# Analyzing the Web: Are Top Websites Lists a Good Choice for Research?

Tom Alby[0000−0002−6696−5185] and Robert Jäschke[0000−0003−3271−9653]

Humboldt-Universität zu Berlin, Germany
{thomas.alby,robert.jaeschke}@hu-berlin.de

**Abstract.** The web has been a subject of research since its beginning, but it is difficult if not impossible to analyze the whole web, even if a database of all URLs would be freely accessible. Hundreds of studies have used commercial top websites lists as a shortcut, in particular the Alexa One Million Top Sites list. However, apart from the fact that Amazon decided to terminate Alexa, we question the usefulness of such lists for research as they have several shortcomings. Our analysis shows that top sites lists miss frequently visited websites and offer only little value for language-specific research. We present a heuristic-driven alternative based on the Common Crawl host-level web graph while also taking language-specific requirements into account.

## 1 Introduction

The web provides an almost endless reservoir of data, but at the same time, it is virtually impossible to get a complete picture of the web due to its size and its continuous changes. As a consequence, researchers that want to leverage web resources for their research have to define how to create a valid data set to answer their questions. This can be easy if a study focuses on a data source itself such as a search engine ("What is the quality of health information in Google's search results?") but becomes more difficult if a more general question needs to be answered ("What are the most common trackers on the web?"). In absence of a database of all currently active hosts of the web, the Alexa Top Sites lists and similar offers have attracted researchers as these lists seem to provide an easy answer to a complex question [27, 12, 39]. In December 2021, Amazon announced that alexa.com would be retired in May 2022 [3].

However, several dimensions should be taken into account when a selection is made:

**Completeness:** Does the selection represent the web or the topic that is analyzed? If it is a sample, is it a random sample of all available hosts?
**Freshness:** In how far does the selection represent the web at the current state? Are new hosts included? Have dead hosts been removed?
**Language:** Is it possible to include only pages in a specific language?
**Locale:** Is it possible to retrieve pages that are relevant to a specific locale? A pizza delivery website in Austria is not relevant to a German user but a pizza recipe site from Austria is.

**Topic distinction:** Is it possible to restrict a selection to a topic?
**User-facing versus embedded resources:** Some sites embed resources from other websites, for example, Google Fonts, and these sites are not intended to be seen by human users. Are these resources required for an analysis or could they distort the results?
**Quality and spam prevention:** Some web pages are excluded from search engines and other repositories because they include spam.

Our goal is to help researchers understand the disadvantages of each approach with respect to the points above and to provide a heuristic-driven solution. Therefore, we review several data sources that can be grouped into three categories:

- search engines (e.g., Google and Bing),
- top sites lists (e.g., Alexa and Majestic),
- other repositories that include URLs and that are freely accessible to researchers (e.g., Twitter, Wikipedia, and Common Crawl).

We only review data sources that are still being updated, excluding historical data sets that are not refreshed anymore such as the dmoz database or del.icio.us data. For comparison reasons, one example of an Alexa list is included despite its shutdown. In addition to a general look at the data sources, we will also analyze the appropriateness of sources for the German locale as an example of a language-specific subset that shares only the language with other locales.

This paper is organized as follows: In Section 2, an overview of related work is provided. In Section 3, different data sources, acquisition approaches and processing methods are described. The results of the analysis of these data sets are provided in section 4. Finally, in Section 5, we propose our solution and close with a recommendation for further research.

## 2 Related Work

While no comparison between the aforementioned data sets exists to the best of our knowledge, each data source has been extensively reviewed for itself. Approaches using search results from both Google and Bing have been analyzed with respect to their limited coverage of different countries and languages [43, 46] and spam, that is, web pages of low quality [13]. The use of search engine results for linguistic research has been criticized due to the lack of transparency about linguistic processing and fluctuation of results [23]. Search engine freshness was critically viewed [25].

Internet top lists are widely used in research [4, 10, 30, 40] but their quality and stability is questionable and subject to manipulation [35, 36, 39] which also has an impact on the results of research [38]. In order to protect research from manipulated lists, top lists provided by Alexa, Cisco Umbrella, Majestic, and QuantCast have been combined in TRANCO, using either a Borda count or Dowdall's rule to let lists rank the inclusion of a site in an output list [36]. A

combination of an Alexa top list with topic categories combined with crawls of the Internet Archive, restricted to .de domains emphasized the pace of changes in the web, stating however that crawling strategies of different indices are biased and will thus result in biased results [20]. A bias in terms of language and country distribution has also been detected in the Internet Archive [5, 17, 44]. Finally, Lo and Sedhain reviewed different website rankings of which none is available anymore [28].

Wikipedia is regarded as a stepping stone between search engines and other websites [34]. Spamming approaches that try to add links of low quality to Wikipedia articles perform poorly, resulting in a rather clean repository of URLs [49]. However, the quality of information on web sites that is being linked to by Wikipedia is inferior to the information provided by official sites [24].

Finally, undesirable content has been found in the Common Crawl corpus and challenges the language models based on this data [29]. Another quality problem in Common Crawl data is caused by near-duplicates [14].

