prefix for the entire year. Requiring domains to be ranked consistently for one year removes a majority of malicious domains, but 27 in Alexa, and 292 in Umbrella remain.

ranking was relatively stable. During our experiments (Section 4.3), we found that as of late 2018, Alexa was not applying any smoothing to the ranks found in the CSV download, resulting in a less stable ranking.

Often, researchers use only a list prefix such as the top 10 k or 100 k instead of the full ranking. Furthermore, Alexa cautions that ranks below 100 k are not statistically meaningful [3, 5]. Figures 1a and 1b show that shorter list prefixes are increasingly more stable. Around 48 % of the Alexa top 1 k, and 63 % of the Umbrella top 1 k domains, for instance, were ranked in the top 1 k every day. An exception are the Umbrella top 10 and top 100, which are less stable than the top 1 k, or the corresponding list prefixes in Alexa. This is due in part to the weekend effect, which causes domains that are much more popular during the weekend to enter the shorter list prefixes and displace other domains. As a result, neither of these domains is present in the short list prefixes during 365 consecutive days.

3.2 Maliciousness of Ranked Domains

To determine the maliciousness of ranked domains, we looked up their status in Google Safe Browsing in the last week of October 2018. ' maliciousness after the year had already elapsed, we do not know about their maliciousness at the time they were ranked. For example, our methodology does not detect domains that were temporarily compromised and subsequently cleaned up. Furthermore, we cannot distinguish compromised domains from those that are intentionally malicious. Yet, our methodology models the strategy used by researchers who first compile a list of presumably "benign" domains and then collect data from these domains, such as downloading "benign" websites and extracting features for model training. In this scenario, it is critical that these domains not be malicious at the time data is collected from them.

In relative terms, a very small fraction of domains that were ranked on any of the 365 days are labelled as malicious – 65,755 out of 24 M domains in Alexa, and 34,974 out of 7 M domains in Umbrella. The top 100 in Alexa, and the top 1 k domains in Umbrella do not contain any domain labelled as malicious. Malicious domains do exist in the Alexa top 1 k and the Umbrella top 10 k, but they appear on no more than 37 and 54 days, respectively. All longer list prefixes contain malicious domains even among those that were ranked every day for one year. Out of the almost 93 k domains in Alexa that were ranked every day, 27 are malicious according to Safe Browsing; they were all marked as "social engineering" or "unwanted software." However, Alexa contained six "malware" domains ranked for over 300 days, out of which two were just one and two days away from the 365-day threshold. In Umbrella, 292 out of 304 k domains ranked the entire year were labelled as malicious: 231 as "unwanted software," 33 as "malware," and 28 as "social engineering."

The ratio of consistently ranked malicious domains over all malicious domains is lower than the ratio of consistently ranked domains over all domains, which suggests that malicious domains leave the ranking faster than benign domains. Yet, at a time scale of one year, the strategy of retaining only domains consistently ranked for a longer time period reduces, but does not completely eliminate malicious domains. A small number of malicious domains may be acceptable in some scenarios, such as when aggregating over large numbers of domains. However, mislabelled training data in a machine learning context (i.e., a few examples labelled as benign despite being malicious) could have a disproportionate effect on classifier performance. Another issue is that requiring domains to be ranked continuously significantly decreases the number (and diversity) of domains,

as a domain's absence for a single day, or during each weekend, would cause it to be eliminated from the final set of domains.

4 Infiltration Attacks

Both Alexa and Umbrella exhibit strong weekend effects, visible in different domains being popular during the weekend as opposed to the workweek [25, 26]. These periodic changes suggest that changes in observed traffic have a direct and immediate impact on the ranking. Consequently, it may be possible for attackers to manipulate the ranks of existing domains, or to infiltrate the lists with new domains. We ran controlled experiments with the primary goal of showing that such attacks are indeed feasible. Since we controlled the (fake) traffic to our domains, we were able to quantify the extent of the weekend effect by comparing weekday and weekend ranks. Further, we explored the effect of various attack parameters, especially when they have an influence on the cost of the attack. We did so by running multiple independent experiments in parallel, one with a reference domain, and additional experiments with separate domains that vary one attack parameter each. We used newly registered domains to avoid any bias due to prior activity. We also created several control domains that were registered but not used in any experiment. To observe the effect of being ranked, we logged incoming web requests on our domains.

4.1 Ethical Considerations

Since our experiments involved domain lists that were in active use, we needed to consider and minimise potential risks due to our activities. To that end, we carefully designed a research protocol prior to starting our experiments.

