Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research

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

Northeastern University, Boston, MA
walter@iseclab.org

Abstract Top domain rankings (e.g., Alexa) are commonly used in security research, such as to survey security features or vulnerabilities of "relevant" websites. Due to their central role in selecting a sample of sites to study, an inappropriate choice or use of such domain rankings can introduce unwanted biases into research results. We quantify various characteristics of three top domain lists that have not been reported before. For example, the weekend effect in Alexa and Umbrella causes these rankings to change their geographical diversity between the workweek and the weekend. Furthermore, up to 91 % of ranked domains appear in alphabetically sorted clusters containing up to 87 k domains of presumably equivalent popularity. We discuss the practical implications of these findings, and propose novel best practices regarding the use of top domain lists in the security community.

1 Introduction

In recent years, security research has seen the emergence of Internet measurements as a subdiscipline aiming to quantify the prevalence of security risks or vulnerabilities in practice. Since many types of security assessments do not easily scale to the entire Internet, researchers typically consider only a subset of registered domains. Often, they decide to cover the most popular domains, that is, those receiving the most visitors [14, 18, 27, 32]. In doing so, they rely on "top site" rankings such as the lists compiled by Alexa [2], Majestic [6], Quantcast [7] and Umbrella [4]. Consequently, these top site lists play a central role in many studies; they decide which domain will or will not be included in the measured sample. Alexa's list in particular has become nearly ubiquitous, with multiple papers using it at any major security and Internet measurement conference [20, 30].

Many authors have commented individually on shortcomings of Alexa's ranking (e.g., lack of reliability in the bottom ranks [29], presence of malicious domains [23,24,28]) and devised their own ad-hoc mitigations to make their research results more robust against these issues (e.g., using only a list prefix [9, 22], using multiple domain lists [13, 19], and using only domains that have been present on the list for a longer time period [9, 22]). Yet, researchers are just beginning to investigate these issues in a more systematic way. In 2018, Scheitle et al. [30] and Le Pochat et al. [20] performed rigourous analyses on the nature of top lists. These works aim to understand the construction of these lists, including: how they model popularity, what their data sources are, how fast they change, and how resilient they are to manipulation attempts. While these papers have shed light on

many important characteristics of top domain lists, several aspects have gone unnoticed, or have received less attention than they deserve.

Specifically, Scheitle et al. mention a periodic *weekend effect* in Alexa's and Umbrella's lists [30], becoming manifest in a higher degree of change each weekend. We conduct a more in-depth analysis of the weekend effect by studying the content categories of the respective websites, confirming the authors' cursory finding that the weekend effect is likely due to a dominance of leisure traffic during the weekend, and office traffic during the workweek. In addition, we show that the weekend effect causes changes even among the highest ranked domains in Umbrella, whereas these domains tend to be more stable in Alexa. The weekend effect also affects country representation in the lists. These phenomena highlight the need for a more robust and stable domain selection process.

Beyond the brief reference by Le Pochat et al. [20], we are the first to quantify in detail how Alexa and Umbrella cluster domain names of equivalent popularity, while assigning them individual ranks. In fact, more than 54 % of domains in Alexa, and 91 % in Umbrella, appear in such alphabetically ordered clusters that can reach a size of up to 87 k domains. If not accounted for, the alphabetic ordering caused by clustering can cause anomalies when correlating a domain's rank with a measured property.

By characterising clustering and the weekend effect, we contribute to a better understanding of the limitations of top domain lists. We distill our findings into concrete recommendations by proposing novel best practices for the use of domain lists.

Overall, this paper makes the following contributions:

- We provide a detailed look at weekend changes in Alexa and Umbrella, the extent of these changes in different parts of the list, and the implications on the content categories and geographical diversity of listed domains.
- We are the first to quantify and explain the presence of alphabetically sorted clusters of domains in Alexa's and Umbrella's rankings.
- We discuss the implications of these phenomena for researchers using the lists in their measurements, and propose novel best practices to minimise unwanted biases.

2 Background & Related Work

In this paper, we often refer to entries of rankings or lists, but language can be confusingly ambiguous as to a "high" rank being good or bad. As a convention, when we write that a rank is *higher*, we mean that it is a *better* rank, *numerically lower*, towards the top of the list with the most popular entries.

2.1 Use of Top Lists in Security Research

Top domain lists such as the Alexa Top Sites are frequently used in security research. Le Pochat et al. [20] found 102 papers using the Alexa ranking at the four main security conferences from 2015 to 2017/2018, and Scheitle et al. [30] found 68 studies using Alexa published at the top measurement, security, and systems conferences in 2017.

Researchers can use top domain lists in different ways. In this paper, we focus on measurement studies that use these lists to select a "representative" sample of domains

Table 1: Data Sources of Common Top Site Lists.

Ranking	Data Source	List Contents
Alexa	browser toolbar	typed-in website domains
Majestic	web crawl	linked website domains
Quantcast	website instrumentation	measured website domains
Umbrella	DNS resolver	resolved (sub)domains

Alexa and Umbrella data from 2018-02-01 to 2018-05-31
Majestic data from 2018-02-28 to 2018-05-31

Table 2: Hidden Entries in Quantcast (2018-06-17).

List Prefix	Hidden #	%
$1-10$	0	0.0
$1-100$	15	15.0
$1-1,000$	136	13.6
$1-10,000$	594	5.9
$1-100,000$	1,892	1.9
$1-511,804$	5,045	1.0

to analyse, in the sense that these lists designate the "largest" or "most popular" domains (e.g., [14, 18, 27, 32]). When measurement studies compute aggregates over the domains on these lists, their results depend on how the lists select and rank domains [20, 30].