## 3 Experiments and Data

### 3.1 Experimental Setup

Data was collected in the first half of 2021. Each dataset has been processed before it was compared to the other sets. Some datasets contain complete *URLs* such as https://www.hu-berlin.de/index.html, other include *hosts* (www.ibi.hu-berlin.de), and Alexa only offers second-level *domains* (hu-berlin.de).[1] Our study focuses on hosts and not on second-level domains since different hosts on the same domain can have different content. As a first step, malformed URLs were removed from the datasets, host names were extracted from URLs and stripped off "www.", since some sources do not include it. The reverse domain name notation of the Common Crawl host web graph has been transformed to the notation of the other datasets. Finally, duplicates were removed from each list, and hosts were matched between the lists. The details of the acquisition and specifics for each dataset are described in the next sections. Research data is available for download [2].

### 3.2 Search

Search results have been analyzed in various papers, for example in order to analyze the information that users see for specific queries [1, 9, 19, 22, 24, 47]. Using queries allows for topic selection and the identification of relevant hosts to that topic, while a full list of all sites in Google's and Bing's indices is not available. Query selection thus plays an important role, involving either the

[1] The term *sites* will be used as a synonym for *hosts*. A *page* is regarded as a single web page document on a *host*.

creation of a keyword cluster around a topic including the search volumes,[2] or creating a sample of general queries. In addition, search engines offer language and locale selection, that is, search engines determine a user's locale and language preference and rank results accordingly.

For our study, we use two different data sets:

- A random sample of 1,000 English search queries of the Million Query Track Overview [8] that accounts for 101,001,690 average monthly search volume on the US version of Google. The most popular query in the data set, *google maps*, has an average monthly search volume of 24,900,000 alone.
- Three sets of German language queries that each cover a topic of different popularity level according to the Google Keyword Planner:
  - Federal Election ("Bundestagswahl") as this was a "hot" topic in Germany in summer 2021 with 614 queries that account for 939,730 average monthly search volume
  - iMac M1 (same in German) with 90 queries and an average monthly search volume of 12,180 as these new iMacs were announced in April 2021
  - Minimalism ("Minimalismus" in German, can be interpreted as an art direction, a music style or a lifestyle approach) with 86 queries and an average monthly search volume of 34,510.

Queries were sent to both Bing and Google, the first data set to the en-US versions of the search engines, the second to the de-DE versions. Only the first results page was considered as less than 1% of clicks go on a result on the second results page [6].

### 3.3 Top Sites

Top lists have been used in many studies [18, 20, 21, 45], in particular the Alexa Top Sites list. Different methods are used by providers to generate each list. Alexa ranks sites based on the number of visits measured by a browser-based panel. While Cisco Umbrella also leverages user behaviour, this list is based on the number of DNS requests for hosts, including those that receive requests from other device types.[3] Looking at Table 1, the Cisco top 10 includes hosts such as Netflix delivery nodes that are typically not visited by user browsers. The main advantage of monitoring user behavior, however, is the freshness of the data.

Majestic provides a list of top sites based on a link graph that is based on links discovered in Majestic's own crawls,[4] that is, it is not created based on user behavior but on the number of links of pages pointing to other pages. Common Crawl (see Section 3.4) follows a similar approach in their web graphs. The

[2] Relevant queries and their search volume can be identified using tools such as the Google Keyword Planner that provides historical data about search volume of specific queries.

[3] https://s3-us-west-1.amazonaws.com/umbrella-static/index.html

[4] https://majestic.com/reports/majestic-million

**Table 1.** Top 10 Sites across different lists.

| Alexa | Cisco Umbrella | Majestic | Common Crawl |
|---|---|---|---|
| google.com | google.com | google.com | facebook.com |
| youtube.com | netflix.com | facebook.com | fonts.googleapis.com |
| baidu.com | microsoft.com | youtube.com | twitter.com |
| facebook.com | www.google.com | twitter.com | google.com |
| instagram.com | ftl.netflix.com | instagram.com | youtube.com |
| bilibili.com | prod.ftl.netflix.com | linkedin.com | s.w.org |
| yahoo.com | api-global.netflix.com | microsoft.com | instagram.com |
| qq.com | data.microsoft.com | apple.com | goo...tagmanager.com |
| wikipedia.org | ichnaea.netflix.com | wikipedia.org | linkedin.com |
| amazon.com | eve...data.microsoft.com | goo...tagmanager.com | ajax.googleapis.com |

assumption behind link-based approaches is that sites with a high link popularity are also more popular for users. Since web sites also embed resources from other hosts that count as a link, the top sites include services such as the Google Tag Manager (see Table 1). The link popularity approach has several disadvantages as it is prone to spam, links in the graph may point to dead sites, and not every active site has an incoming link, especially new sites, so that completeness could be a problem as well.