The main risk was that consumers of the lists would receive invalid data if our experiments were to succeed. We reduced the impact of this risk in the following ways:

- *Limit the number of fake domains used concurrently.* At any time, each ranking contained no more than ten of our domains. This is a negligible fraction compared to the full one million entries of each list.
- *Limit the maximum rank we attempt to achieve.* Since Alexa cautions that the bottom 900 k ranks are not statistically meaningful [3, 5], we need to infiltrate the top 100 k domains to show that the attack can result in a significant rank. However, once a domain crosses that threshold both during the workweek and the weekend, we do not seek any higher rank. Our highest ranks were around 60 k (during the weekend), and we never had more than three domains ranked in the top 100 k at the same time. Our experiments barely impact consumers of the most popular domains.
- *Limit the duration of the experiments.* Due to the strong weekend effect, and to quantify natural rank fluctuation, we need to test each attack parameter for at least one week. Once a stable attack parameter has been found and confirmed, we end the experiment and the domain disappears from the ranking within one or two days, in line with the fast responsivity of the lists. As an exception, we maintained a 200 k Alexa rank for one reference domain that we used to explore long-term effects, convergence between the Alexa rank shown on the website and in the CSV file, and to observe website crawling during an extended period of seven months.

– *Use newly-registered domains under our control.* The experimental domains and their mostly empty websites do not harm any potential visitor. The experiments are not aimed at directing human visitors to our websites. This probability is minuscule since the domains only appear in the ranking among a million other sites and are not advertised elsewhere. Neither we nor any third party unduly benefits from the fake ranks of our experiments.

Our approach involved sending fake data to Alexa and Umbrella. We do not consider overloading their systems as a major risk, as those systems are designed to handle very large numbers of users. For example, Umbrella reports a total of 100 B requests per day [4], whereas our experiments never exceeded more than 42 k requests per day. Similarly, Alexa claims millions of toolbar users [3], and we simulated no more than a dozen daily toolbar users with moderate browsing behaviour. To err on the side of caution, we perform our experiments in an open way from IP ranges in our institutional network. We place a message with contact information on our experimental domains, but have not received any inquiries.

We did not seek IRB approval because our experiments do not involve human data, and our IRB does not review ethics beyond human subjects research.

We strictly followed this protocol throughout our experiments. Given these precautions, we believe that any short-term risks are outweighed by the long-term benefit of showing that the lists can be manipulated with little effort. Furthermore, by raising awareness for the limitations of the lists, our findings may prevent future harm to consumers of the lists.

4.2 Alphabetically Sorted Clusters

Our earlier study showed that the Alexa and Umbrella rankings contained long sequences of alphabetically sorted domains [25]. When considering any sequence of at least 42 alphabetically sorted domains as a cluster, more than 54 % of list entries in Alexa, and more than 91 % in Umbrella were part of such a cluster.

We hypothesised that these clusters correspond to domains that the list publishers cannot distinguish based on their traffic characteristics. Our experiments support this hypothesis, as domains with identical fake traffic were ranked in the same cluster. Furthermore, in Umbrella, subdomains appear to cause their parent domain to be ranked, too. Since our experiments involved only one subdomain per parent domain, and we did not fake any visits to the parent domain, both the subdomain and the parent domain always appeared in the same cluster.

In our experiments, we take advantage of clustering in two ways. First, if two domains with different traffic parameters appear in the same cluster, we know that their traffic is considered equivalent by the list publisher, and the different parameter is likely irrelevant. Second, inside each cluster, the position of a domain is determined only by its lexicographical ordering. This means that it is possible to place a domain at the beginning of the cluster, and thus obtain a minor improvement of the domain's rank, by selecting a name beginning with zeroes. This also reduces the rank distance between similar experimental domains, and makes our tables easier to read.

4.3 Alexa

Alexa primarily collects data from users who install the Alexa toolbar in their browser and give consent to share their browsing history with Alexa. The presumably most straightforward approach for attackers would be to install the toolbar in real browsers and use automation tools to create fake browsing sessions. However, this approach is somewhat expensive to scale, and it is more complicated to vary variables such as screen size and network delays that are collected by Alexa. Another option is to reverse-engineer the browser toolbar, understand its data collection and communication behaviour, and use its internal remote APIs to send fake toolbar traffic to Alexa, without actually visiting any website. We pursue the latter approach for better control of experimental conditions.

When the toolbar is first installed, it requests a new user identifier (AID) from Alexa's servers, and the user must consent to the data collection. We did not automate this process, as only a limited number of AIDs were necessary for our experiments. Instead, we generated AIDs manually and extracted the identifier as well as cookies from the browser profile. When active, the toolbar downloads configuration from Alexa and sends a request with metadata each time the URL in the address bar changes. The data sent to Alexa includes the current and previous URL, the page load time, response status code, the window and tab IDs, a request counter, the screen resolution and the browser window width. While the toolbar collects additional information, it does not seem related to website ranking, thus we do not investigate further. Simply recording an API message and replaying it multiple times does not result in a rank, as Alexa appears to do semantic checks. Therefore, we implemented a script that emulates the toolbar's communication behaviour by increasing counters as necessary, and randomising fields such as the page load time. From Alexa's website, we gather that both the number of users and the number of unique pages visited on a domain may influence its rank. We implement our script such that it can emulate browsing sessions consisting of visits to a predefined list of pages on our experimental domain, optionally interleaved with fake visits to unrelated non-experimental websites. Fake visits consist in data being submitted to Alexa's toolbar API. We do not connect to any of these "visited" domains.

