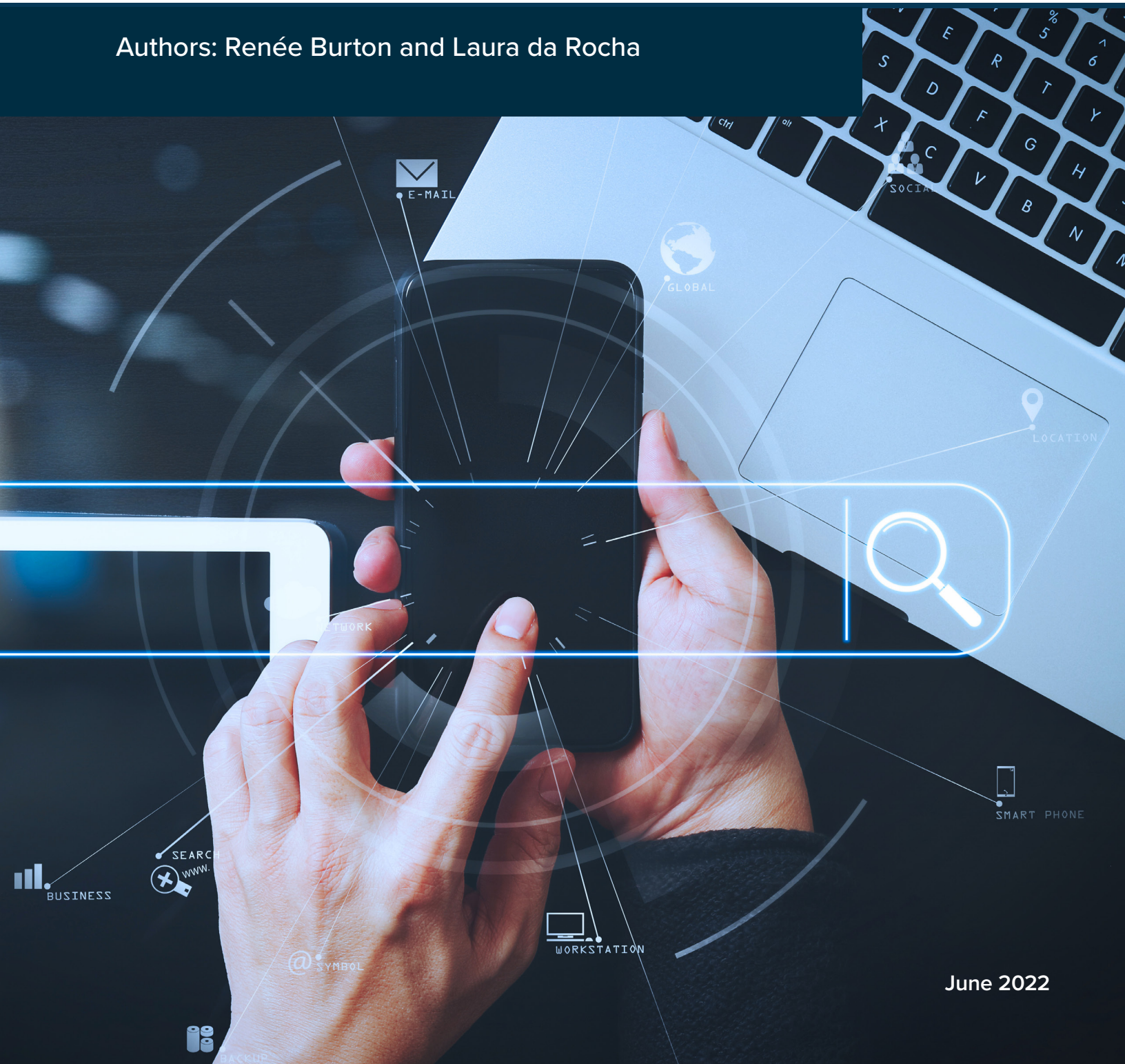
infoblox

WHITE PAPER

No Ranking List Is Perfect: A Top Domains List Comparison

Authors: Renée Burton and Laura da Rocha



June 2022

Disclaimer

Infoblox publications and research are made available solely for general information purposes. The information contained in this publication is provided on an “as is” basis. Infoblox accepts no liability for the use of this data. Any additional developments or research since the date of publication will not be reflected in this report.