TRANCO offered a combination of Alexa, Quantcast (based on traffic measured by a toolbar and an internet service provider), Cisco, and Majestic data but two of their initial data sources, Alexa and Quantcast, are not available anymore today. We use the Tranco list generated on 31 July 2021, including a combination of ranks provided by Alexa, Umbrella, and Majestic from May 1st, 2021 to May 31st, 2021.[5]

These different ways to compile a *top* list lead to the question why most lists include exactly *one million* sites. For Alexa and Cisco, no information about the frequency or length of user visits to the hosts or any other key performance indicator is included. In other words, we do not know if a top one million sites list represents 80%, 50%, or only 2% of overall web traffic, time spent on these sites, or numbers of visits. In Section 4.2, we will approach popularity in terms of whether a host is found for a Google search and, in Section 4.3 with respect to the search volume.

Cisco, Majestic, and TRANCO offer a selection based on TLDs. In Section 4.4, we will review in how far TLDs are suitable for representing a locale or a language. None of the data sources above provides a topic detection.

### 3.4 Common Crawl

The Common Crawl project's data has been widely used for research in different areas [15, 16, 37, 41, 48, 42]. Seed URLs for Common Crawl are collected from

[5] Available at https://tranco-list.eu/list/3X4L.

outlinks or sitemaps [33]. Crawls usually take place on a monthly basis, updating known pages but also crawling new pages. The focus is set on broadness rather than depth of hosts, that is, Common Crawl tries to get a broad sample of hosts and more pages from higher ranking domains.

Common Crawl's June 2021 crawl includes 2.45 billion web pages, the host web graph of the crawls of February/March, April and May 2021 had 514,570,180 nodes.[6] The graph data includes a column with a ranked position based on harmonic centrality [7] and another one based on PageRank. Harmonic centrality takes the distance of a node into account whereas PageRank considers the importance of the neighbourhood of a node. We have used harmonic centrality due to it being less influenced by spam [31].

Common Crawl also offers domain, host, and URL information with detected languages. Language is identified and annotated by Common Crawl using the Common Language Detector 2.[7] Up to three languages can be associated with a page, cascading up to the host. We have downloaded all known hosts that were detected to include German content since the start of the Common Crawl project (10,057,081 hosts), using the Amazon Web Services Athena interface [32].

### 3.5 Wikipedia

Wikipedia includes one of the last manually collected and reviewed URL repositories after the closures of Yahoo's directory in 2010 and the Open Directory in 2017. For this study, data dumps of external links with 3,707,420 unique hosts of the English and 1,159,179 unique hosts of the German version of Wikipedia, published in July 2021, have been analyzed.[8]

The dump of external links does not contain any topics for each URL. These could be extracted from the Wikipedia page that includes such links, requiring additional processing steps. Not all of the links included in the German external links dump refer to German-language content hosts, that is, we cannot conclude that this is a German-language dataset or specific to a locale. Wikipedia has its own link rot detection method to ensure freshness. Resources embedded from other hosts are not included in the Wikipedia links.

### 3.6 Twitter

Twitter offers a sampled stream that covers approximately 1% of all publicly available tweets [11]. A collection of this data with tweets from February 1st, 2021 to July 31st, 2021, has been analyzed for this study, resulting in 21,031,785 discovered unique hosts in tweets. A subset of 3,505,629 tweets in German language as detected by Twitter's own language detection engine resulted in 25,881

[6] Data available at https://commoncrawl.org/2021/05/host-and-domain-level-web-graphs-feb-apr-may-2021/.

[7] https://github.com/CLD2Owners/cld2

[8] Data dumps of Wikipedia External links are available at https://dumps.wikimedia.org/backup-index.html.

unique hosts. However, while the text of a tweet is detected to be in German language, this does not mean that the URL in the same tweet will link to a German language website or that it is specific to the German locale.

Twitter does not offer topics. However, hashtags that annotate tweets could act as a proxy for topic identification. For the German language tweets subset, hashtags were extracted to search for hosts containing the topics that were used for the search data set. In order to identify URLs that could be relevant for the same topics but that are not present in search, we first looked at the URL overlap for each topic between tweets and search engines and then extracted the hashtags for those URLs present in both datasets. Of these tweets, we also extracted their other hashtags in order to broaden the recall.

## 4 Results

### 4.1 Completeness: Overlap of Data Sets

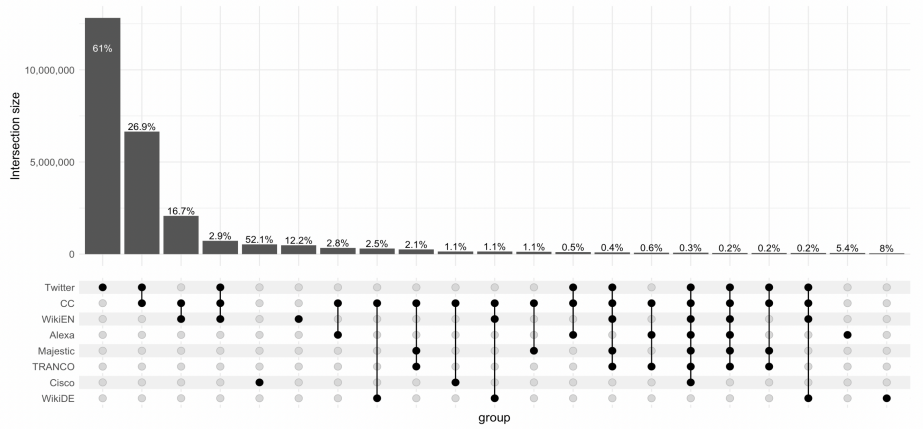


**Fig. 1.** UpSet diagram displaying the overlap of the data sources. Each bar shows the overlap or uniqueness of one or more data sources with a percentage of hosts unknown to the other hosts. The diagram includes only hosts from Common Crawl (CC) that are present in at least one other data source. The graph is built based on a minimum intersection size of 50,000 hosts.