For our experiments, we use newly registered domains with a website that contains only a brief sentence with contact information. We run multiple experiments in parallel, one as a baseline, and others where we vary different parameters to observe their effect. The parameters we consider are the number of users (AIDs), the number of unique pages "visited" on our websites, whether the fake browsing session includes any visits to non-experimental websites, and the number of browsing sessions per user. To create lists of pages to visit on our website, we concatenate random dictionary words to simulate a directory structure; these pages do not actually exist. When an experiment calls for the inclusion of visits to non-experimental websites, we pick domains from the Alexa top 100. Our limited experiments are unlikely to have a noticeable effect on the ranking of domains that are already highly popular.

Attack Parameters In our experiments until end of June 2018, we found that the number of identical browsing sessions did not matter; one or two fake visits per unique URL had the same effect as twenty repetitions, provided all other parameters were the same. In fact, when the number of repetitions was too high, the domain lost its rank. However, Alexa

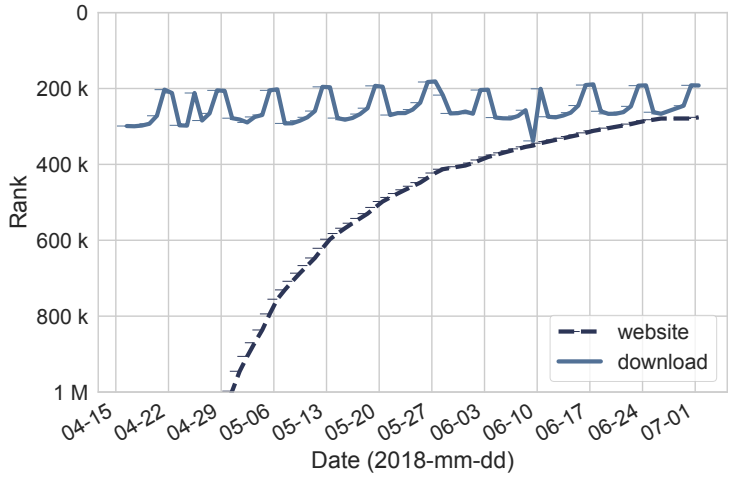


Figure 2: Alexa ranks for the baseline domain (two AIDs, 21 pages). The rank on the website appears to be smoothed; it starts at 7.2 M and slowly converges to the more immediate rank from the CSV file download. Weekend ranks, with constant fake traffic, are around 74 k better than workweek ranks.

does not appear to permanently block traffic from the associated source IP addresses or AIDs. Similarly, interleaving fake visits to the experimental domain with visits to non-experimental websites did not seem to have any effect on whether or not the experimental domain could obtain a rank. A working website was not a requisite either, as submitting fake visits for a domain without any DNS A record resulted in a normal rank.

Since summer 2018, several details appear to have changed. In our more recent experiments, two repetitions per unique URL were necessary in order to *obtain* a rank for a new domain for the first time, whereas a single visit per URL was sufficient to *maintain* the rank on the following days. Similarly, our newer experiments required us to interleave fake visits to our experimental domain with approximately half as many fake visits to other domains. We do not know whether these changes happened in response to the disclosure of the attack by Le Pochat et al. [16]. With minimal changes, attacks can still be carried out successfully.

We use as the baseline a domain with fake traffic from two AIDs visiting 21 pages each (the front page and twenty deeper pages). With constant fake traffic, such a domain reaches a median rank of 270 k during the workweek, and a median rank of 196 k during the weekend, as shown in Figure 2. The ranks found in Alexa's CSV file appear to be immediate and have little long-term variation. The same domain's rank on Alexa's website starts at 7.2 M and gradually approaches the rank from the CSV file; it appears to be smoothed. Alexa's API returns values similar to those found on the website.

Visits to more unique pages on the experimental domain have a moderate effect on the domain's rank. In our experiment, increasing the baseline from 21 to 41 page visits resulted in a slightly better rank than the baseline, 241 k during the week instead of 263 k.