A less frequent, but common use of domain lists in security research is to obtain samples of "benign" domains. In this context, domain lists are sometimes used to train models or evaluate proposed security systems (e.g., [9, 10, 15, 22]). In a few cases, any ranked domain is whitelisted to improve classifier performance [21, 26]. This use is most sensitive to malicious domains not appearing in the ranking, and other list properties such as stability or ordering are less critical. Maliciousness of ranked domains has been studied before [23, 24, 28], and this scenario is beyond the scope of this paper.

2.2 List Compilation Methodology

We are aware of four major measurement-based top site lists: Amazon Alexa Top Sites [2], The Majestic Million [6], Quantcast Top Websites [7], and Cisco Umbrella Top 1 Million [4]. Table 1 summarises the data source and popularity model of each ranking.

Alexa. The data for the ranking originates primarily from "millions of users" [3] who have installed the Alexa toolbar and share their browsing history with Alexa. Its website documents Alexa's methodology as follows: The toolbar only collects URLs that appear in the address bar of the browser window or tab. Sudomains are not ranked separately from the main domain, unless they can be determined to be blogs or personal homepages. 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. Ranks below 100 k are not statistically meaningful because the data collected about those domains is too scarce [3,5]. The ranking is updated daily. Our work uses the ranking from the file download [1]. In contrast to the API and website, ranks in the file do not appear to be smoothed.

Majestic. Majestic's ranking is based on the link graph built from a continuously updated, proprietary web crawl comprising over 528 B URLs as of June 2018 [6]. Domains are ranked by the number of unique /24 IP networks hosting inbound links [17].

Quantcast. Ranks are based on direct traffic measurements through client-side tracking code embedded by Quantcast's customers into their websites and mobile applications, as well as estimated traffic (from unspecified sources) for non-customer websites [7]. Quantcast customers can choose to hide their identity in the ranking. Table 2 shows that around 1 % of all list entries are hidden, but for some list prefixes the percentage can be much higher, such as 15 % in the top 100. These censored entries make it challenging to compare this ranking to others. Therefore, we do not consider it further in this paper.

Umbrella. The ranking is computed 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. According to Umbrella, ranks are based on the unique client IPs looking up a domain [16].

2.3 Related Work

In 2006, Lo and Sharma Sedhain compared the lists available at that time to determine how similar and reliable they were [25]. Out of the lists we initially considered relevant for this study, they included only Alexa. Given the long time that has passed since then, it is likely that the ranking methodology and list composition have changed.

Scheitle et al. [30, 31] study the domains on the lists compiled by Alexa, Umbrella and Majestic, how these lists differ, how they evolve over time, how they are being used in research studies, how list parameters influence the outcome of research studies, and how the rankings could be manipulated. The authors describe a weekend effect in Alexa and Umbrella, a periodic change in list composition between weekday and weekend rankings. While the authors convey an intuition as to why this effect exists, we provide a more detailed analysis of the reasons and implications of this phenomenon. We describe an additional phenomenon, clustering of equivalent domains in Alexa and Umbrella, which is not mentioned by Scheitle et al., and discuss potential implications.

Le Pochat et al. [20] also quantify several properties of domain lists and reproduce prior studies using different lists, but the focus of their work is on attacks to influence the rankings. While the weekend effect is visible in one of the figures, it is not further mentioned or analysed. The authors do mention clustering, but only in an attack context, and without discussing the implications for research studies relying on these lists.

When discussing their results, both papers make high-level recommendations how other researchers should use domain lists in their studies. We believe that this topic warrants more discussion and conclude our paper with several additional recommendations.

3 List Analysis

In the following, we study weekend effects and clustering in the top 1 M rankings of Alexa, Majestic, and Umbrella. We downloaded the respective ranking file every day. We label the data with the date one day prior to downloading, as a list updated and downloaded on Monday, for instance, appears to contain the ranks computed from Sunday data.

3.1 List Stability

We begin our analysis with a look at *how much* and *how fast* the rankings change. In contrast to prior work [20, 30], we divide each ranking into non-overlapping intervals of exponentially increasing length 1–10, 11–100, 101–1,000, etc. This provides a better view on *which parts* of the ranking change. The exact order of domains within each interval does not matter for many uses in security research, thus we allow for reordering or minor rank changes by calculating set intersections. We pick a single reference day, 2018-02-07 for Alexa and Umbrella, 2018-03-28 for Majestic, and compare all subsequent days upto 2018-05-31 to this day. This allows us to visually distinguish long-term drift from transient changes. Figure 1 uses a Wednesday as a representative of the workweek; similar heatmaps using a Sunday for the weekend can be seen in Figure 6 in the appendix.