Figure 1 shows the small overlap between the different data sets in an UpSet diagram [26]; the low number of overlapping unique hosts has also been described by Pochat et al. [36]. Each bar shows the number of overlapping hosts in the data sources, displayed as dots in the matrix below and a percentage of the hosts exclusive to this intersection relative to its potential size. The diagram does not include all hosts of the Common Crawl graph due to 94.9% of hosts in the Common Crawl graph being unknown to the other data sources. Instead,

only those Common Crawl hosts were included that were also present in at least one other data source, resulting in 26,059,931 unique hosts.

Twitter surprisingly has 61% hosts that are unknown to the other data sources while it still has the biggest overlap with the Common Crawl Web graph. However, more than 71% of these hosts found in Twitter occur only once in the data set and their quality needs further review. Without a filter in terms of traffic, link popularity, or visits and no review by other humans as in Wikipedia, Twitter may include hosts that cannot pass the filters of the other data sources. Similarly, the majority of sites in the Cisco list is unknown to the other data sets (52.1%), which may be the result of the inclusion of hosts that are embedded by other services.

Not only is the overlap of the top lists low, they also hardly correlate in terms of ranking as pointed out by Pochat et al. [36]. To illustrate this point, for Alexa and Majestic, less than 50% of their hosts appear within the first million hosts of the Common Crawl host-level web graph; about 95% are found within the first 50 million.

### 4.2 Search Results versus other Data Sets

Given that both Common Crawl and Twitter have large amounts of hosts unknown to the other lists, the question is whether these additional hosts are of minor quality and thus excluded for good reasons by the top lists. As the overlap between the other top lists is small, search results will be leveraged as an additional signal: If a host is regarded as good enough to be included by search engines, we assume that its quality does not indicate its exclusion.

Comparing the two search engines against each other, Bing and Google have surprisingly different result with respect to the hosts included in the ranking as detailed in Figure 2; however, this may be a result of looking at the first results page only. For both locales, Bing includes results from a larger variety of hosts in search results compared to Google.

**Table 2.** Coverage of Search Result Hosts in Data Sources.

| Search | CC Graph | WikiEN | WikiDE | TRANCO | Alexa | Majestic | Twitter |
|---|---|---|---|---|---|---|---|
| Bing DE | 96.92% | 48.92% | 63.39% | 48.00% | 39.45% | 51.54% | 31.58% |
| Bing US | 95.57% | 63.84% | 35.05% | 61.46% | 53.15% | 56.91% | 31.87% |
| Google DE | 98.95% | 57.12% | 71.36% | 57.57% | 46.18% | 62.07% | 23.92% |
| Google US | 98.45% | 70.40% | 39.67% | 66.49% | 57.83% | 63.17% | 24.85% |

While the link-based approach was expected to perform worse in terms of completeness (see Sec. 3.3), the Common Crawl host web graph is by far the most comprehensive data set with respect to the hosts discovered in the search engines as detailed in Table 2. All other data sources perform far worse, also Majestic that also relies on a link-based approach. TRANCO is stronger than Majestic for