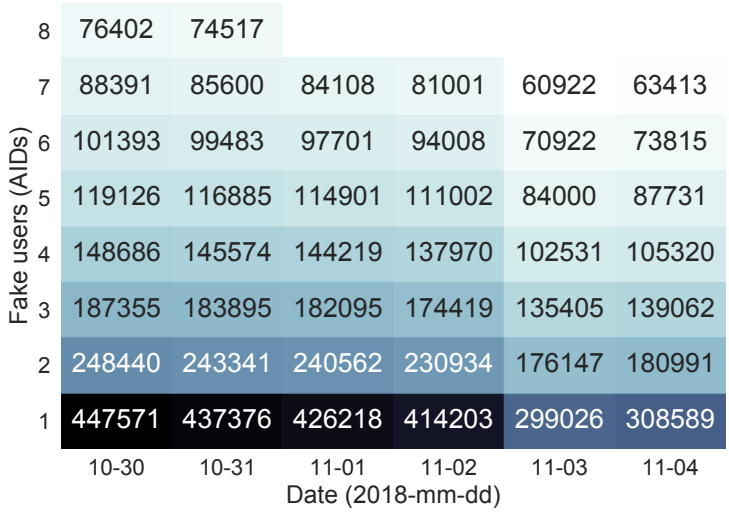


Figure 3: The impact of fake users (requesting 21 pages each) on the Alexa rank. Dates refer to when the fake traffic was sent. A rank in the top 100 k requires visits from 6–7 fake users during the workweek, and 5 users on the weekend. A *single* toolbar user's traffic causes ranks in the top 500 k.

Further increasing to 81 page visits yielded a rank of 220 k instead of the same-day baseline rank of 259 k.

In comparison, emulating more unique toolbar users (AIDs) has a much higher impact on the rank. To explore this effect in detail, we designed a separate experiment. We created eight domain names to be "visited" by one to eight AIDs. Each domain had a long prefix of zeroes so that it would be placed on top of its respective alphabetically sorted cluster in the ranking (see Section 4.2). We used all AIDs from a single IP address, but in sequence such that only one AID was active at any time. Each AID ran between one and eight sessions, where each session consisted of visits to 21 pages on the respective experimental domain, as well as an average of 10 pages on unrelated domains. Initially, each session consisted of two repetitions, and we later reduced them to one repetition without any discernible impact on the ranks. We experimentally determined that Alexa used midnight UTC as the cutoff time for rank computations, thus we scheduled our experiments accordingly. The duration of a single session with one repetition was approximately ten minutes, plus random delays between sessions.

The domain visited by two AIDs had the same effective settings as the baseline domain, and resulted in comparable ranks, as shown in Figure 3. Using only one AID resulted in a rank of about 448 k, which means that a domain can be ranked in Alexa's top 500 k based on traffic from *a single user*. This suggests that the data used by Alexa to compute the lower ranks is quite fragile. The high volatility of the ranking underscores this issue (Section 3.1).

A domain visited by eight users achieves a weekday rank of around 75 k. Data for this domain is only available for two days because we disabled this part of the experiment when the high rank appeared in Alexa's ranking, following our guidelines from Section 4.1. A rank in the top 100 k, which are considered more reliable by Alexa [3, 5], requires between five and seven users, depending on the day. Their fake visits correspond to 155–217 API requests, made from a single IP address in one to two hours. None of these requirements represent any significant cost for an attacker.

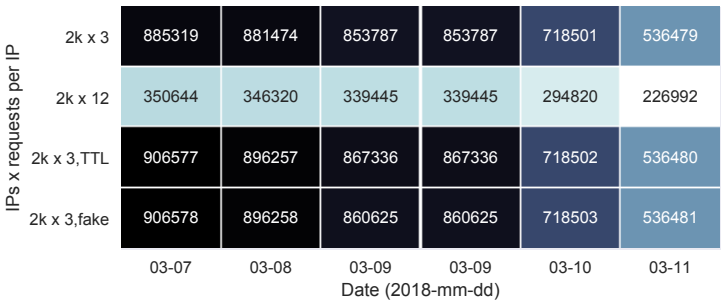
4.4 Umbrella

The Umbrella ranking is derived from DNS lookups using the OpenDNS resolver. Since the resolver is available for public use, attackers can repeatedly look up their own domain to make it appear more popular than it actually is. A blog post about the ranking suggests that the number of source IP addresses may influence the rank more than the total number of lookups [14]. To test this hypothesis experimentally, we obtained permission to utilise four unused network prefixes of our institution for outgoing DNS lookups, totalling 24.5 k possible source IP addresses. We run multiple experiments in parallel with disjoint sets of source IPs and independent, fresh domain names to observe under controlled circumstances the influence of several parameters on the domain rank. We spread out the lookups of each experiment evenly over the full day and deliberately choose parameters such that for all parallel experiments combined, we do not exceed a limit of one lookup every two seconds. Given that OpenDNS reports 100 B daily requests [4], our maximum of 42,000 daily lookups is unlikely to threaten the stability of OpenDNS resolvers.