Table of Contents

Considerations for Replacing Amazon Alexa Rankings

Amazon discontinued production of its popular Internet domain ranking list, Alexa, on May 1st, 2022 and many users of the service are scrambling to find a replacement.[1] Widely used for purposes ranging from search engine optimization to security applications, the website alexa[.]com began providing publicly available, free rankings of domains over twenty five years ago. Infoblox has not utilized Alexa for some years, having found statistical issues with the lists that made them unreliable for our use cases.[2] With users forced to find a new information source or devise their own, we want to share our insight into ranking Internet domains. This paper discusses the security use cases for domain rankings, the difficulties inherent in creating reliable ranking lists, provides a short technical assessment of alternative public ranking lists, and makes recommendations for replacing Alexa in your workflows.

Security Uses for Domain Rankings

Provided as a daily list of the top 1 million website domains, Alexa was incorporated into security products, algorithms, and decision support systems around the globe over the last two decades. The freely available data from Alexa, drawn from devices distributed around the world, became the standard for assessing popularity. Security analysts and researchers assume that if a domain is very popular it is less likely to host malicious content. As a result, the Alexa top 1 million domains was commonly used to:

- create a list of domains that should not be blocked by security appliances such as firewalls,
- assess the likelihood that a domain contained malicious content when analyzing network security events,
- train machine learning models designed to identify malicious domains, and
- attempt to identify anomalous traffic in a network by highlighting domains that were not contained in the Alexa list.

Until a few years ago, it was accepted without question that Alexa provided both insights into the most popular domains on the Internet and a means to easily distinguish legitimate domains.

Infoblox began studying domain rankings and assessing the quality of publicly available ranking lists in order to improve our automated allowlist process.[4] Based on our studies, we developed an algorithm to identify sets of domains which are of critical importance to our customers and, as a result, should not be blocked by the DNS firewall. Allowlists are necessary because the utilization of machine learning, statistical methods, and large scale automation can lead to false positives in a threat feed. Everyone in the security community faces this challenge, and most use an allowlist to mitigate the risk.[5] In order to balance the risks of false positives with the protection of our customers, we designed a Bayesian inference algorithm at the core of our process.[6]

---

1. https://www.theverge.com/2021/12/9/22825744/amazon-retiring-alexa-web-ranking-sevice
2. https://blogs.infoblox.com/security/going-beyond-whitelists-smartlisting-is-required-for-the-modern-enterprise/
3. Whitelists that Work: Creating Dynamic Defensible Whitelists using Statistical Learning, Renée Burton and Laura da Rocha, IEEE Proceedings of the APWG eCrimeX Conference 2019.
4. Infoblox no longer uses the terms whitelist and blacklist, traditionally used to indicate good and bad domains.
5. https://blogs.infoblox.com/security/going-beyond-whitelists-smartlisting-is-required-for-the-modern-enterprise/
6. Whitelists that Work: Creating Dynamic Defensible Whitelists using Statistical Learning, Renée Burton and Laura da Rocha, IEEE Proceedings of the APWG eCrimeX Conference 2019.