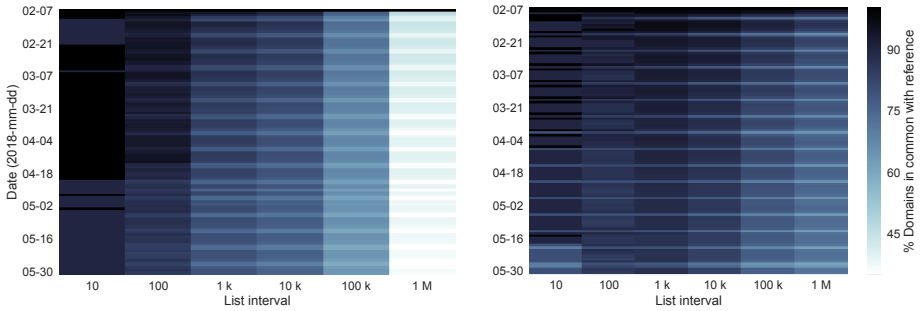
At a high level, the heatmaps show that the top ranked domains exhibit less change than the lower intervals of the ranking. This is in line with Scheitle et al. [30], who showed that longer list prefixes tend to exhibit lower stability. In contrast to prior work, our representation reveals that the higher ranks in Alexa are more stable than in Umbrella, where changes occur within the top 10 domains on a regular basis. The bottom 900 k domains, however, are considerably less stable in Alexa than they are in Umbrella. In the bottom of the plot, most intervals get lighter in color, corresponding to long-term drift.

Scheitle et al. [30] describe a *weekend effect* in Alexa and Umbrella, a weekly pattern where change is highest on the weekend. This pattern appears in the heatmaps as regular horizontal bands. While only implied by Scheitle et al., the heatmaps in Figure 1 confirm that the change is indeed transient, that is, the ranking tends to revert back to the original domains after the weekend. Furthermore, close inspection of the heatmaps shows that the weekend differences are strongest on Sundays. Figure 6 in the appendix contains similar heatmaps using a Sunday as the reference day, and shows the expected inverted pattern of a greater difference during the workweek, and less during the weekend, relative to the Sunday list. Umbrella has the strongest weekend effect, with changes occurring even in the top 10. For example, Table 4 in the appendix shows that Netflix moves from ranks two and three to one and two, and Hola appears with two new entries. Majestic, shown in Figure 7 in the appendix, has no discernible weekend effect, as its ranks appear stable.

3.2 The Weekend Effect

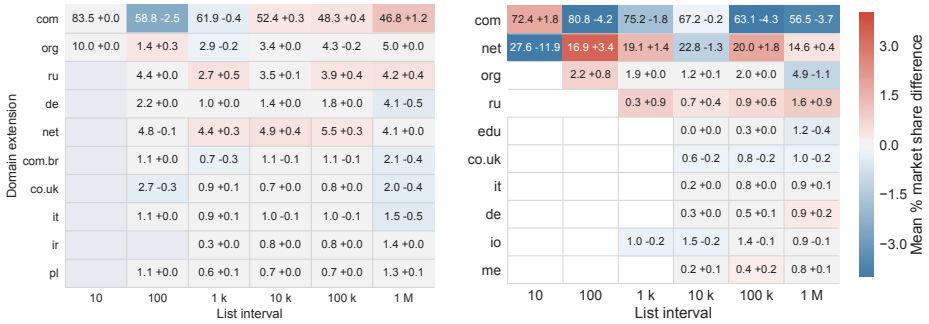
Alexa and Umbrella exhibit strong, temporary changes each weekend. Using domain extensions and website categories, we quantify how this affects the type of listed domains.

Domain Extensions. To judge how the lists represent different geographical regions, we look at country-code domain extensions, or more precisely, public suffixes. The public suffix of a domain is the domain extension under which domains can be registered, such as `.cl` or `.co.uk`. Country-code domain extensions are only a coarse-grained approximation of country-level popularity, as many regions use generic top-level domains such as `.com` in addition to their country-code domain, and the U.S. in particular makes comparatively little use of their `.us` extension. However, the way how each region splits its traffic across generic and country-code domains should be stable, which means that we can use domain extensions to uncover weekday to weekend changes.



(a) Alexa (Wednesday) (b) Umbrella (Wednesday)

Figure 1: Heatmaps showing the set intersection of ranked domains with the reference day, Wed. 7 February, in exponentially increasing list intervals 1–10, 11–100, 101–1,000, etc. Horizontal lines correspond to the weekend effect, which is stronger in Umbrella, whereas Alexa has stronger long-term drift. For Majestic, see Figure 7 in the appendix.



(a) Alexa (b) Umbrella

Figure 2: Heatmaps showing domain extensions' mean Wednesday market share ± the difference to the mean Sunday market share (also used to colour each cell) in exponentially increasing list intervals 1–10, 11–100, 101–1,000, etc., from February to May 2018. Extensions ordered by Wednesday top 1 M mean market share. Weekends cause a change in geographic representation. For Majestic, see Figure 9 in the appendix.

Figure 2 shows the most common public suffixes used in Alexa and Umbrella on Wednesdays from February to May 2018, ordered by their mean market share. Different list intervals often exhibit variation in the relative popularity of domain extensions. For example, `.jp` is the sixth most frequent extension in Alexa's top 100 k, whereas it is ranked twenty-fourth in the full list. Extension diversity differs between the lists, with Alexa containing 33 extensions in the top 100, Majestic 13, and Umbrella only 4.

The weekend effect affects the geographical diversity of Alexa and Umbrella. On weekends, Alexa loses domains from European countries and gains in Russia, India, and for `.com` (from mean of 47.0 to 48.1 %); Umbrella also includes more Russian domains,

Table 3: Top 5 Unresolvable Public Suffixes in Umbrella, Feb. to May 2018.

Suffix	Wednesday (mean freq. / best rank)	Sunday (mean freq. / best rank)
`localhost`	18 / 18,583	7,852 / 11,829
`local`	835 / 2,211	1,080 / 1,530
`home`	705 / 2,629	1,266 / 1,331
`lan`	566 / 6,246	948 / 3,687
`localdomain`	208 / 13,852	315 / 8,723

and more domains with invalid extensions, but has fewer `.com` domains (from 57.1 to 53.4 % in the full list). Only Majestic remains relatively stable, most likely due to its ranking reflecting structural properties of a website link graph and not visitor popularity.