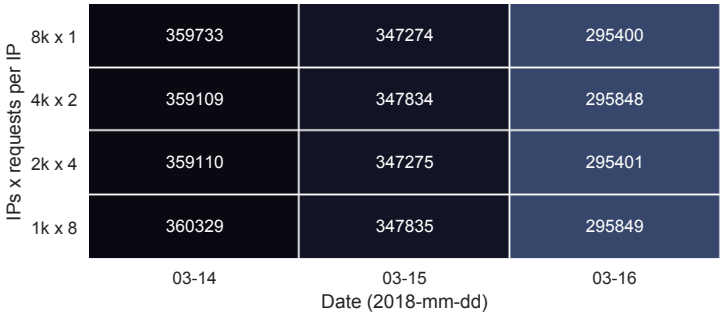
During a successful experiment, our domain appears in Umbrella's next daily update of the ranking. We do not know which cutoff time Umbrella uses to split their data stream into days. Our results suggest that the beginning and end of our experiments are not perfectly aligned with Umbrella's notion of a day, as the first and last days' ranks are always significantly worse than the ranks in between (which are based on 24 h of lookups). For this reason, and to observe "natural" rank fluctuations, we run our experiments with constant lookups for about one week. We then discard the ranks of the first and last day, retaining only those in between that we consider "stable".

Many domains in the ranking appear in alphabetically sorted clusters (Section 4.2). We leverage this characteristic to improve the ranks of our experimental domains by looking up a `0000000000` subdomain, which results in the subdomain being placed at the beginning of the cluster, and the parent domain further down in the alphabetical order of the same cluster. A secondary effect of this approach is that two list entries result from lookups of just one domain. We did not expand this technique to deeper subdomain levels so as not to unnecessarily pollute the list, but it is a possibility for attackers.

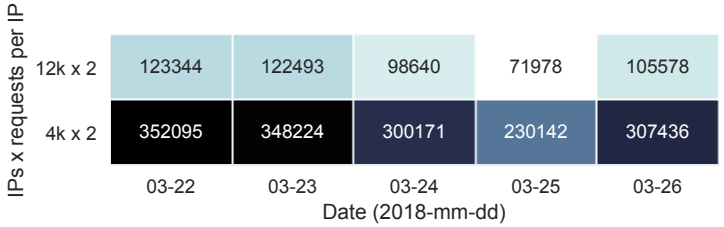
Attack Parameters After an initial exploratory experiment to find a successful combination of parameters that resulted in a (low) rank, we designed three subsequent experiments to determine how the size of the source IP pool, the number of lookups, the TTL of the domain's DNS A record, and the resolvability of the domain influence the resulting rank. In the first experiment, the baseline corresponds to 2 k source IPs each making three daily lookups, using a domain name that resolves to an A record with a 3 h TTL. As



(a) Effect of traffic parameters (top to bottom): Baseline, four times more lookups, longer DNS A record TTL, non-existing domain.



(b) Identical total lookups sent from varying number of source IPs.



(c) Rank in the top 100 k, and a reference domain for comparison.

Figure 4: Ranks obtained in the Umbrella experiments sending lookups to OpenDNS. Adjacent ranks correspond to domains in a single cluster. Minor rank differences for similar traffic parameters denote different clusters, likely due to sampling or packet loss. Two lookups sent from 12 k IP addresses result in a 123 k Thursday rank, and 72 k on Sunday.

shown in Figure 4a, looking up a domain name with a 12 h TTL, or a name that cannot be resolved, results in subdomain ranks comparable to the baseline, between 536 k and 907 k depending on the day of the week.

Contrarily, quadrupling the number of lookups from the same number of source IPs results in a higher rank between 227 k and 351 k.

This finding contradicts our intuition and public documentation [14] that the Umbrella ranking may be based on source IPs rather than lookup counts. To investigate, we designed another experiment with four conditions, a domain looked up once from 8 k source IPs, twice from 4 k IPs, four times from 2 k IPs, and eight times from 1 k IPs, respectively. All four domains achieved nearly identical ranks, as visible in Figure 4b. However, this does not mean that more lookups can be directly substituted for the same number of source IP addresses. In a separate experiment, 1 k source IPs making two lookups each did not result in a rank, whereas 2 k source IPs with one lookup did. Similarly, making two lookups each from 12 k source IPs resulted in a rank between 72 k and 123 k (Figure 4c), considerably better than the same total number of lookups using only 2 k source IPs (Figure 4a, 227 k to 351 k). We hypothesise that Umbrella's ranking is based on a potentially non-linear combination of source IPs and lookup counts, where source IPs have a higher weight. However, this is hardly a hurdle for attackers, as DNS is based on UDP, and source IPs can be spoofed. An attacker can place a domain in Umbrella's top 100 k with about 24 k daily DNS lookups, or potentially even fewer by spoofing more than the 12 k source IPs we had available in our experiment. The technical complexity and cost for an attacker, given a network location that does not filter spoofed source IPs, are very low.

4.5 Limitations

In our experiments, we refrained from pursuing ranks that were significantly higher than 100 k, in line with our ethical guidelines (Section 4.1). Prior work has demonstrated attacks for ranks as high as 31 k in Alexa (using a different technique [16]), and 17 k in Umbrella (using a similar technique [26]), thus it appears reasonable to assume that attackers could reach even higher ranks. To the best of our knowledge, all prior work (including our efforts) only showed that new domains can be added to the ranking, rather than manipulating the ranks of domains already present on the list. It is conceivable that list publishers treat long-term entries differently from new entries, though it seems unlikely, given the extent of change that we observed in the rankings.