Popularity is a critical component of our allowlist algorithm. As a result, we studied every available domain ranking list we could find, performing a large amount of statistical analysis on them, as well as various algorithms for creating ranked lists. Unfortunately, not only do most publicly available lists have limited application to the allowlist use case for our customers, but algorithms to combine various sources into a single aggregate list are fundamentally flawed. These flaws appear to lead ranked list creators to continually try to tweak them to adjust, but this fails to address the core problem with the data. Our algorithm and the research behind it can be found in our paper Whitelists that Work: Creating Dynamic Defensible Whitelists Using Statistical Learning, published in the proceedings of the Anti-Phishing Working Group eCrimeX 2019 conference.[7]

As a direct result of that research, we developed our own algorithm for ranking domains. This second algorithm finds a maximum likelihood rank for a domain and also provides a range for the true rank of the domain. We wrote a white paper that explains the statistics behind our approach and provides statistical analysis of the advantages of our approach. The paper InfoRanks: Statistical Inference for Defining Internet Ranks is available on our website. In the remainder of this paper, we'll discuss the difficulties in creating top N lists and provide current metrics against some popular public lists.

Limitations to Creating Reliable Domain Rankings

Ideally, we could create a list of the most important domains in the world, ranked from the most to the least important. However, there are challenges to this seemingly-simple task. What does it mean to be important? This will vary by use case. If you are interested in the most popular websites, you are only interested in a subset of domains that exist in the world: website domains. Alexa and Majestic Million are two sources specifically focused on this type of popularity. A large portion of the Internet's traffic, though, is not limited to websites, and yet those other types of traffic are at least equally important. If your use case, as ours, needs to consider all kinds of domains, then you can't limit your data sources to those built around website usage.

7. https://blogs.infoblox.com/wp-content/uploads/infoblox-whitelists-that-work.pdf

Logging from the Domain Name System (DNS) is a natural way to obtain information about how all domains are used, because it will contain queries for websites, supporting infrastructure, and the myriad other types of domains that aren't used to support a web page. But this too is fraught with issues. The DNS is a complex, distributed system in which domains are independently managed and the results are transmitted through a tree-like hierarchy that includes a variety of caching mechanisms to ensure performance. Caching reduces the number of queries that are made by using a time-to-live for each domain, which might range from 0 seconds to a week or more. If we simply consider one domain to be more popular than another because there were more queries for the domain, our counts might be impacted by the time-to-live set by the domain owner. Bad actors could take advantage of this by setting their time-to-live values to 0 seconds and creating an unnatural rise in queries to their domains. When using DNS query logging to create ranked lists, you also need to consider the position of the logging within a network.

Cisco Umbrella offers a publicly available list of top queried domains from their OpenDNS networks; this data is collected between end users and recursive resolvers.[8] As such, it contains a lot of information about home queries, including unresolved queries, but does not contain infrastructure queries, for example those related to the delivery of mail or edge services to optimize content delivery in the cloud.

Over the last few years, a number of different researchers, including ourselves, found limitations in the public lists. Alexa data frequently contained questionable domains and appeared to be derived from an increasingly small perspective of the Internet. The Majestic Million was limited to websites and dependent on a web crawler. The Umbrella list contained a lot of noise and a perspective that didn't match many use cases. As individual lists were determined to be problematic, researchers started to cobble lists together in an effort to create a unified ranking, but this is a dangerous practice.

The Drawbacks of Blended Ranking Lists

As appealing as it sounds, it is ill-advised to combine various ranked domain name lists into a single list. Let's consider an example that is easier to understand: book rankings. The familiar New York Times Best Seller list provides a ranked list of the top book sales from various vendors each week, similarly Amazon offers a list of their top sales, as do others.[9] Now let's suppose we want a ranked list of the most read books in the world; this is a good analogy to ranking domain names globally. The challenges are immediately apparent. We can find data sources for sales, but not for books read from libraries, or shared copies distributed between friends and strangers alike. Distinguishing items that are read from those that are purchased is similar to the problems caused by DNS caching.

We might decide to infer which books are read from which books are sold. The reality is that each of these bestseller lists are unrelated data sets. Yes, they all contain books, but beyond that, they have little to do with each other. Each source is influenced by the demographics of its clientele and position in the market. Geography, size, price and publicity all play a role in how books will sell for each vendor. As an example of the variance, even the Amazon Best Selling Books and the Amazon Best Selling Books on the Kindle have little resemblance to one another.[10] Attempting to combine best seller lists into a single one that can be reliably interpreted as the most read books, in rank order, is not feasible because they are not measuring the same information.

