Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists

Kimberly Ruth
Stanford University
kcruth@cs.stanford.edu

Deepak Kumar
Stanford University
kumarde@cs.stanford.edu

Brandon Wang
Independent Researcher
bmw4@illinois.edu

Luke Valenta
Cloudflare, Inc.
lvalenta@cloudflare.com

Zakir Durumeric
Stanford University
zakir@cs.stanford.edu

ABSTRACT

Researchers rely on lists of popular websites like the Alexa Top Million both to measure the web and to evaluate proposed protocols and systems. Prior work has questioned the correctness and consistency of these lists, but without ground truth data to compare against, there has been no direct evaluation of list accuracy. In this paper, we evaluate the relative accuracy of the most popular top lists of websites. We derive a set of popularity metrics from server-side requests seen at Cloudflare, which authoritatively serves a significant portion of the most popular websites. We evaluate top lists against these metrics and show that most lists capture web popularity poorly, with the exception of the Chrome User Experience Report (CrUX) dataset, which is the most accurate top list compared to Cloudflare across all metrics. We explore the biases that lower the accuracy of other lists, and we conclude with recommendations for researchers studying the web in the future.

ACM Reference Format:
Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. In *ACM Internet Measurement Conference (IMC '22), October 25–27, 2022, Nice, France.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3517745.3561444

1 INTRODUCTION

Nearly a quarter of recent Internet measurement papers rely on lists of popular websites like the *Alexa Top Million* [2, 27]. Unfortunately, recent results have called into question the soundness of these "top lists," and, by proxy, the research results derived from them [16, 18, 19, 25–27]. Further exacerbating the concern, Amazon recently announced the planned shutdown of the *Alexa Top Million* [5], the most commonly used list of websites and the cornerstone of amalgam lists like the *Tranco Top Million* [18].

Alexa's retirement presents an opportunity for the research community to establish more trustworthy methodologies for studying the web. However, evaluating top lists is challenging, requiring access to privileged data, either in the form of sensitive browsing data from a large, globally distributed set of web clients or access to server logs from a significant number of independently operated websites. As a result, no work has yet directly evaluated the accuracy of the Alexa Top Million or proposed alternatives, leaving researchers without a clear alternative or a principled path forward.

In this paper, we partner with Cloudflare, a popular CDN and DDoS mitigation provider, to evaluate the relative accuracy of top lists of websites. While Cloudflare authoritatively serves traffic for only about a quarter of top sites, this is significantly more than any other provider, and we find that despite its limitations, the perspective uncovers meaningful differences between top lists.

Each top list employs its own methodology for inferring popularity, and we start by deriving a set of seven server-side metrics that present a diverse set of perspectives on what it means for a website to be popular. These include metrics that approximate both the number of page loads served by each website and the number of unique clients that access each website. Surprisingly, we find that all seven of our metrics evaluate the relative accuracy of the set of sites captured by top lists *identically* (i.e., all metrics agree on which top list captures best the set of most popular sites).

Overall, we find that top lists (including Alexa) capture the set of top websites relatively poorly across all of our metrics with one exception: Google's *Chrome User Experience Report* (CrUX), which has recently begun publishing rank order magnitude buckets (i.e., top 1K, 10K, 100K, 1M) of the most popular websites as seen by Google Chrome. In contrast to other top lists, CrUX is as similar to Cloudflare metrics as the varied Cloudflare metrics are to one another. We investigate recent papers that utilize top lists and find that the vast majority use top lists as only an *unordered set* of websites to study. As such, CrUX is a compelling set of websites to consider in future research studies. Cisco Umbrella—a rank order list of the most commonly queried *names* rather than *websites*—captures the set of popular sites second best, but is not able to capture the relative accuracy of individual websites.

Through further analysis of Cloudflare data along with additional supplementary data from Google Chrome, we analyze the biases in the websites included/excluded by top lists and investigate why lists show inconsistent accuracy in ranking websites. We conclude with recommendations for the research community on how to more accurately study the web moving forward.

IMC '22, October 25–27, 2022, Nice, France

ACM ISBN 978-1-4503-9259-4/22/10…$15.00
https://doi.org/10.1145/3517745.3561444

2 BACKGROUND AND RELATED WORK

Both industry and academic research rely heavily on published lists of popular websites like the *Alexa Top 1 Million Websites* [1] to model web browsing behavior [2, 27]. Indeed, a 2018 study by Scheitle et al. found that more than 10% of papers at top Internet measurement (22% of papers), security (9%), networking (6%), and web (8%) venues used a list of popular websites as the foundation of their analysis [27]. At the extreme, more than a quarter of papers at IMC'17 relied on a list of popular websites like the Alexa Top Million.

There are several competing top lists, each of which employs a unique methodology for inferring popularity. In some cases, lists specifically rank the most popular *websites* whereas others simply compute the most popular *domain names* independent of the application layer protocol. In nearly all cases, lists attempt to provide an approximate rank order list of names. Below, we detail top lists appearing prominently in prior academic literature:

- The Alexa Top 1 Million [1] approximates site popularity by tracking the browsing behavior of several million users through partnerships with a reported 25K browser extensions as well as through websites that install *Alexa Certify* code [4]. While their methodology is private, Alexa states that their rank is calculated daily based on "the average daily visitors and pageviews to every site over the past 3 months" [3, 6]. The Alexa Top Million is the most commonly used list of popular websites [27], but Alexa has announced that it will be discontinued in December 2022 [5].
- The Cisco Umbrella 1 Million [10] is a list of the most popular *names* (e.g., `.com` is ranked #1) looked up using Cisco Umbrella's DNS service. Their ranking is based on a propriety algorithm, which "uses the number of unique client IPs visiting each domain, relative to the sum of all requests to all domains" to calculate popularity [33].
- The Majestic Million [20] is a list of popular websites maintained by Majestic SEO, which is calculated based on the number of backlinks that each site has [21].
- The Secrank list [34] is a researcher-built list that aggregates several features of DNS data from a major resolver in China. Each IP address "votes" for domains based on request volume and frequency of access, and IP addresses are weighted according to their requests' domain diversity and total volume. The list is designed to be stable, transparent, and resistant to manipulation.
- The Tranco Top Million List [17, 18] aggregates data from the Alexa, Umbrella, and Majestic lists over a 30 day window to form a ranking that is more temporally stable and resistant to adversarial manipulation.
- The Trexa Top Million [35] interleaves Tranco and Alexa rankings (i.e., additionally weighting towards Alexa) to better approximate the observed browsing (i.e., intentional URL loads) of 52K Firefox users who opted into a Mozilla research study [35].

There exist other commercial lists, like SimilarWeb [29] and Comscore [12], which are paywalled and rarely used in academic research. Others, like Cloudflare Radar [11], lack public APIs, which have precluded research use.

There is little agreement between top lists in terms of both overlap and rank order of names [27]. Further, the choice of list can lead to dramatically different downstream research conclusions [19, 27]. A series of past studies have shown that lists are unstable, inconsistent, and vulnerable to external manipulation [16, 19, 25–27]. Most recently, Scheitle et al. formalized three key properties for top lists: stability, consistency, and transparency [27]; they show that top lists are unstable, have little intersection, and rarely disclose methodological details. However, without clear sources of ground truth, these studies have stopped short of understanding whether instability is indicative of low accuracy and which, if any of the lists, best represent user behavior.

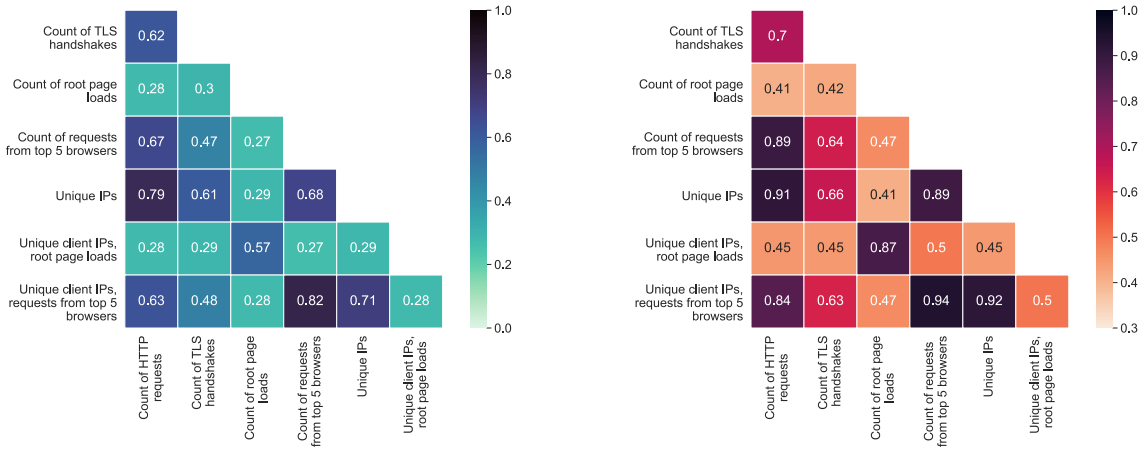
In response to increased skepticism of existing top lists, several alternative ranking schemes have been proposed [7, 16, 22, 23, 35]. Naab et al. propose replacing domain-based rankings with BGP prefix-based rankings by weighting and aggregating domains that belong to the same Internet prefix [23]. Aqeel et al. propose measuring popularity at a page-level instead of domain-level granularity [7]. Most recently, Xie et al. propose a "voting"-based methodology based on DNS request volume and frequency. It is unclear whether these proposals are more accurate and they have seen little widespread adoption.

