# Analysis of Wikipedia External Links

Onur Cakmak-Simic, Patrick Sattler, Johannes Zirngibl*
*Chair of Network Architectures and Services, Department of Informatics
Technical University of Munich, Germany
Email: onur.cakmak-simic@tum.de, sattler@net.in.tum.de, zirngibl@net.in.tum.de

***Abstract*—Network scans are an inherent element of research within the field of Computer Networks. The basis for these scans is a list of targets, commonly referred to as hitlist. There are readily available hitlists and active research on how such lists can be generated.**

**In this paper, we extract domain names from external link datasets provided by the Wikimedia Foundation and use them as the source for generating a hitlist. We assess the general structure of the extracted domains and compare them to the Alexa Top 1M. We find that our list has no apparent structural disadvantages. We also analyze the targets for potential biases regarding their distribution over ASes, prefixes, and IP addresses. Our results show that 52% of the gathered IPv6 addresses are within 30 prefixes of AS13335-CLOUDFLARENET and that the top 10 most occurring ASes contain 45% of all IPv4 targets. We find that 33% of the IPv4 and 42% IPv6 addresses map to more than one domain. Around 5.8% of our domains resolve to the same four IPv4 addresses belonging to AS53831-SQUARESPACE and 3.3% of domains to four IPv6 addresses in AS15169-GOOGLE.**

***Index Terms*—Internet measurement, Internet hitlists**

## 1. Introduction

Network scans and their resulting measurements are important to many stakeholders in a network, from individual clients measuring their provided service, ISPs trying to optimize their operational costs to researchers measuring network characteristics, evaluating their findings, or deploying algorithms on a larger scale. IPv4 scanning and the generation of hitlists date back to the 90s [1,2]. Nowadays, tools like ZMap [3] and MASSCAN [4] enable scanning the entire IPv4 address space in feasible time. Although possible, a full scan might not be suitable. Not all types of scans scale well to that size [5]. We might need domains names, e.g., for TLS scans which generally require domain names due to Server Name Indication, we might have limited infrastructure or have a narrow target group. With IPv6, complete scans of the address space are not feasible [6], so hitlists are a necessity.

Depending on the source, the list of targets might be biased. Detecting and eliminating these biases is not trivial and a field of active research [7,8]. To ensure some form of quality for the list of targets, Gasser et al. [7] suggest gathering addresses that belong to individual hosts and have an even distribution across ASes and prefixes. At the very least, the potential for biases should be considered when conducting research, as a nonrepresentative or skewed list might lead to wrong conclusions.

In this paper, we analyze a potential source for generating a hitlist. For that we extract domain names from external links found in Wikipedia articles. According to Wikipedia community guidelines [9], each article may include an external link section listing the web presence of entities relevant to the article. External links refer to links from articles to web pages outside of Wikipedia.

**Outline** Section 2 briefly presents related work and terminology used throughout this paper. Our methodology for extracting the domains and resolving them to IP addresses is outlined in Section 3. Section 4 covers some structural properties of the extracted domain names. We inspect the list of targets for potential biases towards ASes, prefixes, and IP addresses in Section 5. Finally, Section 6 concludes our paper and suggests possible future work.

## 2. Related Work and Background

There are many sources from which to generate hitlists, including passive [6] and active measurements [10], Certificate Transparency logs [11], and machine learning [12]. As a result of continuous research, a considerable amount of datasets, providing sources or targets, have been accumulated. While some of these datasets are restricted and proprietary [6,13], many are publicly available [14,15]. Frequently used sources are top lists, e.g., the Alexa Top 1 Million list [16] that rank web domains by popularity. Scheitle et al. [17] and Pochat et al. [18] found that some of these lists exhibit characteristics that need to be accounted for prior to their use in research. These include, but are not limited to, significant and frequent churn, a nontransparent ranking mechanism, and a weekend and clustering effect [5]. Attempts to address some of these issues include using prefix top lists [19] or incorporating multiple such lists [20].

To the best of our knowledge, Paul Hoffman's [21] work is the first to generate a hitlist using external links from Wikipedia articles and to evaluate specific network characteristics of the targets.