8. https://umbrella.cisco.com/blog/cisco-umbrella-1-million
9. https://www.amazon.com/gp/bestsellers/books
10. https://www.amazon.com/Best-Sellers-Books/zgbs/books; https://www.amazon.com/Best-Sellers-Kindle-Store/zgbs/digital-text

A list called Tranco has gained a lot of attention in the research and practitioner communities as a means to take data from disparate sources like Alexa, Majestic, and Umbrella, and combine them.[11] Tranco grew out of frustration with Alexa in the security research community and a realization that ranked data was critical to a lot of work being done in the field, but lacked transparency and reproducibility. Their website states that "As the research community still benefits from regularly updated lists of popular domains, we provide Tranco, a new ranking that improves upon the shortcomings of current lists. We also emphasize the reproducibility of these rankings and the studies using them by providing permanent citable references." The authors analyzed several sources and ultimately proposed an intuitive algorithm that appeared at face value to solve many problems.[12] Their analysis of the problems with ranked lists is excellent, and mirrors much of our own unpublished results. They also provide software for their process.[13]

"As the research community still benefits from regularly updated lists of popular domains, we provide Tranco, a new ranking that improves upon the shortcomings of current lists. We also emphasize the reproducibility of these rankings and the studies using them by providing permanent citable references."

Unfortunately, the Tranco algorithm is fundamentally limited in ways similar to those they cite for other individual lists. In particular, they found that individual lists had little in common, could contain a number of malicious domains, and that ranks could be manipulated by bad actors. Their paper highlighted that the origins of these lists differ widely and are created through different methodologies. Because the individual publicly available sources for domain rankings were problematic for various reasons, they theorized that by combining them one could create a more representative and authoritative rank list. This is analogous to creating a single ranked list of most read books in the world.

Tranco takes several independent sources of ranked data and combines them using a positional voting system. The Dowdall rule used in Tranco is a variant of a Borda count, in which each candidate for a ballot is given a variable number of points depending on their rank in a multi-choice election.[14] The small nation of Nauru uses the Dowdall system to allow voters to select multiple candidates in the election of their Parliament, while the Borda Count is used in Slovenia. The Tranco algorithm considers each of the contributing rank lists as a ballot for domain popularity, summing the score for each domain across the ballots to obtain a final ranking. Prior to applying the Dowdall rule, they average data over 30 days for each list and remove domains that are only present for a few days; these filters are an effort to reduce manipulation of the lists by bad actors and eliminate domains with limited lifecycles.

11. https://tranco-list.eu/
12. https://tranco-list.eu/assets/tranco-ndss19.pdf
13. https://github.com/DistriNet/tranco-list
14. https://en.wikipedia.org/wiki/Borda_count

While the algorithm is easily understood, the results are not domain rankings in the sense of global relative popularity. The Dowdall Rule was designed as a mechanism for an electoral community to make multiple selections and arrive at a single result, but it is only one of many such systems. Indeed, Tranco offers lists created using both a Borda Count and the Dowdall Rule, which can produce different results for the same data.[15] The justification of these positional voting systems is based on mathematical principles, including the Law of Large Numbers, in which discrepancies of individual voters are overcome by the large number of voters. In spite of the supporting theory, Nobel Prize laureate Kenneth Arrow demonstrated that all such voting systems are unfair in some way.[16] As a result, positional voting systems are both dependent on the integrity of the underlying data and a large number of inputs.