Invalid Domains. All of the domains in Alexa use a well-known public suffix, but a mean of 0.5 % (Wednesday) and 1.6 % (Sunday) of Umbrella domains and 0.004 % of Majestic domains have a non-delegated domain extension. Such domains cannot currently be registered or resolved on the public Internet. In fact, Umbrella appears to contain domains used internally in corporate networks. These domains can appear quite high in the ranking, such as the domain `tcs` at rank 820. Table 3 shows the five most frequently used invalid domain extensions in Umbrella. Each Wednesday, Umbrella contains a mean of 18 domains with the `localhost` extension, the highest of which was observed at rank 18,583, while each Sunday, `localhost` contains a mean of 7,852 domain with a best rank of 11,829. This trend is consistent with other invalid domains, showing that invalid domains peak on the weekend. The list also contains a mean of 198 `corp` domains, and entries corresponding to the names of networking equipment manufacturers such as `belkin` and `dlink`. Chen et al. [11, 12] describe how internal domain name lookups can leak into the public Internet, where they are susceptible to attacks.

Website Categories. Similar to country-level representations, the lists may exhibit differences in the content-level types of domains they contain. We utilise Symantec/BlueCoat WebPulse [8] to categorise the top 10 k domains of each list, assuming that they are websites. For subdomains, the category usually refers to the registered parent domain.

We successfully retrieve categories for 97.8–98.3 % of domains in the top 10 k from March and April 2018. Domains listed in Alexa and Majestic are classified into 63 and 62 categories, respectively, whereas Umbrella covers only 53 distinct categories. This effect is even more pronounced in the top 1 k, where Alexa contains 48 categories, Majestic 39, and Umbrella only 23. Umbrella contains many subdomains [20, 30], which results in a significantly less diverse set of websites. Figure 3 shows the most frequent categories ordered by their Wednesday market share. The category market share distribution in Alexa is much more balanced than in Umbrella, resulting in a better representation of websites of different categories.

The types of categories also differ between the lists. The Wednesday Alexa in the interval 100–1 k contains 7.5 % websites that could be considered “unsafe for work” environments, whereas in Umbrella, the percentage is only 0.2 %. This suggests that the

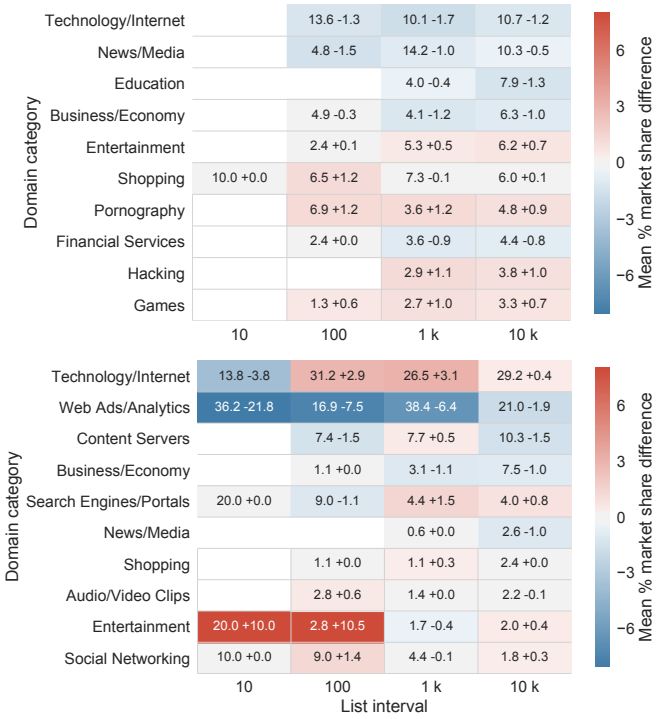


Figure 3: Alexa (top) and Umbrella (bottom) heatmaps showing website categories' mean Wednesday market share ± the difference to the mean Sunday market share (also used to colour each cell) in exponentially increasing list intervals 1–10, 11–100, 101–1,000, etc., from March to April 2018. Categories ordered by Wednesday top 1 M mean market share. Sundays see fewer office-related domains, and more entertainment. For Majestic, see Figure 10 in the appendix.

Umbrella ranking may be based on a larger share of corporate traffic. Similarly, while the News/Media category is ranked first in Sunday Alexa, it appears at rank 12 in Umbrella. In contrast, Umbrella highly ranks several categories that appear to apply to internal subdomains and subresources such as Web Ads/Analytics, the highest ranked category at (38.4 % Wed.), as well as Content Servers (7.7 % Wed.) and Non-Viewable/Infrastructure (4.0 % Wed.). For comparison, in the Alexa top 1 k, the former categories account for only 2.8 % and 0.5 %, respectively, and the latter category does not appear. This further illustrates the effects of Umbrella's subdomain inclusion.