**Background** For the structural analysis in Section 4, we use the notions of base domain and subdomain depth to obtain insights into the depth and breadth of our domains [17]. By base domain, we refer to the public suffix and the first domain prefixing it, e.g., `google.com`. Each subdomain preceding the base domain adds a value of 1 to the subdomain depth, e.g., `www.support.google.com` has subdomain depth 2. For clarity, in this paper, the term bias


doi: 10.2313/NET-2022-01-1_08

TABLE 1: List structures. The $SD_x$ columns indicate the share of domains with subdomain depth $x$. $SD_0$ represents domains with no subdomain, i.e., a base domain [17]. In the third column, 1498 TLDs correspond to 100%. The given numbers are rounded down to the nearest tenth decimal.

| List | Size | TLDs | $SD_0$ | $SD_1$ | $SD_2$ | $SD_{>3}$ | ∩Alexa |
|---|---|---|---|---|---|---|---|
| Joint | 3.5M | 57.8% | 22.2% | 70.3% | 6.2% | 0.8% | 21.0% |
| de | 1.2M | 44.1% | 17.7% | 76.2% | 5.7% | 0.5% | 8.5% |
| en | 3.3M | 57.6% | 23.9% | 68.7% | 6.5% | 0.7% | 20.9% |
| fr | 981.8K | 41.7% | 17.9% | 74.2% | 6.8% | 0.8% | 8.5% |
| ceb | 6.2K | 9.9% | 17.6% | 76.9% | 5.1% | 0.2% | 0.3% |
| sv | 243.7K | 28.1% | 16.5% | 75.5% | 6.3% | 1.5% | 3.4% |
| nl | 317.3K | 31.2% | 14.0% | 78.9% | 5.3% | 1.6% | 3.6% |
| Alexa | 690.9K | 52.2% | 85.4% | 14.4% | 0.1% | 0.0% | 100.0% |

of a hitlist refers to its propensity to certain subsets in the IP address space.

## 3. Methodology

The Wikimedia Foundation, the parent company of Wikipedia, provides a wide range of data dumps. Among them are SQL dumps that provide information about the external links across all articles within a given Wikipedia language edition. As of June 2021, there are 321 Wikipedia editions. For this paper, we used the six largest editions, based on the number of articles. These are the English, Cebuano, Swedish, German, French, and Dutch Wikipedias. To create the lists of domain names we performed the following steps:

- We pulled the external link SQL dump for each language edition on May 04, 2021.
- We extracted the individual URLs from the dumps and pruned those that were nonvalid URLs, had bad syntax, used nonstandard ports, or contained irrelevant protocols.
- We removed any unwanted prefix and suffix leaving the base domain and subdomains.
- We deleted duplicate entries.

In addition, we created a Joint list by merging the individual lists, again removing duplicate entries. To resolve the domains and collect IP addresses, we used MassDNS [22] with an Unbound [23] resolver. We performed the scan on May 17, 2021.

## 4. Structure

We check how many unique Top Level Domains (TLDs) are used and the subdomain depth across the domains in each list. In addition, we compute the intersection between our lists and the Alexa Top 1M, which we retrieved May 26, 2021.

### 4.1. TLD Coverage

As of May 2021, IANA [24] reports the existence of 1498 valid TLDs. Table 1 shows the results for all lists. The Joint and English list with 57.8% and 57.6% cover almost the same number of TLDs, approximately 865. There is a noticeable relation between the size of a list and the amount of TLDs it contains. An exception is the Alexa Top list which at almost half the size of the German list includes 120 TLDs more. This might be attributed to the larger share of base domains in the Alexa list leading to a wider range of targets. The smallest list, Cebuano, misses over 90% of TLDs. This is a consequence of its small number of entries, although it is the second largest Wikipedia edition. It is the smallest list because of an unexpectedly large number of duplicate entries in the SQL dump, which we removed during the list's creation.

TABLE 2: Top 5 TLDs by occurrence. Values are percentages of the number of domains in the respective list.

| | Lists | | | | | |
|---|---|---|---|---|---|---|
| TLD | de | en | fr | ceb | sv | nl |
| com | 25.4 | 48.9 | 38.2 | 44.5 | 32.0 | 26.6 |
| org | 7.7 | 14.2 | 11.2 | 11.9 | 9.3 | 7.6 |
| de | 32.4 | 2.4 | 3.6 | 1.5 | 6.4 | 6.8 |
| net | 3.3 | 4.2 | 4.3 | 4.0 | 3.5 | 3.1 |
| fr | 1.6 | 0.7 | 12.2 | 2.9 | 0.8 | 1.5 |

Table 2 lists the five most frequently occurring TLDs across the language-specific lists. Three of the most commonly used TLDs on the internet, `com`, `net`, and `org` are present. The entries `de` and `fr` are due to an unsurprising bias of the German and French list, the second and third largest lists respectively, towards these TLDs. Around 420 K domains in the German list have `de` as their TLD and around 118 K entries in the French list have TLD `fr`.