4.6 The Weekend Effect

Prior work has shown that both Alexa and Umbrella periodically and temporarily change their composition during each weekend compared to the workweek. One manifestation of this phenomenon is that different sites are popular during the weekend [26]. However, we also show that the lists appear to be based on less traffic data during the weekend. Since our experiments use constant traffic parameters, any rank increase of the experimental domains from the workweek to the weekend must be due to other ranked domains receiving less traffic. All of our domains received better ranks on Saturday and Sunday. In general, this effect was more pronounced in the lower ranks. For example, our one-AID domain in Alexa increased its rank from 448 k on Tuesday to 299 k on Saturday, whereas the seven-AID domain only increased from 88 k to 61 k. Similarly, Umbrella ranked a domain with three lookups from 2 k IP addresses at 885 k on Wednesday, and at 536 k on Sunday, a difference of 349 k ranks, more than one third the length of the list.

For attackers, this means that rank manipulation is less costly during the weekend, as comparable ranks can be obtained with fewer resources. For example, weekend attacks

Table 2: Daily Web Requests to Experimental Domains While Ranked.

Domain/Experiment	Ranked	Min	Avg/day	Max
none (control)	0 days	0	0	0
Umbrella 100 k	7 days	1	1.9	4
Alexa 200 k (reference)	60 days	13	35.1	174
Alexa 100 k	9 days	10	27.8	42

at the operating points in Figure 3 typically require two fake users less than during the workweek.

We cannot determine whether these rank changes correctly mirror the extent of websites receiving fewer visitors during the weekend, or if they are amplified by fewer Alexa toolbar or OpenDNS users being active during the weekend. In either case, they show that the influence of a single user on the composition of the domain ranking grows during the weekend, rendering the ranking less reliable.

4.7 The Aftermath

To immediately observe the impact of a domain being ranked, we logged all web requests to our test domains. Before being ranked, most domains received no requests. We registered these domains many weeks before first using them, and observed only very rare visits from crawlers such as one likely associated with the DomainTools service. Presumably, these crawlers discovered the domains through `.pw` zone files, as our domains were not referenced elsewhere. Once ranked, our domains started receiving regular visits from crawlers. Table 2 shows that an Alexa rank has a much higher impact on crawler traffic than an Umbrella rank. Our domain in the Umbrella top 100 k received at most four web requests on any day, whereas the two Alexa domains received at least 10–13 requests each day, with a maximum of 174. By the end of June 2018, we had observed 125 distinct crawlers.

To estimate the number of distinct crawlers, we identify them by the Autonomous System (AS) of their IP address, which unlike user agent strings cannot be faked easily. We group together multiple ASes used by Amazon cloud services. This approach is clearly an underestimation of the number of crawlers, whereas counting crawlers by the number of distinct source IP addresses would likely overestimate their number due to IP changes in residential networks, and deliberate IP changes or IP pooling in cloud-based crawlers. Furthermore, this would complicate detection of recurring crawlers.

Table 3 shows a selection of crawlers identified by their AS name, and the types and periodicity of requests they made. Some crawlers visited our domains only once, close to the date the domains entered the ranking. These crawls included vendors of software security products, likely to assess the type and maliciousness of the websites. Our domains also received visits from crawlers evidently looking for potential vulnerabilities on our websites, such as unprotected configuration files, database backups, management scripts and vulnerable web applications. As these crawls came from residential access networks, we suspect they were not benign security surveys.

Table 3: Twelve out of 125 Web Crawlers Observed on the Experimental Alexa Reference Domain (until end of June 2018).
URLs visited: Homepage/robots.txt/experimental URLs/other URLs

Crawler/AS Name	First Delay	Active Days	Periodicity	Requests per visit	URLs visited
Google	1.7 days	59	daily	1.5	●/●/○/●
Amazon	1.8 days	47	appr. daily	5.4	●/●/●/●
SPRINT-SDC, PL	1.8 days	13	irregular	26.8	○/○/○/●
Symantec	35.6 days	3	occasional	1.0	●/○/○/○
Cisco Ironport	1.8 days	2	once	2.0	●/○/○/○
Trend Micro	39.2 days	1	once	1.0	●/○/○/○
McAfee	1.6 days	1	once	2.0	●/○/○/○
University of Michigan	2.6 days	58	daily	2.0	●/○/○/○
RWTH Aachen	6.2 days	10	irregular	1.0	●/○/○/○
University of Sydney	13.4 days	2	twice	1.0	●/○/○/○
Colgate University	10.1 days	1	once	1.0	●/○/○/○
KU Leuven	29.6 days	1	once	1.0	●/○/○/○

Only one crawler requested the fake pages our script sent to Alexa. To the best of our knowledge, these URLs do not appear publicly in any data released by Alexa, thus we assume this crawler was affiliated with Alexa. Except for the front page, these pages do not exist, and result in HTTP 404 errors for the crawler. We did not notice any impact on our domain ranks after the visit.