From the weekdays to the weekend, Alexa and Umbrella both lose in business-related categories and gain in various forms of entertainment. In the Umbrella interval 100–1 k, the Business/Economy category loses 1.1 percentage points, whereas the Chat category gains 0.9 percentage points; Games increase their market share threefold. Furthermore, categories appear to be slightly more evenly distributed during the weekend. The categories of the top 10 domains (Table 4) in Alexa, and to some extent also the top 100, remain

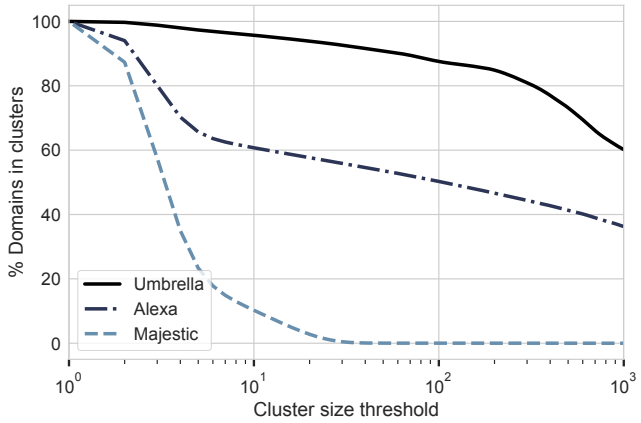


Figure 4: Percentage of the ranking that is part of a cluster, for varying minimum length thresholds for an alphabetically sorted sequence to be considered a cluster. Alexa and Umbrella cluster a large fraction of their respective list, Majestic does not.

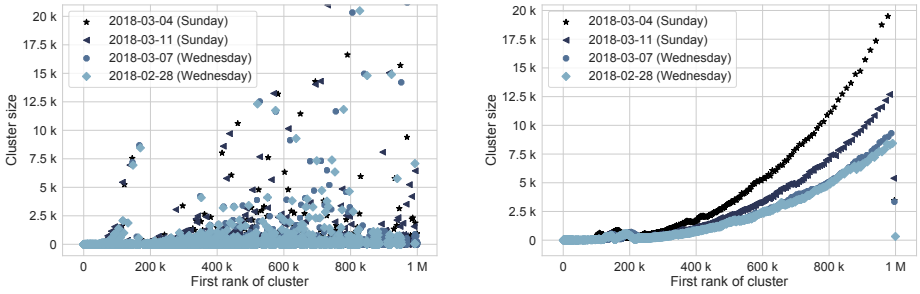
stable between the workweek and the weekend. In Umbrella, however, there is significant change in the categories in the top 10 because of the addition of two new domains. Taken together, these results confirm the preliminary finding by Scheitle et al. [30] (based on popularity changes of a handful of domains listed with many subdomains) that Alexa and Umbrella are dominated by office traffic during the workweek, and leisure traffic during the weekend.

3.3 Clusters

The rankings of Alexa and Umbrella contain large alphabetically sorted clusters of domain names. (Umbrella appears to apply an atypical sorting order when dashes and prefixes are involved: `ab-c` before `ab`.) We assume that these clusters represent domains that cannot be distinguished based on their traffic characteristics.

Alphabetically sorted sublists may occur coincidentally. To explore minimum size thresholds for when a sorted sublist may be considered a cluster, we plot in Figure 4 the resulting percentage of domains that would be considered part of any cluster. Majestic has only very small clusters; fewer than 0.05 % of the list would be part of clusters if they were required to be larger than 42 domains. Applying the same threshold to the other lists, more than 54 % of Alexa, and more than 91 % of Umbrella appear in a cluster.

To understand the sizes and rank locations of clusters, Figure 5 plots the length of each alphabetically ordered sublist against its first rank. In Alexa, larger clusters start appearing at ranks around 49 k. Clusters can grow very large, with outliers of 40 k and 87 k domains, and their size does not increase monotonically. Majestic, shown in Figure 8 in the appendix, has no significant clusters except for a few outliers in the last third of the list. In Umbrella, clusters larger than 42 domains start at rank 83 k (rank 126 k with a threshold of 100). The size of clusters appears to grow exponentially towards the end of



(a) Alexa (b) Umbrella

Figure 5: Scatterplots of each alphabetically sorted cluster's size by its highest rank. No size threshold, but clusters with 42 or fewer domains are downsampled to 1 % for printability (difference invisible). In Alexa, ten very large outlier clusters of up to 87 k domains not shown. Umbrella clusters have a trend of increasing size towards the end of the list, except for the last (truncated) cluster; weekends tend to increase cluster sizes.

the list, but the last cluster of the list is likely truncated as it does not follow the increasing trend. Furthermore, clusters on the two Wednesdays are one third to a half smaller than the clusters observed on Sundays. This suggests that Umbrella's ranking is based on less traffic on Sundays, as larger clusters imply more domains that cannot be distinguished.

These clusters have a number of important implications for users of the lists. First, while one may expect that domains equivalent in terms of traffic would receive the same rank, Alexa and Umbrella do in fact assign individual ranks to each domain in alphabetical order. Inside a large cluster, the first few characters of a domain can cause a large rank difference, such as 87 k in Alexa. The last cluster of the ranking is cut off, as including it entirely would extend the length of the list beyond 1 M entries. Similar effects can occur when researchers use a list prefix without accounting for clusters. In both cases, domains are excluded from consideration not because of their popularity, but because of their relative lexicographical order. Furthermore, clustering effects have implications on the stability of the list. A domain with stable traffic may receive a worse rank when domains with equivalent traffic but a lower lexicographical ordering are added to the list. Similarly, when a domain switches to an adjacent cluster, the rank difference can be consequential, even though the actual change in traffic may be minor.

4 Discussion: Best Practices for Using Top Domain Lists

Our analysis has revealed various characteristics of the lists compiled by Alexa, Majestic, and Umbrella. To minimise any negative impact that these characteristics can have on measurement results, we propose the following best practices.