We would like to note that the extracted URLs contained thousands of invalid TLDs, which was a point of interest in previous research [17] when analyzing such lists. Due to the human component in adding external links to articles, this is to be expected and not further elaborated on in this paper.

### 4.2. Subdomain Depth

Looking at the subdomain depths in Table 1, we notice a significant discrepancy between the Wikipedia lists and the Alexa Top list. With 590 K entries, the Alexa list almost exclusively consists of base domains, whereas our lists comprise around 14% to 23% base domains each. Conversely, up to 78% percent of domains in the Wikipedia lists have subdomain depth 1, compared to Alexa's 14%. Worth noting is that 60-70% of these domains with subdomain depth 1 have the `www.` prefix, which


doi: 10.2313/NET-2022-01-1_08

TABLE 3: Top 10 ASes by the number of contained domains from the Joint list.

| IPv4 | | | | IPv6 | | | |
|---|---|---|---|---|---|---|---|
| AS | Domains | Addresses | Prefixes | AS | Domains | Addresses | Prefixes |
| AS13335 - CLOUDFLARENET | 442K | 80K | 175 | AS13335 - CLOUDFLARENET | 399K | 75K | 30 |
| AS16509 - AMAZON-02 | 209K | 49K | 1.5K | AS6724 - STRATO | 49K | 473 | 2 |
| AS53831 - SQUARESPACE | 208K | 27 | 3 | AS8560 - IONOS-AS | 48K | 2.8K | 3 |
| AS15169 - GOOGLE | 174K | 25K | 266 | AS16509 - AMAZON-02 | 29K | 15K | 143 |
| AS58182 - wix_com | 147K | 25 | 4 | AS8972 - Host Europe | 21K | 2.8K | 2 |
| AS14618 - AMAZON-AES | 140K | 24K | 108 | AS51468 - ONECOM | 20K | 20K | 1 |
| AS16276 - OVH | 137K | 37K | 87 | AS15169 - GOOGLE | 20K | 222 | 14 |
| AS8560 - IONOS-AS | 118K | 10K | 34 | AS20773 - GODADDY | 13K | 10K | 1 |
| AS46606 - UNIFIEDLAYER-AS-1 | 90K | 27K | 129 | AS16276 - OVH | 12K | 1.9K | 5 |
| AS26496 - GO-DADDY-COM | 26K | 9K | 285 | AS54113 - Fastly | 10K | 219 | 16 |

does not provide us with any more interesting targets than base domains. Our lists do contain a considerable amount of domains with subdomain depth 2 or greater, leading to potentially interesting targets. In the Joint list, there are ≈210 K domains with subdomain depth 2 and ≈28 K with a subdomain depth larger than 3. The Alexa Top list has 740 and 35 such domains, respectively. This suggests that our Joint list covers domains beyond an entity's main web presence.

### 4.3. Intersection with the Alexa Top 1M

The intersection between hitlists is an important measure and was studied in previous research [17] as a large overlap may indicate that potential biases and shortcomings in one list are also present in the other. Of the approximately 3.5 M domains in our Joint list, about 150 K can be found on the Alexa Top list. Most of this overlap comes from entries in the English list. All other lists have intersections consistently below 10% and in total only contribute 7K domains to the overlap of the Joint list. This indicates that our lists are a more diverse source for the generation of a hitlist that goes beyond the most popular domains. The generally low overlap might be explained by the fact that the broad diversity of Wikipedia articles results in many external links pointing to niche, regional and unknown domains.

## 5. Biases

In this section, we analyze our target address for potential biases by inspecting their distribution over ASes, prefixes, and IP addresses. We conclude the section by checking IPv6 adoption and the use of privacy extensions across our targets. The following analysis is based on the addresses resolved from the Joint list only.