We observed a number of crawls originating from university networks, including U. Michigan, RWTH Aachen, and a crawl from KU Leuven that we were able to attribute to a concurrent study of domain lists and attacks through the detailed time and user agent description in their paper [16].

Some, but not all, crawlers request `robots.txt`, a convention for websites to tell crawlers which areas may or may not be visited, or indexed by search engines. None of the identified research crawlers respected the convention.

Our websites were highly ranked in Alexa and Umbrella, but do not have any real visitors. The fact that they were already included in research studies shows that the risk of infiltration is real, albeit we do not believe that our limited experiments skewed parallel research efforts in any significant way.

5 Discussion

We assess the likelihood and consequences of manipulation from the perspective of potential attacker motivations.

Distort empirical measurements such as web crawls. Since many security web crawls use top domain lists as their seed, if attackers manage to manipulate the ranks of existing domains, or add additional domains to the lists, they could create artificial scenario, and skew aggregate results [16]. We argue that this risk is relatively low, especially for

academic research, as the prospect of financial gain for attackers is somewhat remote. Vandalism may occur, but there is hope that it would be transient in nature, limited in scope, and could be mitigated by combining multiple data sources.

Intentional distortion of measurements is likely a minor risk, yet it could happen accidentally, as a side effect of other motivations that are more lucrative to attackers.

Bypass security mechanisms. Some research prototypes [9, 17] use features of domains from top domain lists as benign examples for training purposes, or they outright whitelist any domain found on the list. The threat intelligence feed Umbrella Investigate API [6] includes domain ranks; infiltration could make a domain appear more benign than it actually is. Thus, attackers may infiltrate top domain lists to evade detection or bypass such security mechanisms.

In contrast to the vandalism discussed above, it is easier to see how an attacker could financially benefit from a bypass attack, thus we argue that it is a medium-high risk. Fortunately, these systems usually do not depend on a specific source for their list of benign websites, and may not need any rank data at all.

Malicious infiltration of the lists could be addressed by obtaining lists of benign websites from more trustworthy sources and validating them before use, such as by using only domains in the intersection of multiple lists from different sources, and cross-checking them against blacklists, as proposed by Le Pochat et al [16].

Furthermore, as discussed in Section 3, several research studies compiled lists of presumed "benign" domains from Alexa by retaining only domains ranked for at least one year. While not perfect (e.g., this strategy cannot rule out long-lived domains compromised by attackers), it imposes an additional cost on attackers, namely a one-year preparation period for successful attacks.

Increase the value of a domain, gain more visibility and more visitors through a better (fake) ranking. Several online services provide independent estimates for potential sale prices of domain names, and some of them factor in domain rankings. Attackers could manipulate their domain's ranking to artificially inflate the domain valuation. As an extreme example, `worthofweb.com` estimates an unrealistic $ 21,000 value for one of our test domains, even though it does not receive any real visitors and was initially purchased for $ 0.50. While it is unlikely that such an estimate would be used as the sole basis for sale price negotiation, in general rank manipulation could lead to the incorrect belief of more visitors, thus a higher sale price. Similarly, a better rank may lead to higher prices that can be charged for advertising campaigns.

Rank manipulation is, in fact, not a hypothetical risk. Unscrupulous website owners can buy an "Alexa rank boost" from a range of online services, which we do not name to avoid promoting them. Some of these services promise to direct real web traffic to the website, whereas others reassure prospective customers that *"(...) We send alexa desired data to alexa system directly to improve alexa rank. So there won't be any increase in your web traffic and thus no impact on your website."* A rank of 100 k is advertised at about $ 40 per month, with the highest offered target rank of 1 k costing $ 3,300. Some of these services have been in operation for more than six years, citing customer feedback such as *"I sold my site finally at the price 3 times as previous"* and *"It helps me in talking about the ad prices."* Given the existence of these services, it is likely that rank manipulation

is already occurring in practice, but we are not aware of any proven technique to detect manipulations of top domain lists from a list consumer perspective.

6 Conclusion

We have demonstrated that attackers can place domains in Alexa's and Umbrella's domain rankings, even though these domains do not receive any real visitors. Though the lists may not have been designed to withstand attacks, they are frequently used in research in ways that they were not designed for (e.g., [7, 9, 13, 17, 18, 21]). Our research shows these attacks take up negligible resources, and are trivial to execute. A rank in the Alexa top 100 k, for instance, requires a total of 217 fake visits from seven fake toolbar users. This poses a threat to security systems that assume the most popular domains to be benign. Before using domain rankings for such a purpose, some researchers have sanitised them by discarding all domains ranked for less than one year. However, our analysis has shown that this step does not fully eliminate malicious domains. Furthermore, the limited cost of infiltration attacks means that determined attackers can circumvent such measures by mounting long-term attacks. We recommend that researchers reconsider using these rankings when rank manipulation or maliciousness could have a negative impact on their research. Detecting rank manipulation attempts, both from a list provider and list consumer perspective, is an interesting and important topic for future work.