Assessing Public Lists for Security Use Cases

To demonstrate the potential pitfalls of both the individual and combined lists, we performed a brief assessment of current data.[17] As mentioned earlier, domain ranking lists are often used in workflows to identify domains that are more likely legitimate. Some users also consider those that are not part of the Alexa top 1 million list to be more suspicious. In light of these use cases, we assessed the following characteristics:

- presence of malicious domains in the list,
- presence of newly registered domains in the list,
- overlap with each other and various network traffic seen at Infoblox resolvers, and
- variation of individual original ranks compared to final ranks in Tranco

The public lists we evaluated were:

- Alexa Top 1M domains. Alexa measured traffic through browser extensions and via website metrics. Alexa used a proprietary method to combine unique visitors to websites and page views into a single daily ranked list.[18]
- Cisco Umbrella Top 1M domains. This list contains a sample of domains queried through their OpenDNS resolvers, including unregistered domains and those unrelated to websites. Cisco applies some normalization to the sample to adjust for the number of clients.[19]
- Majestic Million domains. The Majestic Million is based on backlinks to a website, rather than traffic to a website. It is created from web crawls of the Internet and operates under the theory that the most important websites will be referred to most often by other websites.[20] The Majestic Million contains supporting raw data in addition to the rank.
- Tranco Top 1M domains. Tranco combines several other publicly available lists into a single ranked list using positional voting systems.[21] Tranco provides a default list and allows registered users to create custom lists based on various filters and algorithms.

15. Fraenkel, Jon & Grofman, Bernard. (2014). The Borda Count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia. Australian Journal of Political Science. 49. 10.1080/10361146.2014.900530.
16. Barnett, Janet Heine. “The French Connection: Borda, Condorcet and the Mathematics of Voting Theory.” (2021).
17. We have evaluated these sources in depth multiple times over multiple years. More details of our previous work can be found in our cited papers.
18. https://web.archive.org/web/20220120112816/https://blog.alexa.com/top-questions-about-alexa-answered
19. https://umbrella-static.s3-us-west-1.amazonaws.com/index.html
20. https://blog.majestic.com/development/majestic-million-csv-daily/
21. https://tranco-list.eu/

The Presence of Malicious Domains in Top 1M Lists

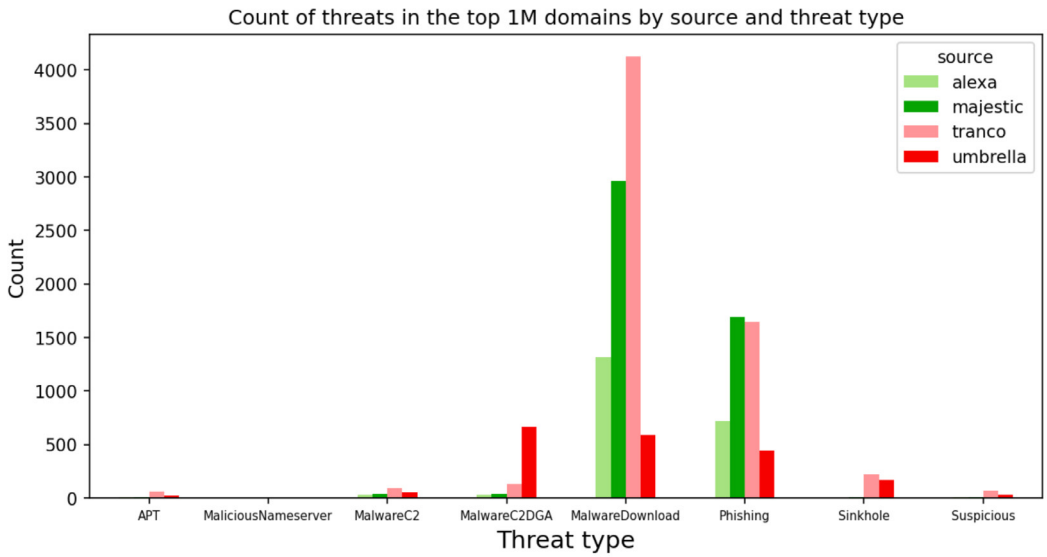
Many researchers, including ourselves, have found malicious domains present in public top ranked domain lists. This is due to a wide range of circumstances, including DNS caching and applications outside of a bad actor's control, as well as potential manipulation to create legitimacy by the actor. The Tranco researchers demonstrated the potential for rank manipulation and aimed to create a final ranking system that hindered manipulation of results. Although their source material was problematic, they estimated that by averaging data over time and combining it with positional voting, the cost of manipulation would at least quadruple and thus the inclusion of malicious domains would be minimized.[22] Using data from May 27th, 2022, and a set of known malicious domains, we found the Tranco combined list to have more high threat domains than any of the individual public sources included in the Tranco list.[23] Overall the number of threats is still fairly low, as the comparative threat set contained nearly 1.6M domains.