### 5.1. AS and Prefix Distribution

Table 3 shows the top 10 ASes most domains within our list belong to. For both IPv4 and IPv6, CLOUDFLARENET is in first place. About 11% of all domains resolved to an IPv4 address and 52% of IPv6 addresses resolved to are within AS13335-CLOUDFLARENET and 175 and 30 of its prefixes respectively. Six of the 10 ASes are found on both sides, while SQUARESPACE, wix_com, AMAZON-AES, and UNIFIEDLAYER-AS drop out of the top 10 when considering IPv6 addresses. With STRATO, IONOS-AS, Host Europe, and GODADDY, about 130 K (17.1%) of all IPv6 domains are located in a German AS and within 8 of their prefixes.

The domains are distributed over 19976 ASes (IPv4) and 2304 ASes (IPv6), yet the top 10 ASes contain 45% and 80% of them respectively. These results carry over to the approximately 69 K IPv4 and 3.4K IPv6 prefixes covered in total. All domains in the top 10 ASes are within 2591 (3.7%) and 217 (6.2%) of all covered prefixes. Figure 1 provides a graphical representation of these results. Beyond the top 10, we find that the top 100 ASes cover 77% and the top 250 approximately 85% of all IPv4 domains. For IPv6, it is more significant as the respective number of top ASes contain 96% and 97% of all domains.

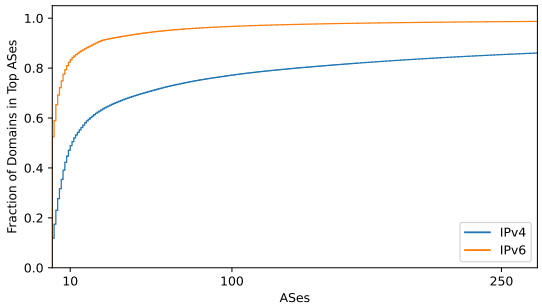


Figure 1: CDF showing address distribution over top ASes.

### 5.2. IP Addresses

To identify a possible bias towards a small set of IP addresses, we check how many domains are resolved to the same address. In total, we gathered 887 K unique IPv4 and 175 K unique IPv6 addresses.

**IPv4** Table 4 shows the top IPv4 addresses appearing the most in our hitlist. Given the purpose of external links, it is no surprise that most addresses belong to well-known web hosters like SQAURESPACE and Wix.com. There are no significant differences among the Top 8, with approximately 50 K occurrences each. In total, around 450 K (12.8%) domains are resolved to these addresses. Figure 2 shows the distribution over the top 5000 IPv4 addresses. We see that the top 100 addresses account for 22%, the top 1000 for 32%, and the top 4000 for 39% of all domains. After the top ≈300 K addresses, we have a one-to-one mapping between domain and address.

**IPv6** In Table 5, the top 10 most occurring IPv6 addresses are listed. Structurally, the table is similar to that of the IPv4 addresses. The top 5 addresses are resolved to from


doi: 10.2313/NET-2022-01-1_08

TABLE 4: Top 10 most frequently occurring IPv4 addresses in the target list.

| Address | # | AS |
|---|---|---|
| 198.185.159.144 | 53 K | SQUARESPACE |
| 198.185.159.145 | 50 K | SQUARESPACE |
| 198.49.23.145 | 50 K | SQUARESPACE |
| 198.49.23.144 | 50 K | SQUARESPACE |
| 185.230.63.171 | 49 K | wix_com |
| 185.230.63.107 | 49 K | wix_com |
| 185.230.63.186 | 49 K | wix_com |
| 184.168.131.241 | 47 K | GO-DADDY-COM-LLC |
| 3.223.115.185 | 29 K | AMAZON-AES |
| 192.0.78.24 | 23 K | AUTOMATTIC |

around 4800 domains each where the top 4 belong to AS15169-GOOGLE. In total, the top 10 addresses cover around 39 K (5.2%) of all domains. Looking again at Figure 2, we find that the distribution over IPv6 addresses follows a similar slope to that of the IPv4 addresses.

TABLE 5: Top 10 most frequently occurring IPv6 addresses in the target list.

| Address | # | AS |
|---|---|---|
| 2001:4860:4802:32::15 | 4886 | GOOGLE |
| 2001:4860:4802:36::15 | 4842 | GOOGLE |
| 2001:4860:4802:34::15 | 4840 | GOOGLE |
| 2001:4860:4802:38::15 | 4837 | GOOGLE |
| 2a05:d014:9da:8c10:306e:3e07:a16f:a552 | 4650 | AMAZON-02 |
| 2a01:238:20a:202:1086:: | 3599 | STRATO |
| 2a01:238:20a:202:1162:: | 3218 | STRATO |
| 2003:2:2:15:80:150:6:143 | 2840 | DTAG |
| 2606:4700:90:0:b518:199c:8a1f:d33b | 2736 | CLOUDFLARENET |
| 2a01:238:20a:202:1064:: | 2393 | STRATO |

The top 100, 1000, and 4000 IPv6 addresses account for 14%, 25%, and 29% of all domains respectively. Here we have a one-to-one mapping between domain and address after 74 K addresses. Interestingly another German AS, DTAG, is the only one appearing in either address table, which is not part of Table 3.