Avoid direct correlation with a domain's rank. Alexa and Umbrella contain large clusters of domains with the same popularity, yet each domain is assigned an individual rank in alphabetical order. For example, 56 % of Alexa, and 99.9 % of Umbrella entries in

the bottom 900 k are part of clusters. This can cause anomalies, e.g. when looking for a linear correlation between the rank and a security property ("do more popular websites have a higher security score?"). Furthermore, especially the lower domain ranks can fluctuate considerably on a daily basis. Instead of using the rank directly, we suggest looking at aggregates based on exponentially increasing rank intervals, such as 1–10, 11–100, 101–1000 etc., which perhaps results in less precision, but more robustness.

Use contemporaneous rankings to label historical datasets. A domain that was popular in the past is not necessarily highly ranked today, and vice versa. When labelling a dataset with domain ranks, it is important to use the rank that was current at the time of the recorded event. For example, a Web vulnerability database may contain entries spanning multiple years, and a website's popularity should be assessed based on the ranking when the vulnerability was discovered ("do popular websites receive more vulnerability reports?"). The fast responsiveness of Alexa and Umbrella implies that this precaution is also necessary at shorter time scales. A malicious domain, for instance, may be active and popular for just a few days before it is blacklisted [28] and traffic subsides.

Measure a static set of domains, if possible. We have shown that the weekend effect in Alexa and Umbrella causes different types of domains to be included (e.g., changing country and content category distributions). This pattern is also visible in multiple network, transport and application layer measurements reproduced daily with the newest domain lists by Scheitle et al. [30]. However, such a measurement setup does not allow to distinguish changes due to list composition from changes that occur on a measured domain. For example, a domain that is present in the ranking during all days may cease to use a certain form of tracking, or a domain that always uses this form of tracking may drop out of the ranking, to the same overall effect. We argue that *short-term* noise from list composition, such as the weekend effect, is usually undesirable in measurements. It makes it challenging to interpret observed changes, and it is typically of little interest to break down the prevalence of tracking, for instance, based on sites' weekend or workweek popularity. We believe that measurements can often be carried out with a static list of domains, such as to study the evolution of tracking on a fixed set of sites. *Medium* and *long-term* list changes may be more relevant to account for permanent popularity changes.

To create a set of domains to be measured, we suggest collecting list data over the course of one or more weeks, and using the union or intersection of all days, depending on the scenario. To improve comparability and reproducibility of measurements, researchers could agree on a common list of domains that is updated on a quarterly or yearly basis.

Account for subdomains. Umbrella contains so many subdomains that the set of unique registered domains is only around 28 % [30], three times smaller than in Alexa or Majestic. In some contexts, measuring all subdomains may be desirable. For example, subdomains may serve different web content, and subdomains include mail servers that may use a different TLS configuration than web servers. In other contexts, subdomains may be aliases, or may be configured and managed identically since they are part of the same infrastructure. For example, if they share the same authoritative name server, they likely have identical DNSSEC capabilities. This can result in duplicates, and bias aggregates towards services that are listed with more subdomains. In such cases, it may be preferable

to use only one (sub)domain per unique registered domain. Similar issues exist due to Content Distribution Networks [30]. The large difference in unique registered domains also makes it challenging to compare results derived from Umbrella to Alexa or Majestic.

For completeness, we discuss recommendations from prior work in the paragraphs below.

Use lists according to what they represent. Each list uses different data, which implies a different definition of popularity. Alexa contains (desktop) type-in website domains, Majestic models website popularity by inbound links instead of visitors, Umbrella ranks domains that may not host web content, and Quantcast allows customers to hide their identity. These missing ranks are not uniformly random (Table 2), and we recommend against the use of Quantcast when representativity or comparability are desired.

Use only the highest-ranked 100 k domains, or fewer. The publishers of the Alexa list caution that only the top 100 k domains are statistically significant [3, 5]. The reverse conclusion is a degree of imprecision, or randomness, in the remaining 900 k list entries; this could refer both to their relative ranking, and to their presence or absence. Research results aggregated over the full 1 M list are based on 90 % "unreliable" data points.

Use multiple sources, including unranked sets of domains. The limitations of individual lists could be mitigated by measuring domains selected from multiple lists in parallel and contrasting the results, as suggested [30] and done [19] in prior work. Researchers could base their analysis on one ranked domain list and a random sample of the `.com` zone (or IPv4 address space). The first set of domains would be "representative" in terms of visits and mirror the security aspects that users face, while the second set would be "representative" in terms of sites and reflect security from the point of view of developers.

Do not assume that ranked domains are benign. This paper and the previous recommendations focus on measurement studies, where domains do not need to be benign. In fact, several prior studies have reported evidence that malicious domains exist in the Alexa ranking [23,24,28]. This can be an issue for security systems using domain lists as sources of "benign" examples for model training, validation or whitelisting [9, 10, 15, 21, 22, 26].

5 Conclusion

Many security research papers utilise top domain rankings such as the ones published by Alexa, Majestic, or Umbrella to select domains or websites to consider in their study. Each list models popularity in a different way. Alexa contains only type-in website domains based on their popularity with toolbar users, Majestic ranks websites based on structural properties rather than popularity with actual visitors, and Umbrella includes any type of domain observed at a large public DNS resolver, including many internal, non-web domains. Consequently, the lists differ in the country and category distribution of their domains, and some exhibit immediate reactivity to momentary changes in traffic volume and distribution, making weekday and weekend rankings look quite distinct. If not properly accounted for, these characteristics can hamper reproducibility, and introduce unwanted bias into research results derived from the domains in the rankings. To that end, we have proposed best practices for the use of top domain lists in security measurements.