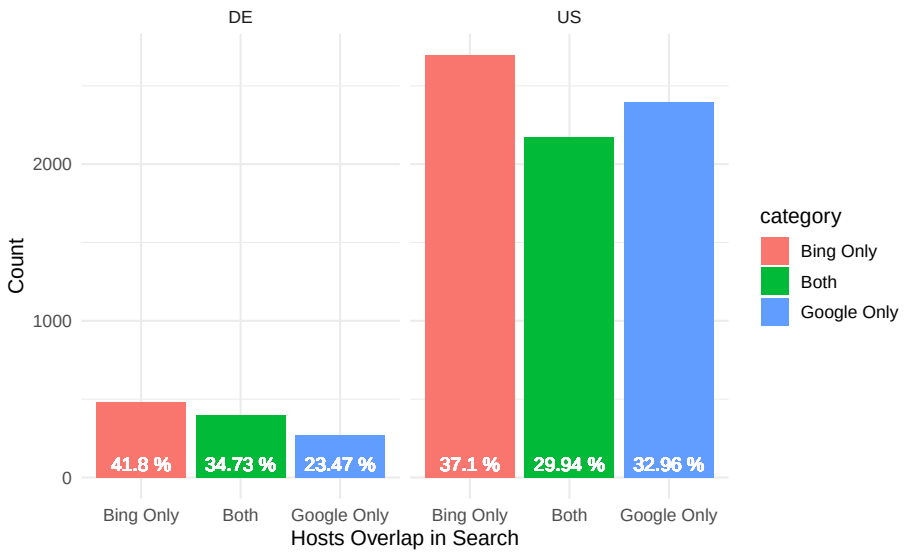


**Fig. 2.** Overlap of unique hosts in search engine results, divided by locale.

the hosts of the US search engines, indicating that the other data sets combined in TRANCO push German language hosts out of the list as it performs worse for this subset. Given that only the first results page of the search engines was used for the host lists and that search engines leverage web graphs as well, it is assumable that the restriction to 1 million sites in the Majestic and TRANCO lists is responsible for the mediocre coverage.

Twitter has the lowest coverage for the hosts found in search, and for a focus on a topic, URLs in tweets have not added considerable value. For the minimalism and the iMac M1 topics, no tweets in German language and thus no hashtags were found at all. While tweets about these topics do exist, they seem to be so rare in the German Twitter community that they probably did not occur in the 1% sample. For the German Federal Election topic, about 100 hosts were identified that were not in the first page search results hosts, mostly regional news websites but also NGOs and alternative media. All of them were found in the Common Crawl web graph.

Hosts in search results are distributed over the Common Crawl host-level web graph, shown in Figure 3. Web graph position data has been logarithmized in order to accommodate for the large scale. Slightly more than 50% of the hosts found in the US versions of Google and Bing come from the first one million hosts in the web graph (first vertical line on the left), more than 80% come from the first 10 million hosts (second vertical line, but even with 50 million hosts (third vertical line), we have not covered 100% of the search results. Google seems to prefer hosts with a higher link popularity with around 60% of hosts

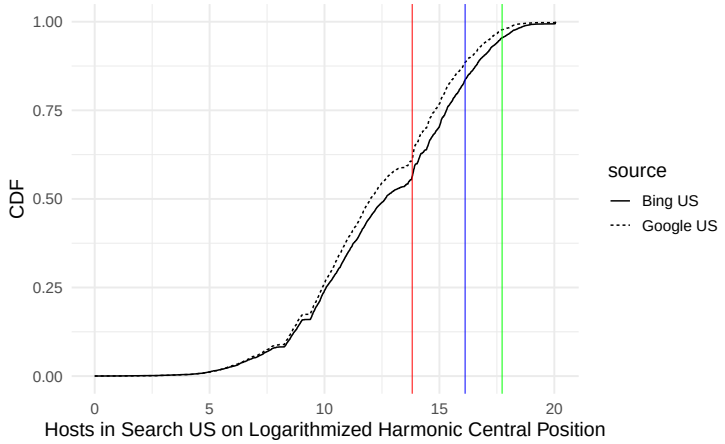


**Fig. 3.** Distribution of hosts found in US search engine results in the logarithmized harmonic central position displayed in a cumulative density function (CDF) plot, vertical lines at 1 , 10, and 50 million computed to the logarithmized number.

coming from the first million hosts, compared to Bing where around 55% of the hosts are in the same segment.

The German language query set with three different topics shows even more differentiated results. A popular local topic (in terms of search volume) such as the German Federal Election brings up hosts that are found in the first 50% of the Common Crawl web graph, whereas the Apple iMac M1 topic benefits from globally linked sites such as apple.com that also serves German language pages. For a less popular topic such as minimalism, both search engines go beyond the 1st million as only 25%, respectively 30%, of the results were found in the top sites of the graph. While the Common Crawl host-level web graph is the most complete data source, are hosts not included in the top lists unpopular?

### 4.3 Is a List of 1 Million Popular Sites enough?

A standard definition of *popular* and *less popular* does not exist, and different top sites lists have different ways to compute their flavour of popularity that is not completely transparent as discussed in section 3.3. Search, however, offers a transparent currency for popularity, the number of searches for a query. If a host appears in a search engine's results for a popular search query or appears more frequently, it is more likely to be seen by a user. In Figure 4, each host is displayed as a point with reference to the highest search volume query it is found for and its rank in the Common Crawl Web Graph, differentiated according to whether it is in the TRANCO list or not.[9] The darker a point is, the more often this host appears in search results.

[9] Only search volume data from Google has been taken into account, and, as a consequence, only the hosts found in Google Search.

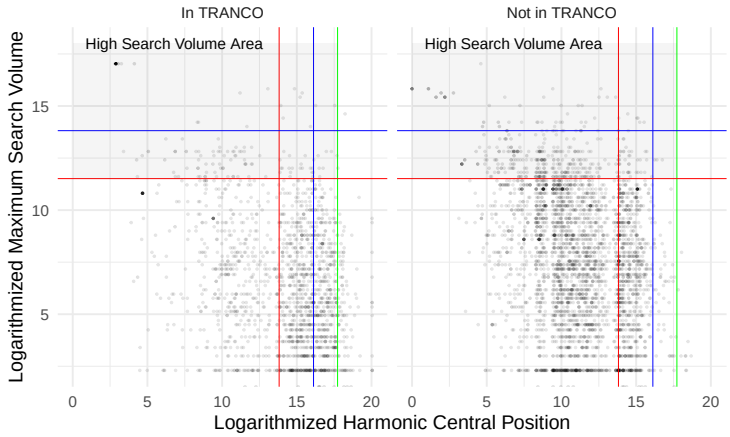


**Fig. 4.** Highest search volume for which a host is found versus its rank in the harmonic central web graph, differentiated by (non)-occurrence in TRANCO. Vertical lines at 1, 10, and 50 million, horizontal lines at 100 000 and 1 million monthly search volume.