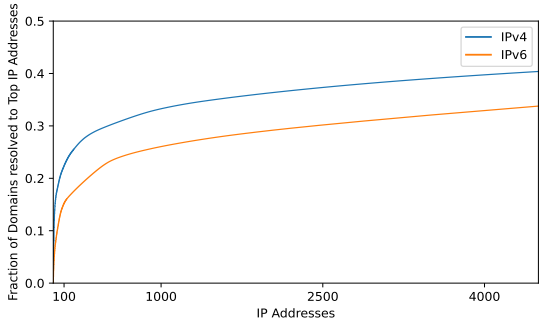


Figure 2: CDF showing domain distribution over top IP addresses.

**IPv6 adoption** IPv6 adoption across the internet was a network characteristic of interest in previous research [25]. Of our ≈3.5 M domains, around 750 K could be resolved to an IPv6 address. This represents an adoption of 21.7%. We take the native IPv6 traffic google receives [26] as a reference for the adoption on the internet, which is 31% as of June 02, 2021. Our list falls well below that. This again might be because of the diverse, possibly niche, and regional nature of external links. Additionally, we determine how many domains can be resolved to an IPv6 address, whereas Google passively measures user traffic.

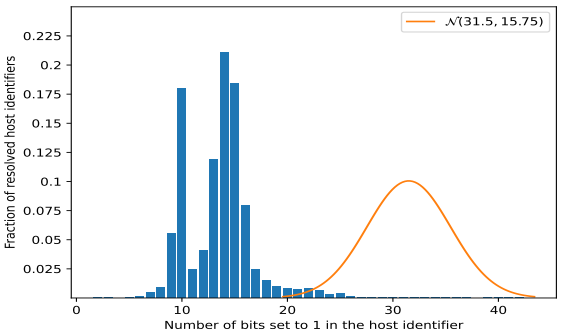


Figure 3: Bit distribution over IPv6 host identifiers.

**Privacy Extensions** RFC 4941 introduced privacy extensions to reduce the traceability of MAC addresses due to the use of Stateless Address Autoconfiguration. Using privacy extensions, the interface identifier, i.e., the last 64 bits, are replaced by random bits. An approximation for the sum of these single bit distributions is the normal distribution $\mathcal{N}(31.5, 15.75)$ [6]. We analyzed the interface identifiers of our IPv6 addresses. The distribution for the sum over the bits is shown in Figure 3. We see that our sample of host identifiers does not match the normal distribution. This shows that most of our targets are not using privacy extensions. Considering that most of our targets are presumably web servers having no need to mask their host identifiers for the sake of reducing traceability, this is not too surprising.

## 6. Conclusion and Future Work

In this work, we analyzed external links from Wikipedia articles as a source for creating a hitlist. We found that our Joint list has similar TLD coverage and higher average subdomain depth than the Alexa Top list. Around 21% of the domains in the Alexa list are also present in the Joint list. When evaluated for biases, our hitlist showed a significant propensity towards a small number of ASes and prefixes. In addition, a large portion of domains are resolved to a small set of IPv4 and IPv6 addresses. We have seen that the IPv6 adoption of our targets is below the general adoption and that most of them do not use privacy extensions.

We note that this work evaluated the hitlist in isolation without comparing it with existing alternatives. This could be addressed in future work to determine the relative value of this method. An attempt to eliminate found biases might increase the quality of the hitlist. Other potential aspects for future work include assessing possibilities to manipulate external links, monitoring the change of the domain names over a longer period of time, considering further network characteristics, checking for additional biases, and incorporating additional Wikipedia language editions.


doi: 10.2313/NET-2022-01-1_08

# References


[1] J.-J. Pansiot and D. Grad, "On Routes and Multicast Trees in the Internet," *SIGCOMM Comput. Commun. Rev.*, vol. 28, no. 1, p. 41–50, Jan. 1998.

[2] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet map discovery," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, vol. 3, 2000, pp. 1371–1380 vol.3.