Table 1: The number of active threats found in each public list on May 27th, 2022. The active threat domains used in this table are high threats, originating from Infoblox Threat Intelligence, available in the Threat Intelligence Data Exchange (TIDE), and are second level domains only. The total number of threats considered was approximately 1.6M.

Top 1M List	Number of Infoblox Active High Threats
Tranco	6354
Alexa	2118
Majestic	4757
Umbrella	1970

The popularity of domains, including malicious ones, can vary widely over time. Phishing domains and domains associated with malware distributed through spam are two types of malicious domains that are frequently seen to have high rankings according to queries at a DNS resolver. The figure below shows the different types of threats observed in the public lists in the sample set.

Figure 1: The distribution of types of high threats observed in the public top 1 million domains on May 27th, 2022.



22. https://tranco-list.eu/assets/tranco-ndss19.pdf

23. This is limited to high threat domains in the Threat Intelligence Data Exchange (TIDE) that were contributed by Infoblox, are active as of this date, and in which the second-level-domain (sld) is the indicated threat.

Another way to measure the potential impact of manipulation of malicious domains is to consider the number of newly registered domains that occur in the list. Domains that are recently registered are often considered suspicious and many organizations block access to such domains out of concern that they may host malicious content. We compared registrations for the month of May 2022 with the public lists. The table below shows the number of recently registered, or updated, domains in each list, along with data about the popularity of these domains. Overall a fairly low number of recently registered domains are found in each list.[24] Notice that Alexa contains the largest number of such domains and that a newly registered domain reached a rank of 1963 in their list. All four of the top ranking recently registered domains in the table below are associated with phishing attacks.

Table 2: Presence of recently registered or updated domains in public top domains lists on May 27th, 2022. The most popular domains have a low rank in the list.

	Tranco	Alexa	Majestic	Umbrella
Recently Registered Domains	553	1,377	437	272
Top Rank	18,843	1,963	104,296	101,396
Top Ranked Domain	digimove[.]com	digimovie[.]one	prednisolone[.]directory	waterneed[.]click
Mean Rank	568,677	430,089	595,219	177,023

Overlap Between Top 1M Lists

We also considered the overlap between different sources. The Tranco researchers observed that the overlap between individual public lists was very low; only 2.48% of the domains existed in every source list. This is not surprising. In essence, it is analogous to comparing the top sellers in a bookstore in Madrid with the New York Times Best Seller list. By combining different sources, the Tranco authors hypothesized that the new list would have more relevance. Domains that were seen in multiple, but not all, sources could make their way into their final list and rankings

In reality, networks are highly unique, similar to the way that individual book sellers have unique demographics. The distribution of DNS queries, or website visits, is so distinct that we generally estimate about 60-70% of domains in any one network, on any given day, will not be visible in any other network; that is, only 30-40% of the domains will overlap. The so-called “long tail” in ranking lists, due to Zipf’s law, contain domains that individually do not significantly contribute to the total volume in a network, and vary across time and networks. To demonstrate, we used Infoblox DNS queries and measured the overlap with public ranking lists.

24. The WhoisXML Service was used to validate the domain registration data and some domains may be missing or have incomplete registration data.

Table 3: Overlap percentage between the top 1M domains in the public lists and Infoblox products on May 27th, 2022.