In February 2021, Google Chrome began releasing rank order of magnitude popularity data (i.e., Top 1K, 10K, 100K, 1M, and >1M buckets) in their publicly accessible Chrome User Experience Report (CrUX) [8]. The dataset is curated monthly by aggregating browsing data from Chrome users who have *opted in* to history syncing, have no history sync passphrase set, and have usage statistic reporting enabled [8, 13]. CrUX is ranked by completed pageloads (measured by First Contentful Paint) and aggregated by web origin, and it adheres as closely as possible to user-initiated pageloads (e.g., it excludes traffic from iframes).

While Chrome provides only rank order magnitude data (e.g., Top 1K) rather than a full rank order list, we find that most research papers use top lists only as an ordered set of websites for study. We survey papers at USENIX Security, IMC, NSDI, SOUPS, NDSS, and WWW in 2021, and we find that of the papers using top lists, 50 (85%) use top lists only as a set, typically as a proxy for "popular websites"—only 9 (15%) use website rank directly. (A small handful of papers (5, 8%) leverage the lists as both a set and rank ordered list.) As such, we evaluate CrUX alongside other ranked lists as a possible alternative.

3 SERVER-SIDE POPULARITY METRICS

At the core of our analysis are server-side HTTP request logs collected from Cloudflare, a global content delivery network (CDN) and DDoS-mitigation provider that serves an estimated 10% of all web traffic and the plurality of websites in the Alexa Top Million (≈20%, Table 1). Cloudflare works by acting as the authoritative DNS provider and reverse proxy for customer websites. As such, Cloudflare's vantage point provides authoritative data on the number of requests that its customers receive. In this section, we detail how we leverage Cloudflare's server-side perspective to build a set of metrics that we can use to evaluate top lists of websites. In the next section, we describe how we use these metrics to evaluate top lists, their limitations, and the ethics of our study.



(a) Jaccard Index **(b) Spearman Correlation**

Figure 1: Intra-Cloudflare Metric Consistency—We compare seven metrics derived from Cloudflare request logs for measuring website popularity against one another to determine internal consistency: (1) All HTTP Requests, (2) TLS Handshakes, (3) HTTP Requests for Root Page, (4) HTTP requests limited to the five most popular browsers, (5) Unique IPs, (6) Unique IPs accessing the root page, and (7) Unique IPs limited to the five most popular browsers. We find that there is disagreement among metrics, likely because websites make a variable number of requests to load a page. Accordingly, we compare top lists against all seven Cloudflare metrics.

3.1 Filters and Aggregations

Our primary goal is to compare *rank-ordered lists* provided by sources like Alexa, Majestic, and Tranco to the authoritative vantage point that Cloudflare offers. However, there are a number of complications that make these comparisons difficult to conduct fairly. Chiefly, each top list uses a slightly different methodology, meaning that comparing a top list against just a single metric (e.g., all HTTP requests a domain sees) may erroneously conclude that the top list performs poorly, when the underlying comparison is itself unfair. In addition, there are many ways to "count" requests on the server-side—ranging from a raw count to unique client IPs—that may affect downstream results if not chosen appropriately. To address these methodological complications, we begin by considering two aspects of the metrics that could be computed from the Cloudflare perspective, which we refer to as *filters* and *aggregations*.

Filters. A filter is a condition imposed to collect a portion of the dataset, often for data quality. For example, one filter looks only at HTTP requests that returned a `200 OK` status message, while another restricts analysis only to TLS handshakes. In sum, we consider seven different filters:

1. All HTTP(S) Requests
   1.1. Limited to MIME-type `text/html` resources
   1.2. Limited to response code of 200
   1.3. Limited to non-null `Referer` header
   1.4. Limited to top 5 most popular browsers
2. TLS Handshakes
3. Root Page Loads (i.e., `GET /`)

Aggregations. An aggregation is a way to count server-side request logs in the dataset after filtering. For example, this may be the raw count (e.g., full number of requests) or the unique client IPs that connect to a website. We consider three aggregations:

1. Raw Count (e.g., number of requests)
2. Unique Client IPs (per day)
3. Unique Client IP and User Agent tuples

3.2 Evaluating Filters and Aggregations

Taken together, the Cloudflare data perspective affords us 21 filter–aggregation combinations. However, many of these combinations provide similar results, both in terms of being closely correlated with one another and evaluating top lists similarly. As our primary goal is to build a *diverse* set of metrics that cohesively measure performance, we first consider which filter-aggregation combinations provide the most variance in website popularity.

We compare each pair of filter-aggregations in two steps. First, we generate *rank-ordered lists*, similar to top lists, for each filter-aggregation combination. We then cross-compare each two filter-aggregation rank lists using two measures of similarity: Jaccard Index (JI) and Spearman's Rank Correlation (r_s). Jaccard Index measures the intersection of two lists divided by the union of the lists, providing a 0–1 score that quantifies the *unordered* similarity of two lists. Spearman's Rank Correlation is a nonparametric measure of how well the rank orders of two lists correlate, but operates on only their intersection. We show the full comparison between all 21 filter-aggregation combinations in Appendix A, Figure 8.

We observe significant redundancy in the 21 filter-aggregation combinations. Filtering out requests that lead to unsuccessful responses (i.e., non-200 HTTP status code) does not appreciably affect results compared to all HTTP requests because the vast majority of requests are successful ($r_s = 0.97$, $JI = 0.84$). Requests with an empty or missing `Referer` header is similar to requests from the

top five browsers ($r_s = 0.92$, $JI = 0.77$). We focus on the latter since it is a more direct measure of browsing behavior. We exclude `text/html` filter, because it acts similarly to TLS handshakes and requests from top browsers. Across all request aggregations, unique client IP addresses is nearly identical to our unique (IP, User Agent) aggregation. We choose the simpler of the two—unique IP address.

3.3 Selecting Final Measurements

We select seven final filter-aggregation combinations in Figure 1 that capture the most diversity in our dataset: (1) all HTTP(S) requests, (2) HTTP(S) requests from top five browsers, (3) HTTP(S) requests for root page, (4) TLS handshakes, (5) unique client IP addresses per day, (6) unique client IP addresses requesting root page, and (7) unique IP addresses from top five web browsers.

For these seven metrics, both list composition and rank varies between Cloudflare metrics (Figure 1), but in all cases, we see *moderate* to *very strong* rank correlation between metrics and 0.28–0.82 Jaccard Indices. There is strong correlation between the metrics found to be most similar by Jaccard Index and by Spearman Correlation (i.e., metrics that have high set intersection also have high Spearman rank correlation).

The two metrics with the lowest correlation are *all HTTP(S) requests* and *root page requests* ($r_s = 0.41, JI = 0.28$). This is not inherently surprising since websites make a varying number of sub-requests to load additional web resources. However, we note that the metrics bookend the number of *page loads* for a website because all page loads result in at least one request and a root page load is inherently a page load. TLS handshakes is more closely correlated with both since multiple HTTP requests can be sent in a single TLS session and each request requires a TLS handshake for HTTPS websites.

As we will show in Section 5, there is *perfect agreement* between all four request-based metrics when rank ordering the accuracy of top lists by inclusion of popular domains (i.e., $r_s = 1.0$ for all pairs). We argue that given this agreement and the fact that *all requests* and *root page requests* over- and under-estimate page loads, that the set of request-based metrics together can serve as a rough estimate for the *page loads* for each site. We further argue that our client-IP roughly models the number of unique visitors that a website serves per day, especially when considering that our unique (user agent, IP) metric is nearly identical ($r_s = 0.99$, $JI = 0.95$).

3.4 Summary

Cloudflare metrics vary when compared against each other. This is not unexpected and we optimize for a set of metrics to provide differing perspectives on what it means for a website to be popular. Two of our metrics, (3) Root Page Loads and (1) HTTP Requests, are chosen to be lower and upper limits on the number of page loads that a site receives since total page loads cannot exceed the number of requests a site receives and page loads cannot be smaller than the number of root page loads. If a top list is more closely correlated with one Cloudflare metric than another, it may simply indicate a difference in popularity metric. However, if top lists correlate similarly with both bookends, they likely serve as a rough indicator for the number of page loads that a site receives. In addition, our client IP metrics serve as an indicator for the number of visitors of each website. Together, these popularity criteria form a framework to benchmark top lists against.