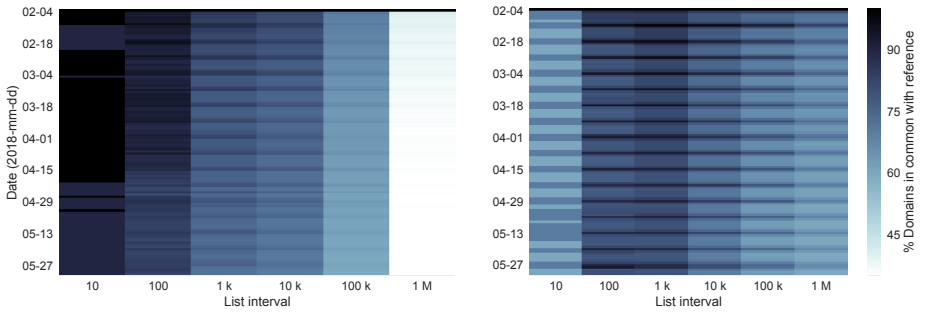
Acknowledgements. This work was supported by Secure Business Austria and the National Science Foundation under grants CNS-1563320, CNS-1703454, and IIS-1553088.

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. Majestic million. https://majestic.com/reports/majestic-million
7. Quantcast top websites. https://www.quantcast.com/top-sites/
8. Symantec BlueCoat WebPulse site review. https://sitereview.bluecoat.com/
9. 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)
10. Bilge, L., Kirda, E., Kruegel, C., Balduzzi, M.: EXPOSURE: Finding malicious domains using passive DNS analysis. In: NDSS (2011)
11. Chen, Q.A., Osterweil, E., Thomas, M., Mao, Z.M.: MitM attack by name collision: Cause analysis and vulnerability assessment in the new gTLD era. In: Security & Privacy Symposium (2016)
12. Chen, Q.A., Thomas, M., Osterweil, E., Cao, Y., You, J., Mao, Z.M.: Client-side name collision vulnerability in the new gTLD era: A systematic study. In: CCS (2017)
13. Durumeric, Z., Kasten, J., Bailey, M., Halderman, J.A.: Analysis of the HTTPS certificate ecosystem. In: IMC (2013)
14. Englehardt, S., Narayanan, A.: Online tracking: A 1-million-site measurement and analysis. In: CCS (2016)
15. Heiderich, M., Frosch, T., Holz, T.: IceShield: Detection and mitigation of malicious websites with a frozen DOM. In: RAID (2011)
16. Hubbard, D.: Cisco Umbrella 1 million. https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/ (2016)
17. Jones, D.: Majestic million CSV now free for all, daily. https://blog.majestic.com/development/majestic-million-csv-daily/ (2012)
18. 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)
19. Lauinger, T., Chaabane, A., Arshad, S., Robertson, W., Wilson, C., Kirda, E.: Thou shalt not depend on me: Analysing the use of outdated JavaScript libraries on the Web. In: NDSS (2017)
20. Le Pochat, V., van Goethem, T., Tajalizadehkhoob, S., Korczynski, M., Joosen, W.: Rigging research results by manipulating top websites rankings. In: NDSS (2019)
21. Lee, S., Kim, J.: WarningBird: Detecting suspicious URLs in Twitter stream. In: NDSS (2011)
22. Lever, C., Kotzias, P., Balzarotti, D., Caballero, J., Antonakakis, M.: A lustrum of malware network communication: Evolution and insights. In: Security & Privacy Symposium (2017)
23. Lever, C., Walls, R.J., Nadji, Y., Dagon, D., McDaniel, P., Antonakakis, M.: Domain-Z: 28 registrations later. In: Security & Privacy Symposium (2016)

24. Li, Z., Zhang, K., Xie, Y., Yu, F., Wang, X.: Knowing your enemy: Understanding and detecting malicious web advertising. In: CCS (2012)
25. Lo, B.W.N., Sharma Sedhain, R.: How reliable are website rankings? Implications for e-business advertising and Internet search. Issues in Information Systems **VII**(2), 233–238 (2006)
26. Nadji, Y., Antonakakis, M., Perdisci, R., Lee, W.: Connected colors: Unveiling the structure of criminal networks. In: RAID (2013)
27. Pearce, P., Ensafi, R., Li, F., Feamster, N., Paxson, V.: Augur: Internet-wide detection of connectivity disruptions. In: Security & Privacy Symposium (2017)
28. Pitsillidis, A., Kanich, C., Voelker, G.M., Levchenko, K., Savage, S.: Taster's choice: A comparative analysis of spam feeds. In: IMC (2012)
29. Porter Felt, A., Barnes, R., King, A., Palmer, C., Bentzel, C., Tabriz, P.: Measuring HTTPS adoption on the Web. In: Usenix Security (2017)
30. 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)
31. Scheitle, Q., Jelten, J., Hohlfeld, O., Ciprian, L., Carle, G.: Structure and stability of Internet top lists. In: eprint arXiv:1802.02651 [cs.NI] (2018)
32. Starov, O., Nikiforakis, N.: XHOUND: Quantifying the fingerprintability of browser extensions. In: Security & Privacy Symposium (2017)

Appendix



(a) Alexa (Sunday)

(b) Umbrella (Sunday)

Figure 6: Alexa and Umbrella changes over time in exponentially increasing list intervals, using Sunday 4 February as the reference day. See Figure 1 for full legend.