Acknowledgements. We thank David Choffnes and Northeastern University's ITS for assisting the authors in obtaining permission to use the university's IP space. We also thank Ahmet Buyukkayhan for running Google Safe Browsing experiments on our behalf. This work was funded by the National Science Foundation under grant IIS-1553088.

References

1. Alexa top 1 million download. `http://s3.amazonaws.com/alexa-static/top-1m.csv.zip`
2. Amazon Alexa top sites. `https://www.alexa.com/topsites`
3. Are there known biases in Alexa's traffic data? `https://support.alexa.com/hc/en-us/articles/200461920-Are-there-known-biases-in-Alexa-s-traffic-data-`
4. Cisco Umbrella top 1 million. `https://s3-us-west-1.amazonaws.com/umbrella-static/index.html`
5. How are Alexa's traffic rankings determined? `https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-`
6. Umbrella Investigate API documentation. `https://investigate-api.readme.io/docs/top-million-domains`
7. Alrwais, S., Liao, X., Mi, X., Wang, P., Wang, X., Qian, F., Beyah, R., McCoy, D.: Under the shadow of sunshine: Understanding and detecting bulletproof hosting on legitimate service provider networks. In: Security & Privacy Symposium (2017)
8. Baker, L.: Manipulating Alexa traffic ratings. `https://www.searchenginejournal.com/manipulating-alexa-traffic-rankings/3044/` (2006)
9. Bilge, L., Kirda, E., Kruegel, C., Balduzzi, M.: EXPOSURE: Finding malicious domains using passive DNS analysis. In: NDSS (2011)

10. Digital Point Forums: Alexa is a scam? https://forums.digitalpoint.com/threads/alexa-is-a-scam.2016206/ (2010)
11. Englehardt, S., Narayanan, A.: Online tracking: A 1-million-site measurement and analysis. In: CCS (2016)
12. Hao, S., Thomas, M., Paxson, V., Feamster, N., Kreibich, C., Grier, C., Hollenbeck, S.: Understanding the domain registration behavior of spammers. In: IMC (2013)
13. Heiderich, M., Frosch, T., Holz, T.: IceShield: Detection and mitigation of malicious websites with a frozen DOM. In: RAID (2011)
14. Hubbard, D.: Cisco Umbrella 1 million. https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/ (2016)
15. Larisch, J., Choffnes, D., Levin, D., Maggs, B.M., Mislove, A., Wilson, C.: CRLite: A scalable system for pushing all TLS revocations to all browsers. In: Security & Privacy Symposium (2017)
16. Le Pochat, V., van Goethem, T., Tajalizadehkhoob, S., Korczynski, M., Joosen, W.: Tranco: A research-oriented top sites ranking hardened against manipulation. In: NDSS (2019)
17. Lee, S., Kim, J.: WarningBird: Detecting suspicious URLs in Twitter stream. In: NDSS (2011)
18. Lever, C., Kotzias, P., Balzarotti, D., Caballero, J., Antonakakis, M.: A lustrum of malware network communication: Evolution and insights. In: Security & Privacy Symposium (2017)
19. Lever, C., Walls, R.J., Nadji, Y., Dagon, D., McDaniel, P., Antonakakis, M.: Domain-Z: 28 registrations later. In: Security & Privacy Symposium (2016)
20. Li, Z., Zhang, K., Xie, Y., Yu, F., Wang, X.: Knowing your enemy: Understanding and detecting malicious web advertising. In: CCS (2012)
21. Nadji, Y., Antonakakis, M., Perdisci, R., Lee, W.: Connected colors: Unveiling the structure of criminal networks. In: RAID (2013)
22. Pearce, P., Ensafi, R., Li, F., Feamster, N., Paxson, V.: Augur: Internet-wide detection of connectivity disruptions. In: Security & Privacy Symposium (2017)
23. Pitsillidis, A., Kanich, C., Voelker, G.M., Levchenko, K., Savage, S.: Taster's choice: A comparative analysis of spam feeds. In: IMC (2012)
24. Rahbarinia, B., Perdisci, R., Antonakakis, M.: Segugio: Efficient behavior-based tracking of malware-control domains in large ISP networks. In: DSN (2015)
25. 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: PAM (2019)
26. 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: IMC (2018)
27. SEO Chat Forums: Alexa ranking is fake? http://forums.seochat.com/alexa-ranking-49/alexa-ranking-fake-10828.html (2004)
28. Starov, O., Nikiforakis, N.: XHOUND: Quantifying the fingerprintability of browser extensions. In: Security & Privacy Symposium (2017)