Top List	Rank Magnitude			
	1K	10K	100K	1M
Alexa	14.97	23.16	26.63	23.12
Majestic	10.12	15.86	23.44	17.58
Secrank	0.57	3.65	6.37	7.8
Tranco	9.98	15.69	24.83	19.65
Trexa	11.62	18.75	25.19	21.5
Umbrella	1.99	4.09	6.75	10.86
Chrome UX Report (CrUX)	24.0	31.97	30.67	23.57

Table 1: Cloudflare Coverage of Top Lists—Percent of websites in the top lists that we evaluate whose content is served by Cloudflare. Cloudflare serves the largest fraction of top sites compared to any other single provider.

4 EVALUATION METHODOLOGY

In the last section, we derived a set of server-side metrics that can be used to used to measure the relative popularity of websites served by Cloudflare. In this section, we describe how we use this data to evaluate the accuracy of top website lists. We start by discussing how we collect and normalize lists of popular websites, and then we present our methodology for using our limited perspective into only Cloudflare sites to evaluate broader lists.

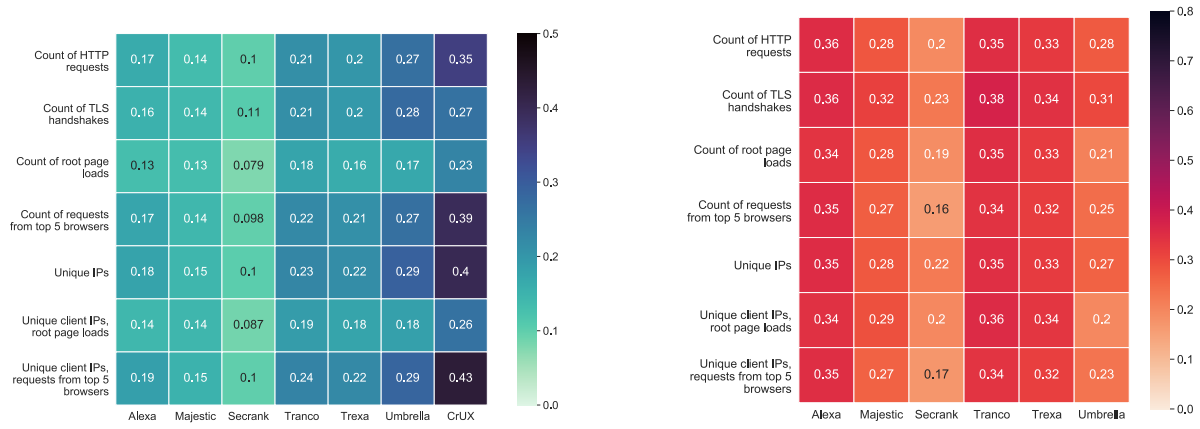
4.1 Collecting Top Lists

We retrieved daily snapshots of Alexa, Umbrella, and Majestic from the archive provided by Scheitle et al. [30], Tranco rankings from the project's archive [31], and Secrank rankings from the project's archive [28] for February 1–28, 2022. Using these rankings, we replicated the Trexa list's construction algorithm as specified by Zeber et al. [32]. We retrieve the public February 2022 CrUX global data from their public BigQuery project [9]. We leverage Cloudflare request data and Chrome telemetry data from the same date range. Unless noted otherwise, our evaluation is based on data from this time period; we average the results across days in the month.

4.2 Normalizing List Formats

We can only compare lists fairly if they contain the same representations of names and websites. Empirically, we find that Alexa, Majestic, Secrank, Tranco, and Trexa are aggregated by domain name, near identically to the Mozilla Public Suffix List (PSL), as shown in Table 2. On the other hand, Umbrella is aggregated by fully qualified domain name (FQDN), which is natural given that the list is of the most queried names, not the most popular websites. CrUX is aggregated by web origin (e.g., `https://google.com`).

These discrepancies complicate directly comparing lists, but without raw data, we cannot recompute lists using the same aggregation function. Instead, we group names by PSL-defined domain and choose the smallest rank (i.e., most popular) value as the rank for each domain to normalize list formats. This affects the Alexa, Majestic, Seclist, Tranco, and Trexa lists very little since they are aggregated by domain, which are nearly all public suffixes. This normalization affects Umbrella and CrUX lists the most since they



(a) Jaccard Index **(b) Spearman Correlation**

Figure 2: Correlation Between Top Lists and Cloudflare—We evaluate top lists against our seven Cloudflare metrics to evaluate their accuracy. We find that list overlap between Cloudflare domains in top lists and Cloudflare request-based metrics is poor, and that CrUX most closely matches Cloudflare for all seven metrics based on Jaccard index (an intersection-based metric) while Alexa and Majestic perform worst. Only CrUX achieves Jaccard indices comparable to agreement between the Cloudflare metrics in Figure 1a. List ordering between top lists and Cloudflare metrics shows at best weak correlation, with Alexa performing the best and Umbrella and Majestic performing the worst. Note: we cannot compute Spearman correlations for CrUX since it is rank-magnitude bucketed.

Top List	Rank Magnitude			
	1K	10K	100K	1M
Alexa	0.3	0.32	1.05	2.31
Majestic	5.87	1.30	0.28	0.10
Secrank	0.0	0.0	0.0	0.0
Tranco	0.0	0.0	0.0	0.00
Trexa	0.18	0.19	0.44	1.31
Umbrella	71.00	77.05	78.25	74.11
Chrome UX Report (CrUX)	75.4	72.09	70.54	66.49

Table 2: Percent of Domains deviating from Public Suffix List—Our PSL-based list normalization affects Alexa, Majestic, Secrank, Tranco, and Trexa relatively little, but could cause us to underestimate the accuracy of Umbrella and CrUX.

are aggregated by name rather than domain. Despite this methodology worsening their accuracy, we show in the next section, that both Umbrella and CrUX evaluate to be more accurate than the other lists. Thus, the discrepancy does not meaningfully change our conclusions or recommendations. Without normalization, all correlations are lower and this appears to be a strictly worse alternative.

4.3 Evaluating Lists Against Cloudflare

We hope to evaluate top lists against Cloudflare using the same metrics as we compared Cloudflare metrics against one another: Jaccard Index and Spearman Rank Correlation. As before, Jaccard Index (JI) quantifies the number of elements shared between lists; Spearman's Rank Correlation (r_s) quantifies the correlation between the ranks of shared elements.

However, unlike in the last section, because Cloudflare serves traffic for only a subset of top sites (Table 1), we cannot directly compare the two ranked lists. To build comparable lists of sites, we filter out non Cloudflare-sites from each top list and compare the subset of Cloudflare sites against the same number of top sites from Cloudflare. For example, if n of Alexa Top Million sites are powered by Cloudflare, we compare that set of n ranked sites against the top n Cloudflare sites. We use the same methodology whether considering the Top 1K, 10K, 100K, or 1M sites. This method would provide only very rough results if Cloudflare served a small fraction of websites in each top list, but we find that in most cases, Cloudflare serves hundreds of thousands of the top sites, which provides enough signal to show meaningful differences between top lists.

To filter top lists down to only Cloudflare-powered sites, we perform a HTTP `HEAD` request against each website in our top lists, and remove any website that does not include the `cf_ray` HTTP header that Cloudflare includes on any website that they proxy and serve authoritatively. We show the number of websites served by Cloudflare in Table 1.

4.4 Interpreting Results

There do not exist recommendations for interpreting list intersection; we caution readers that Jaccard Index often appears pessimistic at first glance. For example, if two lists of 100 websites have 90 sites shared, $JI = 0.82$. Spearman's correlation coefficients range from $[-1, +1]$, indicating a positive or negative monotonic correlation. Interpretation is typically similar to that of Pearson correlation: <0.10 is negligible, 0.10–0.39 weak, .40–.69 moderate, 0.70–.89 strong, and >.90 very strong. All p-values for Spearman's rank correlations that we present are significant ($p \ll 0.05$).

We suggest that readers primarily focus on Jaccard index as the more important metric for several reasons: (1) researchers primarily use top lists as an *unordered set* of websites rather than investigate the ranks of individual sites, (2) rank correlation is secondary to inclusion when studying popular websites (i.e., if a top list fails to include the most popular sites, it is typically not meaningful that it correctly ranked the subset of sites it finds), and (3) set intersection can be computed for all top lists in our study. Because Google CrUX provides only rank order magnitude, we cannot compute Spearman rank correlation to evaluate it against our Cloudflare metrics.

4.5 Limitations

There are several limitations to our study's methodology:

Incomplete coverage. Our study is premised on data from Cloudflare, which serves only a fraction of top websites. While this is a large enough sample to detect differences in top lists, it is not the full set of popular websites nor a random sample. There is likely some bias in the sites that choose to use Cloudflare. For example, none of the top ten sites use Cloudflare. It is unclear whether this could cause bias in our evaluation.