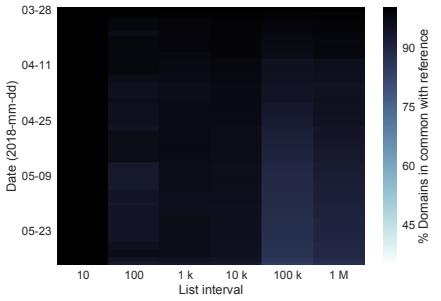


Figure 7: Changes in Majestic over time in exponentially increasing list intervals, using Wednesday 24 March as the reference day. See Figure 1 for full legend. Majestic is remarkably stable.

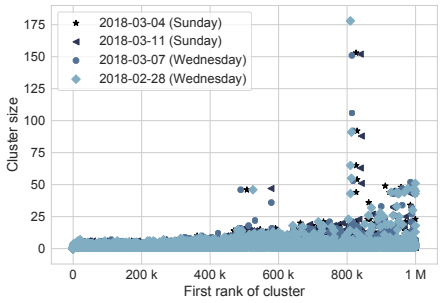


Figure 8: Scatterplot of each alphabetically sorted cluster's size by its highest rank in Majestic. Partially visible downsampling of small clusters. See Figure 5 for full legend. Majestic has only small clusters.

Table 4: Top 10 Domains on Wed. 4 and Sun. 8 April 2018 in Alexa and Umbrella.

	Alexa	**Umbrella**	
	Wednesday & Sunday	Wednesday	Sunday
1	`google.com`	`google.com`	`netflix.com`
2	`youtube.com`	`netflix.com`	`api-global.netflix.com`
3	`facebook.com`	`api-global.netflix.com`	`google.com`
4	`baidu.com`	`www.google.com`	`microsoft.com`
5	`wikipedia.org`	`microsoft.com`	`ichnaea.netflix.com`
6	`yahoo.com`	`facebook.com`	`www.google.com`
7	`reddit.com`	`doubleclick.net`	`facebook.com`
8	`google.co.in`	`g.doubleclick.net`	`hola.org`
9	`qq.com`	`googleads.g.doubleclick.net`	`dns-test1.hola.org`
10	`taobao.com`	`google-analytics.com`	`doubleclick.net`

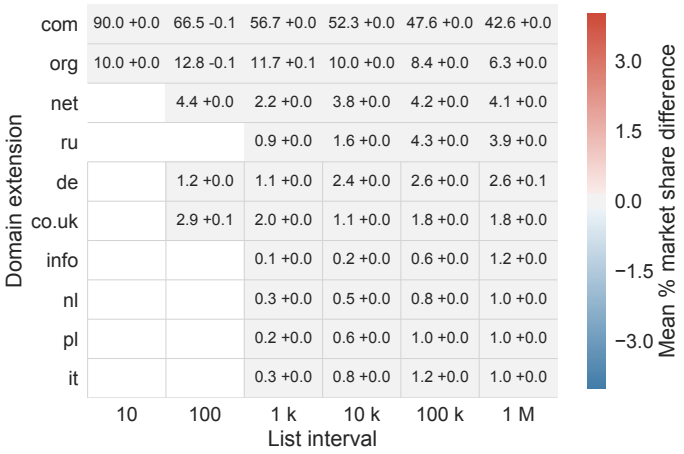


Figure 9: Heatmap showing Majestic domain extensions' mean Wednesday market share ± the difference to the mean Sunday market share (also used to colour each cell) in exponentially increasing list intervals 1–10, 11–100, 101–1,000, etc., from March to May 2018. Extensions ordered by Wednesday top 1 M mean market share. Due to Majestic's high list stability, differences are not visible.

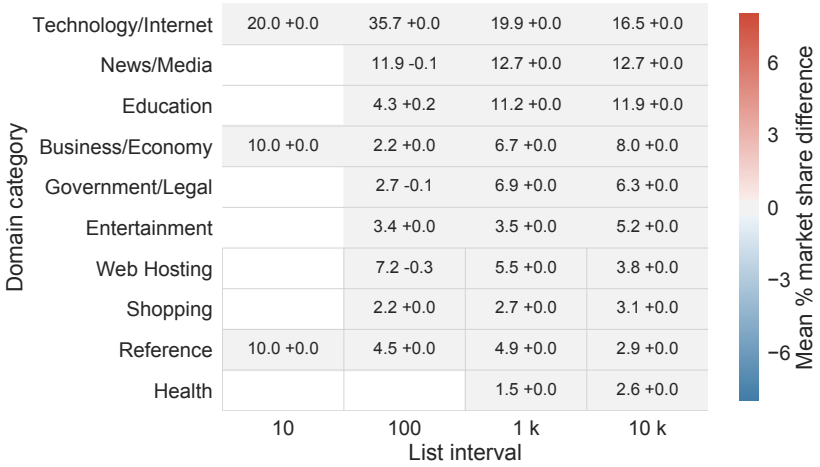


Figure 10: Heatmap showing Majestic website categories' mean Wednesday market share ± the difference to the mean Sunday market share (also used to colour each cell) in exponentially increasing list intervals 1–10, 11–100, 101–1,000, etc., from March to April 2018. Categories ordered by Wednesday top 1 M mean market share. Due to Majestic's high list stability, differences are not visible.