The grey-shaded area includes high-volume search queries, beginning with 100 000 monthly search queries at the lower vertical line and the upper line mirroring a threshold of 1 million monthly search queries. Not only does the TRANCO list contain less hosts that are found for high-volume search queries, it also includes less hosts with a more frequent occurrence in search results. At the same time, most TRANCO hosts found in search come from the lower positions of the Common Crawl Web Graph, even from the area of position 10 million and above (second vertical line), missing more frequent hosts that are in the upper area of the graph.

### 4.4 TLDs are neither a Good Proxy for a Language nor for a Locale

Using the results of the German search query dataset, we look at the distribution of TLDs of those hosts that contain German-language content. Search results differ in TLD distribution, depending on the topic as detailed in Table 3.

**Table 3.** Distribution of the top 5 TLDs in German-language search results.

| Google | Election | Minimalism | iMac M1 |
|---|---|---|---|
| .de | 83.6% | 75.6% | 68.8% |
| .com | 8.0% | 13.6% | 27.3% |
| .ch | 1.5% | 6.0% | 1.4% |
| .net | 0.7% | 1.5% | 0.1% |
| .org | 5.4% | 1.2% | 0.1% |

| Bing | Election | Minimalism | iMac M1 |
|---|---|---|---|
| .de | 85.6% | 71.8% | 63.5% |
| .com | 5.1% | 14.0% | 34.3% |
| .ch | 0.0% | 6.1% | 0.0% |
| .net | 1.2% | 2.0% | 0.3% |
| .blog | 0.0% | 1.2% | 0.0% |

The German Federal Election content is served from more .de domains than other content such as minimalism, the lowest number of .de domains is found for the iMac M1 topic. Apple, for example, serves German content from its apple.com domain and has unsurprisingly been dominant for the iMac search results. The TLD distribution for the Common Crawl German-language dataset includes an even lower share of .de domains, but these results may have to be taken with a grain of salt as Common Crawl does not only include German external links. As a further step, we analyzed how many hosts we would lose if we used the TLD selection provided in TRANCO: 151 out of 667 hosts (22.6%) of the Google results were not included, apart from the fact that Alexa and the other providers did not have all hosts anyway. Using the language detection of Common Crawl, only 62 out of 667 hosts (9.3%) were not included.

## 5 Discussion and Conclusion

Our findings challenge the value of research that uses top sites lists or TLDs for a focus on a country. Apart from the fact that there is no generally accepted definition of popularity and different sources have different ways of defining it, our results show that by cutting at 1 million, hosts are missing that are displayed in Google and Bing for popular queries and also hosts that more frequently occur in search results. As a consequence, a sampling strategy using the Common Crawl host web graph will provide more robust results for research than the use of top sites lists. The Common Crawl host-level web graph contains 95% all hosts that were present in other data sources within the top 50 million threshold. We provide sample code and a test tool to leverage this massive data set.[10]

Instead of using TLDs for language/locale restriction, language detection should be used as the TLD approach prevents local sites from being included. The Common Crawl language identification approach is far from perfect, and it does not solve the problem of identifying relevant sites for a locale, but we see more than twice as many better results using the Common Crawl language detection compared to using TLDs.

The approach described in this paper has several drawbacks that require further research. The amount of spam and dead hosts in the host-level web graph has not been analyzed on a larger scale yet, and web graphs based on link popularity are not a perfect mirror of usage. Depending on the intended use, the inclusion of technical hosts and spam has to be configurable as some studies will be interested in these hosts. Topic identification is unsolved for most data sources. Finally, locale affiliation may be identifiable by using link networks of sites within a locale.