Server-side request-based metrics. We compare top lists to a set of metrics that we derive from HTTP requests, which imperfectly estimate web requests. Further, without exact knowledge of how each top list is constructed, some lists may have more or less similar metrics to ours. Both of these misalignments could affect our evaluation.

Multi-CDN site configurations. Cloudflare may not host all resources on popular websites and websites could potentially use multiple CDNs to serve content. Given that Cloudflare authoritatively serves DNS for domains, and provides multi-CDN functionality to only a small number of customers, we know this configuration is rare and do not expect it to noticeably affect our results.

Temporal bias. Our data collection period is relatively short, occurring during the COVID-19 pandemic and Russia's full-scale invasion of Ukraine. These black swan events may affect the popularity of websites during the study time period. It is likely worth re-measuring the accuracy of top lists again in the future.

4.6 Ethical Considerations

To protect the privacy of Cloudflare customers and their customers' clients, the Stanford research team did not have access to any individual customer or website records or any identifying data (e.g., popular domains). All analysis was performed by providing analysis scripts that output aggregate summary results (e.g., Jaccard indices and Spearman correlations of a top list with internal Cloudflare metrics) to Cloudflare. Results were inspected by Cloudflare staff before being shared to ensure that no customer or user information was leaked. No analysis required Cloudflare to investigate individual websites or clients, and all analysis was performed in accordance with Cloudflare's privacy policy. No additional data was collected or stored by Cloudflare for the purpose of our study. Because the Stanford research team analyzed only aggregate data "without any individually identifiable information", our analysis did not constitute human subjects research per our institution's IRB guidelines.

4.7 Summary

Evaluating top lists fairly against our Cloudflare baseline requires some nuance. Top lists do not present sites in a directly comparable format, but we find that truncating all list elements to PSL-defined domain name provides a favorable common denominator for evaluation. To account for Cloudflare's incomplete coverage of the Internet, we formulate our list comparisons to only operate on the subset of sites controlled by Cloudflare. In the next section, we perform our evaluation of top lists against our set of metrics, ultimately finding that despite our metrics showing significant variance amongst themselves, they evaluate the accuracy of top lists with perfect agreement.

5 EVALUATING TOP LISTS

In this section, we evaluate top website lists against our Cloudflare popularity metrics. We find that lists evaluate consistently against our set of Cloudflare page views and client IP metrics. While most lists have relatively low raw intersection, Google CrUX performs notably better than other lists, inline with the differences amongst individual Cloudflare metrics. Beyond CrUX, we find inconsistent evaluation results when evaluating the ranks of individual sites.

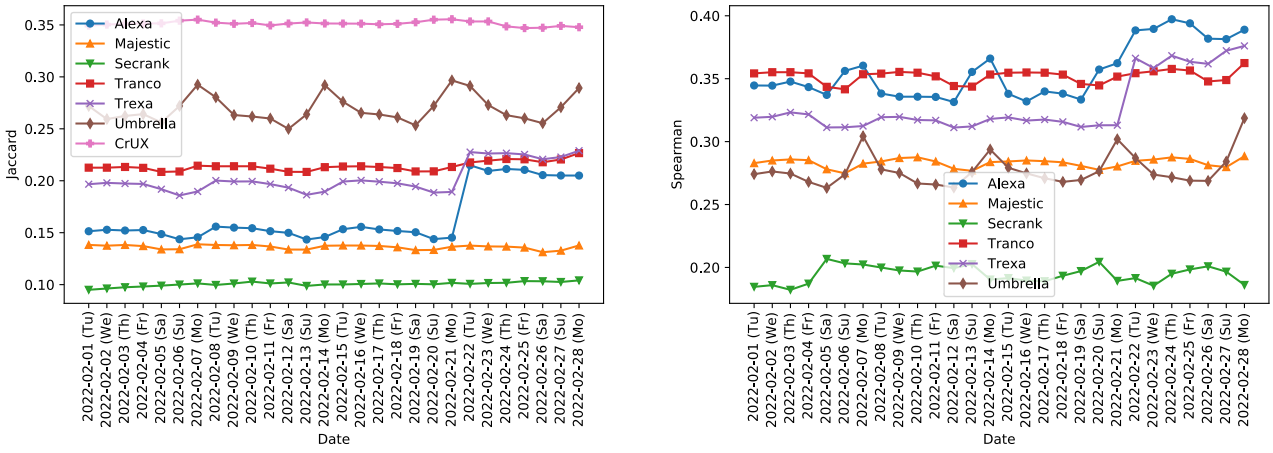
5.1 Websites Captured By Top Lists

We first compare the set of websites captured by top lists with our seven Cloudflare request and requestor metrics. Surprisingly, we find that our seven Cloudflare metrics rank order the accuracy of top lists perfectly consistently ($r_s = 1.0$ for all pairs of metrics) when measuring unordered list intersection (Jaccard Indices). Restated, our set of Cloudflare metrics agree on which top lists best capture the *unordered set* of most popular sites.

There is relatively small intersection between Cloudflare-powered sites in top lists and the Cloudflare sites we expect, even in light of differences between Cloudflare metrics (Figure 2). However, given the non-negligible disagreement amongst Cloudflare metrics, imperfect measurement methodology, and lack of interpretation guidelines for Jaccard Index, we suggest that readers consider relative values rather than evaluate the raw numerical values.

Secrank ($JI = 0.08$–0.11), followed by Majestic ($JI = 0.13$–0.15) and Alexa ($JI = 0.13$–0.19) show the least overlap across all request and client metrics (Figure 2a). CrUX evaluates *notably* better than all lists by intersection ($JI = 0.23$–0.43). Umbrella, which comes in second place ($JI = 0.17$–0.29), is notably worse than Chrome, but better than Alexa and Majestic. Tranco and Trexa approximately average the scores of the lists they combine, falling in the middle. The Jaccard Indices we observe for CrUX (0.23–0.43) fall inline with the differences between Cloudflare metrics (0.27–0.82), strongly hinting that it captures popular websites and differences are within realm of error that arises from methodological differences in measuring popularity—it would be unfair to expect an external list to have better agreement with individual Cloudflare metrics than they do with one another. Umbrella in the best case barely reaches the lower end of the intra-Cloudflare JI range. None of the other lists reach that range with $JI = 0.08$–0.24.

It is difficult to ascertain exactly why certain lists capture popularity better than others, especially when many methodological details are not public. Secrank is based on DNS data from China,



(a) Jaccard Index **(b) Spearman Correlation**

Figure 3: Popularity Metrics Over Time—For each day in February 2022, we compute correlation scores between top lists and all HTTP requests for the top 1M domains. (Note that the CrUX list is fixed since it is aggregated monthly.) We find that list correlations are somewhat periodic, with Jaccard indices best on weekdays and Spearman correlations best on weekends for most lists, but that variations in correlation scores largely do not affect which lists more closely approximate server-observed requests.

which introduces significant geographical bias (Section 6.3). Majestic computes popularity based on backlinks to websites, and there is little evidence to support that the number of links to a websites correlates strongly with page views or number of visitors. Alexa utilizes both page views and visitors, but has a small install base, and is likely biased based on the browser extensions that the company has partnered with. In contrast, Umbrella and CrUX are computed off of a significantly larger set of users, and CrUX is computed directly based off of Google Chrome usage.

5.2 Evaluating Rank Order

Rank correlation between Cloudflare and top lists is less consistent than our unordered list intersection metric, likely because the set of intersecting websites is relatively small. Spearman Ranks do not correlate perfectly by list, but we find that Alexa, Tranco, and Trexa show the highest rank correlations and that Umbrella, Majestic, and Secrank perform poorly (Figure 2b). Secrank is the least correlated with all Cloudflare metrics by both JI and r_s. We cannot evaluate CrUX by rank correlation because it provides only rank-order magnitude rather than individually ranked sites.

It is initially surprising that Umbrella does poorly in rank order correlation relative to Alexa since it had a relatively high intersection with popular Cloudflare sites. Part of this could be due to how the list is constructed and our measurement methodology. Umbrella has been observed to break ties in list ordering with long strings of alphabetically sorted domains [25], which may drive the poor Spearman correlation among a reasonably large list intersection. Further, our process to normalize site names likely disadvantages Umbrella's origin-based rankings and artificially weakens the Spearman correlation.

Inaccuracies could also arise from our Cloudflare metrics being web-driven, whereas Umbrella is DNS based. Popular domains queried by a large number of users likely bubble to the top, but caching, TTLs, and other DNS complexities prevent capturing fine grained popularity. There could also be bias in data collection locations—Cisco Umbrella has a significant enterprise user base. Investigating our Chrome data broken down by country, we see that Umbrella has considerably better rank correlation for clients in the United States than other top lists, but worse rank correlation in other countries (Section 6.3). Umbrella may see the presence of popular sites in other regions, but be unable to discern their exact popularity.