[3] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast Internet-wide Scanning and Its Security Applications," in *22nd USENIX Security Symposium (USENIX Security 13)*. Washington, D.C.: USENIX Association, Aug. 2013, pp. 605–620.

[4] R. Graham, "MASSCAN: Mass IP port scanner," Available at https://github.com/robertdavidgraham/masscan, [Online; accessed 01-June-2021].

[5] 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*, D. Choffnes and M. Barcellos, Eds. Cham: Springer International Publishing, 2019, pp. 161–177.

[6] O. Gasser, Q. Scheitle, S. Gebhard, and G. Carle, "Scanning the IPv6 Internet: Towards a Comprehensive Hitlist," in *In Proceedings of the Traffic Monitoring and Analysis Workshop*, 2016.

[7] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczyński, S. D. Strowes, L. Hendriks, and G. Carle, "Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists," in *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 364–378.

[8] A. Murdock, F. Li, P. Bramsen, Z. Durumeric, and V. Paxson, "Target Generation for Internet-Wide IPv6 Scanning," in *Proceedings of the 2017 Internet Measurement Conference*, ser. IMC '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 242–253.

[9] Wikipedia, "Wikipedia:External links," Available at https://en.wikipedia.org/wiki/Wikipedia:External_links, 2021, [Online; accessed 24-July-2021].

[10] P. van Dijk, "Finding v6 hosts by efficiently mapping ip6.arpa," Available at https://web.archive.org/web/20161121215042/http://7bits.nl/blog/posts/finding-v6-hosts-by-efficiently-mapping-ip6-arpa, [Online; accessed 01-June-2021].

[11] F. Marquardt and C. Schmidt, "Don't Stop at the Top: Using Certificate Transparency Logs to Extend Domain Lists for Web Security Studies," in *2020 IEEE 45th Conference on Local Computer Networks (LCN)*, 2020, pp. 409–412.

[12] P. Foremski, D. Plonka, and A. Berger, "Entropy/IP: Uncovering Structure in IPv6 Addresses," in *Proceedings of the 2016 Internet Measurement Conference*, ser. IMC '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 167–181.

[13] D. Plonka and A. Berger, "Temporal and Spatial Classification of Active IPv6 Addresses," in *Proceedings of the 2015 Internet Measurement Conference*, ser. IMC '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 509–522.

[14] R. P. Sonar, "Forwards DNS Data," Available at https://opendata.rapid7.com/sonar.fdns_v2/, [Online; accessed 27-May-2021].

[15] R. NCC, "IPMap," Available at https://ftp.ripe.net/ripe/ipmap/, [Online; accessed 27-May-2021].

[16] Alexa, "Top 1M sites," Available at https://www.alexa.com/topsites, [Online; accessed 18-May-2021] http://s3.dualstack.us-east-1.amazonaws.com/alexa-static/top-1m.csv.zip.

[17] 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 *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 478–493.

[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 *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, ser. NDSS 2019, 2019.

[19] 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 *Proceedings of the Internet Measurement Conference*, ser. IMC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 351–357.

[20] C. Lever, P. Kotzias, D. Balzarotti, J. Caballero, and M. Antonakakis, "A Lustrum of Malware Network Communication: Evolution and Insights," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 788–804.

[21] P. Hoffman, "Collecting Typical Domain Names for Web Servers," Available at https://www.icann.org/en/system/files/files/octo-023-24feb21-en.pdf, 2021, [Online; accessed 02-May-2021].

[22] T. U. of Munich, "MassDNS," Available at https://github.com/blechschmidt/massdns, 2021, [Online; accessed 09-June-2021].

[23] N. Labs, "Unbound," Available at https://github.com/NLnetLabs/unbound, 2021, [Online; accessed 09-June-2021].

[24] IANA, "TLD Directory," Available at https://data.iana.org/TLD/tlds-alpha-by-domain.txt, 2021, [Online; accessed 01-June-2021].

[25] J. Czyz, M. Allman, J. Zhang, S. Iekel-Johnson, E. Osterweil, and M. Bailey, "Measuring IPv6 Adoption," in *Proceedings of the 2014 ACM Conference on SIGCOMM*, ser. SIGCOMM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 87–98.

[26] Google, "IPv6 Statistics," Available at https://www.google.com/intl/en/ipv6/statistics.html, 2021, [Online; accessed 01-June-2021].