[10] https://alby.link/ccsample

## References


1. Alby, A., Bauknecht, H., Weidinger, S., Mempel, M., Alby, T.: Muster und Limitationen der Internet-basierten Selbstdiagnose bei häufigen Dermatosen. In: JDDG: Journal der Deutschen Dermatologischen Gesellschaft. vol. 19 (2021)
2. Alby, T.: Analyzing the web: Are top websites lists a good choice for research? (0.1) [data set] (2022), https://doi.org/10.5281/zenodo.6821240
3. Alexa Internet, I.: We will be retiring alexa.com on may 1, 2022 (2021), https://support.alexa.com/hc/en-us/articles/4410503838999
4. Allen, G., Downs, B., Shukla, A., Kennington, C., Fails, J.A., Wright, K.L., Pera, M.S.: BiGBERT: Classifying educational web resources for kindergarten-12th grades. In: Advances in Information Retrieval. pp. 176–184. Springer International Publishing, Cham (2021)
5. AlSum, A., Weigle, M., Nelson, M., Sompel, H.: Profiling web archive coverage for top-level domain and content language. International Journal on Digital Libraries **14** (09 2013). https://doi.org/10.1007/s00799-014-0118-y
6. Backlinko: We analyzed 5 million Google search results (2021), https://backlinko.com/google-ctr-stats
7. Boldi, P., Vigna, S.: Axioms for centrality. CoRR **abs/1308.2140** (2013), http://arxiv.org/abs/1308.2140
8. Carterette, B., Pavluy, V., Fang, H., Kanoulas, E.: Million query track 2009 overview. In: TREC (01 2009)
9. Craigie, M., Loader, B., Burrows, R., Muncer, S.: Reliability of health information on the internet: An examination of expert's ratings. Journal of Medical Internet Research **4** (2002), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1761929/
10. Englehardt, S., Narayanan, A.: Online tracking: A 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 1388–1401. CCS '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2976749.2978313
11. Fafalios, P., Iosifidis, V., Ntoutsi, E., Dietze, S.: Tweetskb: A public and large-scale rdf corpus of annotated tweets. In: European Semantic Web Conference. pp. 177–190. Springer (2018)
12. Felt, A.P., Barnes, R., King, A., Palmer, C., Bentzel, C., Tabriz, P.: Measuring HTTPS adoption on the web. In: 26th USENIX Security Symposium (USENIX Security 17). pp. 1323–1338. USENIX Association, Vancouver, BC (Aug 2017), https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt
13. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In: Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004. pp. 1–6. WebDB '04, Association for Computing Machinery, New York, NY, USA (2004). https://doi.org/10.1145/1017074.1017077
14. Fröbe, M., Bevendorff, J., Gienapp, L., Völske, M., Stein, B., Potthast, M., Hagen, M.: CopyCat: Near-Duplicates Within and Between the ClueWeb and the Common Crawl, pp. 2398–2404. Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3404835.3463246
15. Funel, A.: Analysis of the web graph aggregated by host and pay-level domain. CoRR **abs/1802.05435** (2018), http://arxiv.org/abs/1802.05435

16. Giannakoulopoulos, A., Pergantis, M., Konstantinou, N., Lamprogeorgos, A., Limniati, L., Varlamis, I.: Exploring the dominance of the english language on the websites of eu countries. Future Internet **12**(4) (2020). https://doi.org/10.3390/fi12040076, https://www.mdpi.com/1999-5903/12/4/76
17. Hale, S.A., Blank, G., Alexander, V.D.: Live versus archive: Comparing a web archive to a population of web pages, pp. 45–61. UCL Press (2017), http://www.jstor.org/stable/j.ctt1mtz55k.8
18. He, K., Fisher, A., Wang, L., Gember, A., Akella, A., Ristenpart, T.: Next stop, the cloud: Understanding modern web service deployment in ec2 and azure. In: Proceedings of the 2013 Conference on Internet Measurement Conference. pp. 177–190. IMC '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2504730.2504740
19. Höchstötter, N., Lewandowski, D.: What users see – structures in search engine results pages. Information Sciences **179**(12), 1796–1812 (2009). https://doi.org/https://doi.org/10.1016/j.ins.2009.01.028, special Section: Web Search
20. Holzmann, H., Nejdl, W., Anand, A.: The dawn of today's popular domains: A study of the archived german web over 18 years. CoRR **abs/1702.01151** (2017), http://arxiv.org/abs/1702.01151
21. Iqbal, U., Shafiq, Z., Qian, Z.: The ad wars: Retrospective measurement and analysis of anti-adblock filter lists. In: Proceedings of the 2017 Internet Measurement Conference. pp. 171–183. IMC '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3131365.3131387
22. Kakos, A., Lovejoy, D., Whiteside, J.: Quality of information on pelvic organ prolapse on the internet. International urogynecology journal **26** (10 2014). https://doi.org/10.1007/s00192-014-2538-z
23. Kilgarriff, A.: Googleology is bad science. Comput. Linguist. **33**(1), 147–151 (mar 2007). https://doi.org/10.1162/coli.2007.33.1.147
24. Leithner, A., Maurer-Ertl, W., Glehr, M., Friesenbichler, J., Leithner, K., Windhager, R.: Wikipedia and osteosarcoma: a trustworthy patients' information? Journal of the American Medical Informatics Association **17**(4), 373–374 (07 2010). https://doi.org/10.1136/jamia.2010.004507
25. Lewandowski, D.: A three-year study on the freshness of web search engine databases. Journal of Information Science **34**(6), 817–831 (2008). https://doi.org/10.1177/0165551508089396, https://doi.org/10.1177/0165551508089396
26. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H.: Upset: Visualization of intersecting sets. IEEE Transactions on Visualization and Computer Graphics (InfoVis) **20**(12), 1983–1992 (2014). https://doi.org/10.1109/TVCG.2014.2346248
27. Libert, T.: Exposing the hidden web: An analysis of third-party HTTP requests on 1 million websites. CoRR **abs/1511.00619** (2015), http://arxiv.org/abs/1511.00619
28. Lo, B., Sedhain, R.: How reliable are website rankings? implications for e-business advertising and internet search. Issues in Information Systems **7**, 233–238 (01 2006)
29. Luccioni, A.S., Viviano, J.D.: What's in the box? an analysis of undesirable content in the common crawl corpus. CoRR **abs/2105.02732** (2021), https://arxiv.org/abs/2105.02732
30. Mason, A.M., Compton, J., Bhati, S.: Disabilities and the digital divide: Assessing web accessibility, readability, and mobility of popular