Ultimately, we shy away from drawing significant meaning from rank correlations for several reasons: (1) the intersections between non-CrUX top lists and Cloudflare are small (e.g., Alexa has $JI = 0.13$–0.19) and it is unclear if there is meaning in the rank correlation metric within the small intersection, (2) the values show little variance between lists and it is unclear at what point they are more driven by noise than signal, and (3) researchers typically use lists as an unordered set of websites, which Jaccard Index best captures. For these reasons, we still consider Umbrella to be the best alternative to CrUX, despite its lackluster rank order evaluation.

Summary. Despite a messy set of metrics from Cloudflare, we find that CrUX is *notably better* at capturing popular websites than other top lists as defined by visit and visitor metrics. Umbrella does not portray itself as capturing the most popular websites—it captures the most queried domains—but we find that it better captures the set of popular websites than alternatives. However, Umbrella is not accurate enough to capture the ranks of individual websites. This is likely due to combination in how it ranks large sets of websites alphabetically within the list when it cannot determine rank accurately as well as biases in Umbrella's user base. Majestic's link-based methodology does not capture well the set of sites deemed popular by visits and unique visitors. Alexa does

better than Majestic, but notably worse than CrUX, and regardless, will soon be deprecated.

5.3 Small Inaccuracies or a Major Problem?

Although top lists offer ranked lists of domains, most researchers analyze domains as an unordered set of rank-magnitude buckets. For example, researchers may be interested in studying some property of the top 10K or top 100K websites. As such, a pressing question is: to what extent do top lists place domains into *incorrect* rank-magnitude buckets? We investigate the *movement* of domains from one rank-magnitude bucket to another to understand whether differences could alter downstream research results.

We begin by identifying the sets of domains for four popular rank-magnitude buckets—1K, 10K, 100K, and 1M—from the Cloudflare list. As noted in Section 4, many of the Cloudflare metrics do not agree—to account for this, we restrict our analysis to the set of domains that two metrics that bookend pageloads (root page fetches and all HTTP requests) both place into a given bucket. We then compare these rank-buckets to those created by each top list we study. We only consider movement of domains that are Cloudflare operated. We visualize rank-movement for the Alexa list and CrUX list in Figure 5. The results for Majestic, Tranco, Trexa, and Umbrella follow a similar pattern to Alexa and produce similar figures. We discuss both below:

Alexa Top Million. There is a significant amount of rank magnitude movement between the intersection of Cloudflare and Alexa. For instance, of the 1,790 domains we measure in the Alexa top 10K, 70% of them are ranked by Cloudflare in a lower rank-magnitude bucket, and 27.2% of them are ranked by Cloudflare in a bucket *two or more orders of magnitude* less popular. This phenomenon is even more striking at smaller rank-magnitudes: 87.1% of the Alexa top 1K are overranked (based on an average of 210 sites), and 56.7% are overranked by two or more orders of magnitude. These numbers likely underestimate how poorly the top lists comport within each rank bucket, because the rank magnitude of websites that strictly intersect with Cloudflare are likely lower in the global top list. Researchers who use top subsets of Alexa (or Majestic, Tranco, Trexa, and Umbrella rankings) must thus contend with large volumes of non-mainstream sites cluttering their datasets.

Chrome User Experience Report (CrUX). The CrUX list much more closely approximates the Cloudflare list by rank-magnitude movement: 47.1% of the 1410 domains in the CrUX top 10K are overranked compared to Cloudflare, and only 1% of them are overranked by two or more orders of magnitude. While CrUX is not a perfect match with Cloudflare, it shows significantly more rank-magnitude agreement than the other top lists.

5.4 Temporal Stability

We next turn to study if top list performance is affected by window of measurement or changing the platform that measurements are drawn from. All of our measurements presented thus far are computed on a daily basis and averaged over all days in February 2022. However, top lists vary significantly by day [27], and may evaluate differently based on the day of measurement. To investigate these differences, we quantify the stability and periodicity of correlation between top lists and our Cloudflare metrics on a daily basis (Figure 3). Jaccard intersection remains temporally stable across all lists except for Cisco Umbrella, which exhibits weekly periodicity in Jaccard index. Spearman correlations are periodic and less stable, with Alexa and Umbrella showing higher accuracy on weekends. This may be because Alexa is computed primarily from browser extension data, which is likely less prevalent in corporate environments. The remaining lists show some degree of periodicity; however, they are largely consistent day over day. The order of top lists from most to least well correlated is also largely consistent over time. However, we see Alexa improve in accuracy, by both intersection and rank order correlation, in late February. It is unclear why—no announcements have been made about any methodological change.

6 BIASES IN MISSED WEBSITES

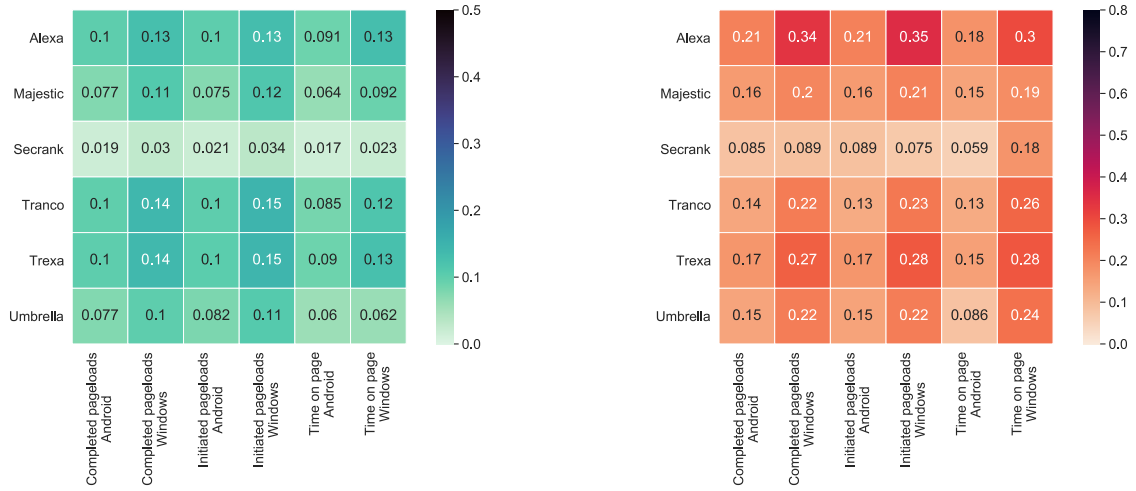
In the last section, we quantified the accuracy of top lists. Next, we explore the hypothesis that list inaccuracy is due to, at least partially, systematic biases in list composition (i.e., errors are not simply random). Specifically, we consider whether lists are biased toward certain client platforms, client countries, or website categories.

6.1 Google Chrome Telemetry Data

Given the accuracy of CrUX demonstrated in the last section, we worked with Chrome to better understand the bias in client platform and country that other lists exhibit. Chrome specifically provided us with rank-order popularity lists for several client telemetry metrics that are not publicly accessible through CrUX [8]: (1) initiated page loads, (2) completed page loads, and (3) total time on sites aggregated by country and OS platform from February 1–28, 2022. The public CrUX data used previously in this paper is based on completed page loads.

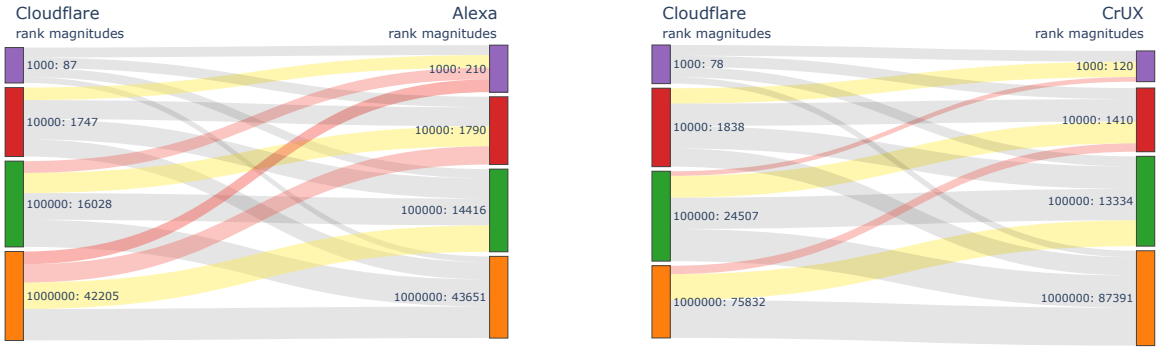
We consider 11 countries: 10 countries that the Chrome team designated as providing high fidelity data and geographic diversity (Brazil, Germany, Egypt, United Kingdom, Indonesia, India, Japan, Nigeria, the United States, and South Africa), plus China as a comparison point for Secrank. We focus on one desktop (Windows) and one mobile (Android) platform because the Chrome team indicated that these have the largest, most representative Chrome install base. Note that Chrome mobile telemetry captures traffic only from the browser and native Android apps that use Custom Tabs and WebAPKs, not from most native apps [13]. Also, Chrome telemetry excludes visits to *non-public domains*—domains that are not hyperlinked from public websites or specify that they may not be crawled per `robots.txt` [13]. We do not evaluate CrUX against these additional Chrome-based metrics because they are derived from the same data source.

We first consider the internal consistency of the three client metrics from Chrome (Figure 6). We find stronger correlation between Chrome metrics than between Cloudflare even though time and page and completed page loads are considerably different: Spearman correlations between Chrome metrics are strongly correlated (0.66–0.98). In almost all cases, Jaccard indices are notably higher than for Cloudflare (0.73–0.86). We note that since our Chrome lists are aggregated by client country and platform rather than globally, we compute the Jaccard index and Spearman correlation



(a) Jaccard by Platform **(b) Spearman by Platform**

Figure 4: Top List Performance by Platform—We compare top lists to Chrome data broken down by client platform, averaging results across client countries. Top lists best approximate client behavior sourced from desktop vs. mobile users. However, the delta between both platforms is small, suggesting that platform alone does explain why top lists poorly approximate client behaviors.

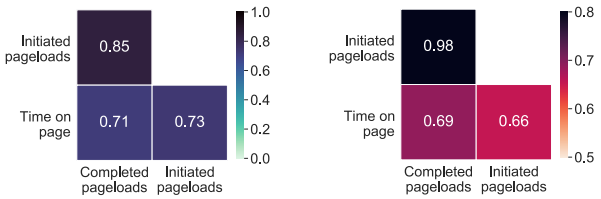


(a) Cloudflare and Alexa 1M **(b) Cloudflare and CrUX 1M**

Figure 5: Magnitude movements between Cloudflare, Alexa, and CrUX—For the set of sites in the Alexa 1M that Cloudflare metrics (1) and (3) place into the same rank-magnitude bucket, we compute the movement of those sites from their Cloudflare rank-magnitude bucket (left) to their Alexa rank-magnitude bucket (right). The width of each link is on a log scale. Yellow links indicate off-by-one rank magnitudes while red links indicate more drastic mismatches. We find that a large fraction of Alexa top-ranked domains are counted in a lower rank-magnitude bucket by multiple Cloudflare metrics. Results for Majestic, Tranco, Trexa, and Umbrella are very similar. CrUX exhibits significantly less of a mismatch with Cloudflare rank-magnitude buckets.

between popularity metrics for each country and platform and report average correlations across all (country, platform) pairs; we are unable to directly compute similarity between global lists. This process may raise correlations somewhat if it corrects for browsing behavior differences across countries and platforms.

Ethics. The Chrome data that we analyze is subject to several privacy protections. All data was collected and shared with us in aggregate form only. We received no data from Google about individual users or the raw amount of traffic that any site receives, and Google did not collect or analyze any additional data for this study. Instead, Chrome provided us with rank order list of most



(a) Jaccard Index **(b) Spearman Correlation**

Figure 6: Intra-Chrome Metric Consistency—We compare three metrics from Chrome telemetry for measuring popularity: (1) Completed Pageloads, (2) Initiated Pageloads, (3) Time On Page. Each value is averaged over all client platforms and countries.

popular websites by country and platform. To prevent inadvertently deanonymizing users, only websites with above a set threshold of unique visitors from each country are included; this threshold is not disclosed, but is set both to protect privacy and to ensure sufficient samples to be confident in the statistical distributions for included pages [13]. Similar to the public CrUX lists, the Chrome data we analyze is based on Chrome users who have opted into sharing URLs with Chrome and have usage statistic reporting enabled. Chrome's privacy policy explains in plain language under what circumstances URLs will be shared, including that data may be used for research and development, and users can control whether usage statistic reporting is enabled. No additional data was collected or stored by Google for the purpose of this study.

6.2 Client Platform Biases

We first investigate whether top lists are biased toward traffic from specific client platforms. To measure the role of client platform in top list performance, we compare top lists with ranked popularity data collected from Chrome telemetry on Windows and Android. Similar to Section 6.1, we compute correlations over (platform–country) pairs and average results across countries (Figure 4).

Top lists better approximate client behavior on desktop platforms than mobile platforms. This is true across every non-CrUX list and Cloudflare metric. Jaccard coefficients range from 0.017—0.1 for Android clients, less than 0.023—0.15 for Windows clients. These results are consistent for Spearman's rank coefficients, which range from 0.059—0.21 for Android clients, compared to 0.075—0.35 for Windows clients. In the most severe case, Alexa has Jaccard Indices for desktop users that are nearly double those of mobile. Unsurprisingly, given the methodological approach, we see the least difference for Majestic. We note that we do not evaluate CrUX here because it is based on the same data as we use for the evaluation. Furthermore, because these results are averaged over per-country lists rather than compared to global lists, the values in Figure 4 are not expected to appreciably outperform those in Figure 2; rather, we focus on the bias shown by the relative values in the heatmap, and observe that despite this bias, platform alone does not fully explain why top lists perform so poorly.

Cat.	Alexa	Majest.	Tranco	Trexa	Umbrel.	CrUX	Secrank
Gov't	–	5.45	17.62	6.81	2.3	5.74	0.24
News	–	1.6	2.87	2.51	0.69	3.93	0.16
Educ.	0.84	–	1.93	1.7	0.67	2.79	0.38
Science	0.87	0.77	1.85	1.66	0.81	1.09	0.41
News	–	–	–	–	–	–	–
Comm.	0.75	0.9	1.8	1.6	0.64	2.93	0.27
Bus.	0.75	0.84	1.72	1.5	0.66	1.33	0.19
Gaming	0.7	0.66	1.61	1.37	0.67	3.0	0.38
Kids	0.62	–	1.92	1.63	–	5.41	0.16
Life	0.67	0.68	1.6	1.38	0.53	3.17	0.19
Arts	0.63	–	1.64	–	0.57	2.64	0.14
Health	0.62	–	1.74	1.5	0.6	3.37	0.09
Blog	0.65	0.54	1.3	1.18	0.44	2.46	0.15
Sports	0.62	–	1.84	1.49	0.51	4.19	0.09
Travel	0.55	1.66	2.26	1.7	0.63	3.96	0.12
Shop	0.55	0.74	1.23	1.08	0.43	5.3	0.08
Cars	0.49	–	1.48	1.22	0.5	4.36	0.05
Adult	0.27	0.14	0.46	0.44	0.16	2.83	–
Abuse	0.26	0.15	0.62	0.52	0.41	0.5	0.3
Gambl.	0.22	0.23	0.38	0.36	0.13	1.84	0.08
Parked	0.11	0.03	0.2	0.19	0.15	0.2	0.2

Table 3: Odds of Website Inclusion by Category—We show the odds that a particular category of website is included by each top list. Top list has a unique set of categories that it is biased to include, however, some categories are universally included, such as government websites and news websites. Conversely, some categories are commonly excluded, such as adult or abuse (e.g., spam) websites. A missing entry means the regression result was not statistically significant at $p < 0.01$ with Bonferroni correction.

6.3 Client Country Biases

Next, we investigate top list biases by country, using the same Chrome data described in Section 6.1. Because global browsing behavior is unevenly distributed, there are two plausibly "correct" results for a globally aggregated popularity list: equally similar to all countries (weighting traffic by country of origin to counterbalance population differences), or systematically biased toward countries with larger user bases (weighting traffic equally irrespective of origin country). We observe neither pattern in the data.

Top lists exhibit noticeable and irregular geographic biases (Figure 7). For example, all top lists poorly represent Japan. Umbrella is biased toward the US, which makes sense given US-centric customer base. Secrank best matches China, though performs relatively poorly, and does extremely poorly elsewhere—likely due to its Chinese DNS data source. Majestic also best matches the US. Alexa performs best when compared to the US, China, and sub-Saharan Africa. Finally, the composite Tranco and Trexa lists inherit the biases of their inputs.