health websites. Journal of Health Communication **26**(10), 667–674 (2021). https://doi.org/10.1080/10810730.2021.1987591, pMID: 34657585
31. Nagel, S.: Common crawl's first in-house web graph (May 2017), https://commoncrawl.org/2017/05/hostgraph-2017-feb-mar-apr-crawls/
32. Nagel, S.: Index to warc files and urls in columnar format (March 2018), https://commoncrawl.org/2018/03/index-to-warc-files-and-urls-in-columnar-format/
33. Nagel, S.: August 2019 crawl archive now available (2019), https://commoncrawl.org/2019/08/august-2019-crawl-archive-now-available/
34. Piccardi, T., Redi, M., Colavizza, G., West, R.: On the value of wikipedia as a gateway to the web. In: Proceedings of the Web Conference 2021. pp. 249–260. WWW '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442381.3450136
35. Pochat, V.L., van Goethem, T., Joosen, W.: Rigging research results by manipulating top websites rankings. CoRR **abs/1806.01156** (2018), http://arxiv.org/abs/1806.01156
36. Pochat, V.L., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., Joosen, W.: Tranco: A research-oriented top sites ranking hardened against manipulation. arXiv preprint arXiv:1806.01156 (2018)
37. Robertson, F., Lagus, J., Kajava, K.: A COVID-19 news coverage mood map of Europe. In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. pp. 110–115. Association for Computational Linguistics, Online (Apr 2021), https://aclanthology.org/2021.hackashop-1.15
38. Rweyemamu, W., Lauinger, T., Wilson, C., Robertson, W., Kirda, E.: Clustering and the weekend effect: Recommendations for the use of top domain lists in security research. In: Choffnes, D., Barcellos, M. (eds.) Passive and Active Measurement. pp. 161–177. Springer International Publishing, Cham (2019)
39. 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: Proceedings of the Internet Measurement Conference 2018. pp. 478–493. IMC '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278532.3278574
40. Silva, C.E., Campos, J.C.: Characterizing the control logic of web applications' user interfaces. In: Murgante, B., Misra, S., Rocha, A.M.A.C., Torre, C., Rocha, J.G., Falcão, M.I., Taniar, D., Apduhan, B.O., Gervasi, O. (eds.) Computational Science and Its Applications – ICCSA 2014. pp. 263–276. Springer International Publishing, Cham (2014)
41. Srinath, M., Wilson, S., Giles, C.L.: Privacy at scale: Introducing the privaseer corpus of web privacy policies. CoRR **abs/2004.11131** (2020), https://arxiv.org/abs/2004.11131
42. Tahir, B., Mehmood, M.A.: Corpulyzer: A novel framework for building low resource language corpora. IEEE Access **9**, 8546–8563 (2021). https://doi.org/10.1109/ACCESS.2021.3049793
43. Thelwall, M.: Web impact factors and search engine coverage. Journal of Documentation (03 2000). https://doi.org/10.1108/00220410010803801
44. Thelwall, M.: A fair history of the web? examining country balance in the internet archive. Library and Information Science Research **26**, 162–176 (05 2004). https://doi.org/10.1016/S0740-8188(04)00024-6
45. Varvello, M., Schomp, K., Naylor, D., Blackburn, J., Finamore, A., Papagiannaki, K.: Is the web http/2 yet? In: Karagiannis, T., Dimitropoulos, X. (eds.) Passive

and Active Measurement. pp. 218–232. Springer International Publishing, Cham (2016)
46. Vaughan, L., Thelwall, M.: Search engine coverage bias: evidence and possible causes. Information Processing and Management **40**(4), 693–707 (2004). https://doi.org/https://doi.org/10.1016/S0306-4573(03)00063-3
47. Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y., Xu, D.: Using internet search engines to obtain medical information: A comparative study. J Med Internet Res **14**(3), e74 (May 2012). https://doi.org/10.2196/jmir.1943
48. Wenzek, G., Lachaux, M., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: Ccnet: Extracting high quality monolingual datasets from web crawl data. CoRR **abs/1911.00359** (2019), http://arxiv.org/abs/1911.00359
49. West, A.G., Chang, J., Venkatasubramanian, K., Sokolsky, O., Lee, I.: Link spamming wikipedia for profit. In: Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. pp. 152–161. CEAS '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2030376.2030394