6.4 Category Modeling

In this section, we examine another potential bias: the type of the website itself. We label each Cloudflare-managed website using the domain categorization feature of Cloudflare's Domain Intelligence API [14], which has been demonstrated to be reasonably accurate [24]. We identify whether a top list erroneously *excludes*

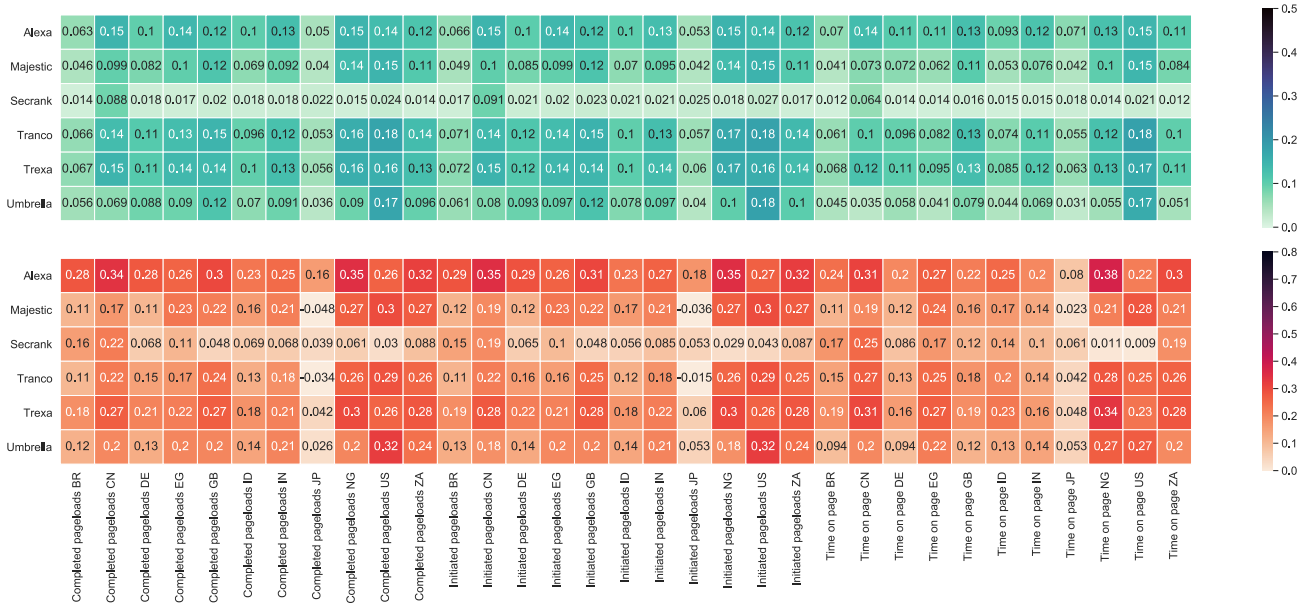


Figure 7: Top List Performance by Country—We compare top lists to Chrome data (Jaccard above, Spearman below) broken down by client country, averaging results across desktop and mobile platforms. Top lists demonstrate significant geographical bias. All top lists poorly match Japan. Secrank best matches China, Umbrella and Majestic best match the US, Alexa matches sub-Saharan Africa surprisingly well, and Tranco and Trexa inherit the biases of their component lists.

a popular website using a simple heuristic: we identify the *least* popular domain (D_{least}) in a top list L that appears in our Cloudflare list. We consider a domain D to be erroneously excluded if it receives more traffic volume than D_{least} based on HTTP requests but does not appear in L. For the scope of this analysis, we evaluate list inclusion on a single day: February 1, 2022. We restrict our analyses to consider only the Top 100K domains from Cloudflare, as inclusion rates significantly decrease at higher thresholds, which reduces the interpretability of our results.

Modeling Website Exclusion. To evaluate the impact that website category has on list exclusion, we perform a simple logistic regression analysis with website category as an input. For each domain in the Cloudflare list, we model the outcome of list exclusion as a binary outcome—1 if the domain was included by a top list L, and 0 if the domain was excluded by the top list. As an input to the regression, we provide the category C of the domain, represented by a one-hot encoded vector. We build a regression between Cloudflare and each top list we study. We report regression results as odds ratios, with the category in question measured against all other domains as a control (Table 3). For example, the odds of a government website being included by the Umbrella top list are 2.3 times that of a non-government website being included by the Umbrella list. All reported results are statistically significant with $p < 0.01$ and a Bonferroni correction of 22 (the number of website categories we consider).

Each top list is biased towards a different set of categories:

Alexa. Alexa is most biased away from adult (0.27× odds of inclusion), abuse (0.26×), gambling (0.22×), and parked domains (0.11×). Adult websites are likely not included because of Alexa's data collection methodology, which depends on data collected from installed browser extensions. Prior research has demonstrated that users typically visit adult sites in a private browser mode, where browser extensions are disabled by default [15]. For other categories, like abuse and gambling, we do not have a clear reason why they are underrepresented. We note that almost every category appears to be underrepresented in Alexa because in aggregate, inclusion rates are relatively low.

Majestic. Majestic skews towards government, news, and travel websites, for which the odds of inclusion range from 1.6–5.45×. Outside of these categories, almost every other category of website is underrepresented. The least likely to be included in the Majestic list are adult websites (0.14×) and abuse websites (0.15×).

Secrank. The Secrank list has poor odds of inclusion for nearly every category, with all but two having statistically significant underrepresentation (0.05–0.41×). Intuitively, this implies that the Secrank list broadly misses many domains from the Cloudflare list, suggesting widespread underrepresentation of popular domains.

Umbrella. Umbrella skews towards government websites (odds of inclusion are 2.3×), but underrepresents every other category.

Tranco and Trexa. The Tranco and Trexa lists follow similar inclusion patterns. In particular, we observe statistically significant differences in the odds of inclusion across all but one category in the Tranco list (News) and all but two in the Trexa list (News and Arts). Given that both the Tranco and Trexa lists have the second highest rates of inclusion from the Cloudflare list (57%, 54%), the odds of inclusion are high in almost every single category. The most

notable category is Government, for which the odds of inclusion in Tranco are 17.62 times the odds of a non-Government website being included. However, Tranco and Trexa both regularly exclude adult websites and gambling websites, which aligns closely with Alexa, Majestic, and Umbrella.

CrUX. The CrUX list has the highest rate of inclusion with Cloudflare (71%), and nearly every category of website has high inclusion rates. Notably, CrUX is the only top list to also account for adult websites and gambling websites in their list.

Taken together, our analysis reveals certain categories of websites (e.g., adult and gambling) are often underrepresented in top lists. CrUX does not have this pitfall, and it has the best overall odds of inclusion among top lists. Beyond this, our results show that while amalgam lists may be an effective strategy to prevent against adversarial manipulation [18], the biases present in their underlying lists are not mitigated by any one aggregation strategy—for example, Tranco and Trexa still suffer from underrepresented adult websites, despite having high rates of inclusion broadly.

7 DISCUSSION AND CONCLUSION

For more than twenty years, the Internet measurement community has approximated the web using lists of popular websites like the *Alexa Top Million*. Recent studies have cast doubt on the accuracy of such lists and, by proxy, the research results derived from them [27]. In addition, Amazon's recent deprecation of the Alexa Top Million—the list used by the vast majority of prior research studies—forces researchers to choose an alternative. But, despite these concerns, it has been impossible to directly evaluate the accuracy of top website lists or to sensibly select a replacement for Alexa without ground truth data about websites—a daunting task given the distributed nature of the web.

For better or for worse, the web has become more centralized over the past decade. A small handful of content providers like Amazon, Akamai, Cloudflare, and Google now serve a significant fraction of popular websites and users have gravitated towards a select few web browsers. This concentration provides us with a unique new opportunity to evaluate lists of popular websites. In this paper, we partnered with Cloudflare, the content provider that serves—by far—the largest fraction of top websites and Google Chrome, the most popular web browser, to evaluate the accuracy of lists of top websites.

Even with these industry datasets, the results we find are messy and inconsistent. To start, there is no shared definition of website popularity. Each list employs a unique, and often proprietary, methodology to compute popularity. In many cases, lists are ranking different objects, which range from web origins to DNS names. Further, it is difficult to reverse model even simple metrics like page loads using server requests as seen by a server provider. Despite these complexities, we find several notable conclusions.

Using a set of metrics from Cloudflare that estimate *page loads* and *unique visitors*, we find Google Chrome's recently released CrUX dataset captures the *unordered set* of most popular websites *significantly* more accurately than other top lists, with correlations inline with the differences we see amongst multiple measures of popularity derived from the same Cloudflare data. No other top list enters this range. This, paired with the internal consistency of Chrome metrics, suggests that Chrome does not simply use a metric more similar to Cloudflare's, but rather that their data is more accurate.

Chrome provides only rank order magnitude buckets. While initially appearing to be a shortcoming of the CrUX dataset, the lack of individual site ranks is not an issue for most research studies. We survey past work that uses top lists and find that the vast majority do not use site ranks, but rather use top lists as an *unordered set* of websites to study. As such, the structure of the CrUX dataset is typically suitable for research use cases. After CrUX, we find that the Umbrella Top Million—a list that measures popular names rather than websites—next best captures the most popular websites, but that their methodology may not be accurate enough to capture the relative popularity of individual sites.

It is difficult to ascertain exactly why some lists are more accurate than others, but our work documents that this is not simply due to random noise. We find that there are biases in top lists, but we do not answer conclusively why these biases arise. Other work [24] provides potential clues. For example, differences in dominant use cases between mobile and desktop browsers may partially explain site category bias; we leave it to future work to pursue these leads. Future work may also be able to combat these biases and to build more representative sets of websites (e.g., by combining data from multiple accurate data sources or hardening accurate data sources against external manipulation).

ACKNOWLEDGEMENTS

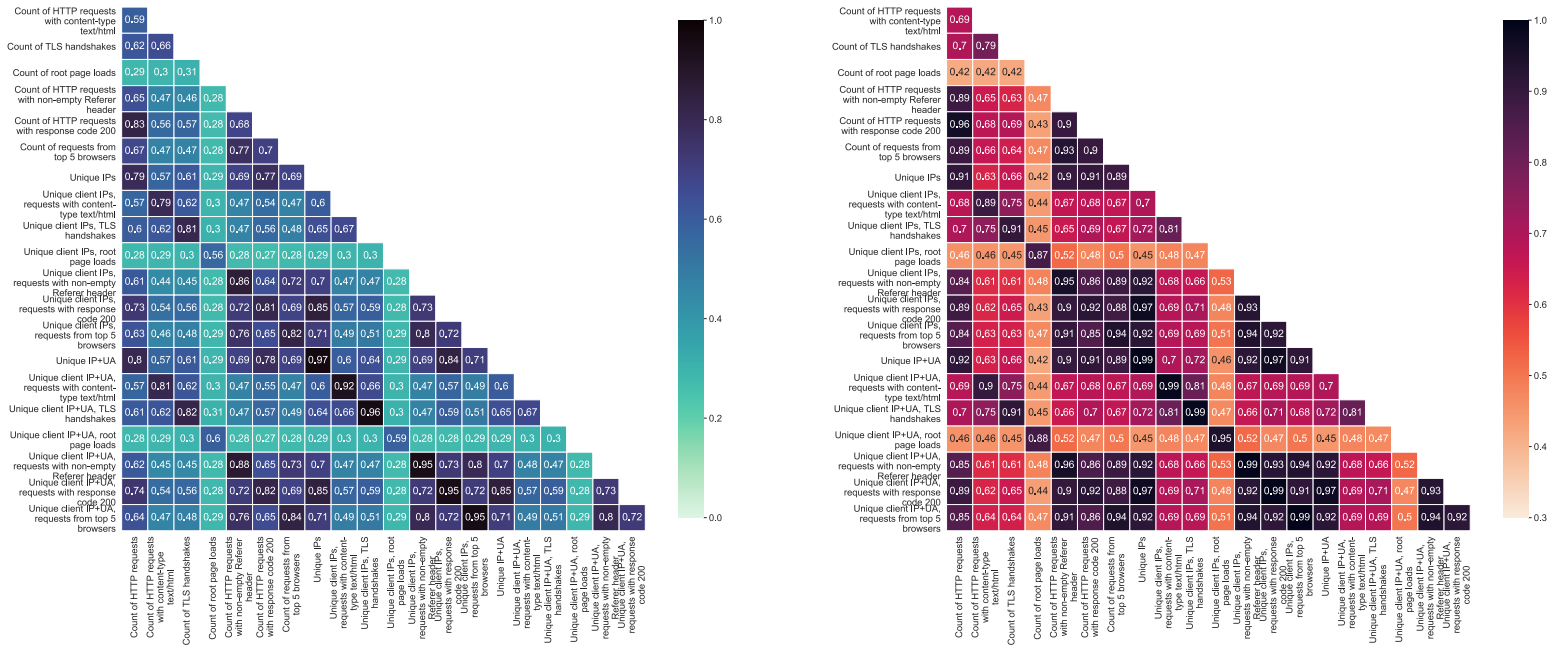
We thank the Google Chrome team for providing supplementary data. We thank Emily Austin, Michael Bailey, Aurore Fass, Liz Izhikevich, Marc Light, Parker Ruth, Nick Sullivan, and Maya Ziv for providing feedback on the paper. We also thank our shepherd Jingjing Ren and the anonymous reviewers for their helpful comments. This work was supported in part by an NSF Graduate Research Fellowship DGE-1656518 and a gift from Digicert, Inc.

REFERENCES

[1] Alexa top million. https://www.alexa.com/.
[2] Alexa. Free website analytics – how marketers use Alexa rank today. https://blog.alexa.com/free-website-analytics-alexa-rank/.
[3] Alexa. How are Alexa's traffic rankings determined? https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-.
[4] Alexa. Top 6 myths about the alexa traffic rank. https://blog.alexa.com/top-6-myths-about-the-alexa-traffic-rank/.
[5] Alexa. We will be retiring the Alexa.com APIs on December 15, 2022. https://support.alexa.com/hc/en-us/articles/4411466276375.
[6] Alexa. Your top questions about Alexa data and ranks, answered. https://blog.alexa.com/top-questions-about-alexa-answered/.
[7] W. Aqeel, B. Chandrasekaran, A. Feldmann, and B. M. Maggs. On landing and internal web pages: The strange case of Jekyll and Hyde in web performance measurement. In *ACM Internet Measurement Conference (IMC)*, 2020.
[8] Chrome User Experience Report. https://developers.google.com/web/tools/chrome-user-experience-report.
[9] Chrome User Experience Report: Getting started. https://developers.google.com/web/tools/chrome-user-experience-report/bigquery/getting-started.
[10] Cisco Umbrella 1 million. https://umbrella.cisco.com/blog/cisco-umbrella-1-million.
[11] Cloudflare Radar. https://radar.cloudflare.com/.
[12] Comscore latest rankings. https://www.comscore.com/Insights/Rankings.
[13] CrUX methodology. https://developer.chrome.com/docs/crux/methodology/.
[14] Cloudflare API v4 documentation: Get domain details. https://api.cloudflare.com/#domain-intelligence-get-domain-details.
[15] X. Gao, Y. Yang, H. Fu, J. Lindqvist, and Y. Wang. Private browsing: An inquiry on usability and privacy protection. In *13th Workshop on Privacy in the Electronic Society*, 2014.

[16] J. Kline, A. Aelony, B. Carpenter, and P. Barford. Triangulated rank-ordering of web domains. In *International Teletraffic Congress*, 2020.
[17] V. Le Pochat, T. Van Goethem, and W. Joosen. Evaluating the long-term effects of parameters on the characteristics of the Tranco top sites ranking. In *12th USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, 2019.
[18] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Network and Distributed Security Symposium (NDSS)*, 2019.
[19] B. W. N. Lo and R. S. Sedhain. How reliable are website rankings? implications for e-business advertising and Internet search. *Issues in Information Systems*, 7(2):233–238, 2006.
[20] The Majestic million. https://majestic.com/reports/majestic-million.
[21] Majestic million CSV now free for all, daily. https://blog.majestic.com/development/majestic-million-csv-daily/.
[22] F. Marquardt and C. Schmidt. Don't stop at the top: Using certificate transparency logs to extend domain lists for web security studies. In *IEEE 45th Conference on Local Computer Networks*, 2020.
[23] J. Naab, P. Sattler, J. Jelten, O. Gasser, and G. Carle. Prefix top lists: Gaining insights with prefixes from domain-based top lists on DNS deployment. In *ACM Internet Measurement Conference (IMC)*, 2019.
[24] K. Ruth, A. Fass, J. J. Azose, M. Pearson, E. Thomas, C. Sadowski, and Z. Durumeric. A world wide view of browsing the world wide web. In *ACM Internet Measurement Conference (IMC)*, 2022.
[25] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda. Clustering and the weekend effect: Recommendations for the use of top domain lists in security research. In *Passive and Active Measurement (PAM)*, 2019.
[26] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda. Getting under Alexa's umbrella: Infiltration attacks against Internet top domain lists. In *Information Security*, 2019.
[27] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez. A long way to the top: Significance, structure, and stability of Internet top lists. In *ACM Internet Measurement Conference (IMC)*, 2018.
[28] Secrank. https://secrank.cn/topdomain.
[29] SimilarWeb top websites ranking. https://www.similarweb.com/top-websites/.
[30] Top lists. https://toplists.github.io/.
[31] Tranco list. https://tranco-list.eu/.
[32] Trexa service. https://github.com/mozilla/trexa-service.
[33] Umbrella popularity list – top million domains. https://docs.umbrella.com/investigate-api/docs/top-million-domains.
[34] Q. Xie, S. Tang, X. Zheng, Q. Lin, B. Liu, H. Duan, and F. Li. Building an open, robust, and stable voting-based domain top list. In *USENIX Security*, 2022.
[35] D. Zeber, S. Bird, C. Oliveira, W. Rudametkin, I. Setall, F. Wollsén, and M. Lopatka. The representativeness of automated web crawls as a surrogate for human browsing. In *International Conference on World Wide Web*, 2020.

A ALL INTRA-CLOUDFLARE METRICS



(a) Jaccard Index

(b) Spearman Correlation

Figure 8: Intra-Cloudflare Popularity Metrics—We evaluate the full suite of 21 Cloudflare-derived popularity lists on a single day of data (February 1, 2022) as described in Section 3.2; from there we choose a set of 7 popularity definitions that capture a diverse set of lists, which we run on a full month of data and use for our evaluation of top lists. Our results on a month of data for the 7 lists (Figure 1) are consistent with our findings here.