May 27, 2022	Tranco Overlap	Alexa Overlap	Majestic Overlap	Umbrella Overlap
Infoblox DNS Forwarding Proxies (DFP)	34%	19%	24%	26%
Infoblox BloxOne Clients (laptops, mobile devices)	45%	27%	35%	35%

As shown in the table above, the overlap between Tranco and the top 1 million domains queried in two of our category products is higher than either Umbrella or Alexa. The overlap, however, remains remarkably low with only 34% of the domains occurring in our top 1 million resolutions from the Infoblox DNS Forwarding Proxies (DFPs) being found in the Tranco list. Notice that in all cases, the overlap between the public lists and our endpoint client devices is somewhat higher. This demonstrates the bias that the public lists have to websites and endpoint devices. For reference, the overlap between the DFP and remote client top domains on this day is only 22%, again highlighting the unique nature of different networks with respect to DNS queries.

Another way to compute similarity is to use the Jaccard similarity coefficient. This metric provides a unitless mechanism to compare different sets. It is calculated by dividing the number of domains that intersect between the two sets by the total number of domains in both lists, i.e. their union. Jaccard similarity is useful when the lists have approximately the same number of items. The metric varies from 0 to 1, the closer the value is to 1 the more similar the two lists are; similarity, values close to 0 mean lists are not similar to one another.

To demonstrate the uniqueness of the top domains in different lists, and the influence of specific network traffic, we computed the Jaccard similarity for top 30k domains in the public lists, some of our products and within industry sectors on May 27, 2022. We normalized the domains using the ICANN only public suffix list (PSL) for consistency.[25]

From the table below we can observe that even at a considerably high rank threshold for comparing top N lists, i.e. highly popular domains, top 30k, we can observe that the overall similarity is still low, demonstrating the uniqueness of domains across networks.

May 27, 2022	Tranco	Umbrella	Alexa	Majestic	Infoblox DFP	Infoblox BloxOne Clients	Infoblox top N list (InfoRanks)	Education (Infoblox DFP)	Banking (Infoblox DFP)
Tranco	1.0								
Umbrella	0.30	1.0							
Alexa	0.32	0.16	1.0						
Majestic	0.34	0.15	0.18	1.0					
Infoblox DFP	0.25	0.37	0.12	0.12	1.0				
Infoblox BloxOne Clients	0.23	0.33	0.17	0.13	0.33	1.0			
Infoblox top N list (InfoRanks)	0.32	0.32	0.12	0.10	0.26	0.21	1.0		
Education (Infoblox DFP)	0.25	0.42	0.12	0.13	0.36	0.25	0.26	1.0	
Banking (Infoblox DFP)	0.16	0.22	0.07	0.08	0.22	0.18	0.16	0.21	1.0

Table 4: Jaccard similarity between the top 30k domains in the public lists, Infoblox products and within industry sectors on May 27th, 2022.

This table also demonstrates that considering domains outside of the Alexa list, or any other, as suspicious is not advisable. Because domain usage inside of networks is highly unique, the number of domains in the user's network that do not exist in a publicly available top 1 million list can be extremely large. Even limiting the comparison to the network's own top 1 million queried domains, were we to consider domains outside of the public list to be suspicious, the result would contain 500-600k suspicious domains each day.

25. Specifically we allowed wildcards but required that the TLD be one of the ICANN TLDs. This ensures that subdomains of domains like blogspot[.]com are not included. https://publicsuffix.org/

Comparison of Original to Final Rank Position

The figure below compares how the original ranks in each separate list vary from the final rank in Tranco for the top 250 domains, and highlights domains that are present in Tranco but not in the individual lists. The red dots represent domains that are in the Tranco top 250 domains, but that were not present in the top 1M original lists. We can observe that Umbrella is the most representative of Tranco for the top 250 domains, and that Alexa seems to be the least representative. This showcases that highly popular domains can be unique in different networks. Moreover, in cases that a domain is not present in all of the lists, it becomes easier for manipultion to occur because the data sample has decreased.

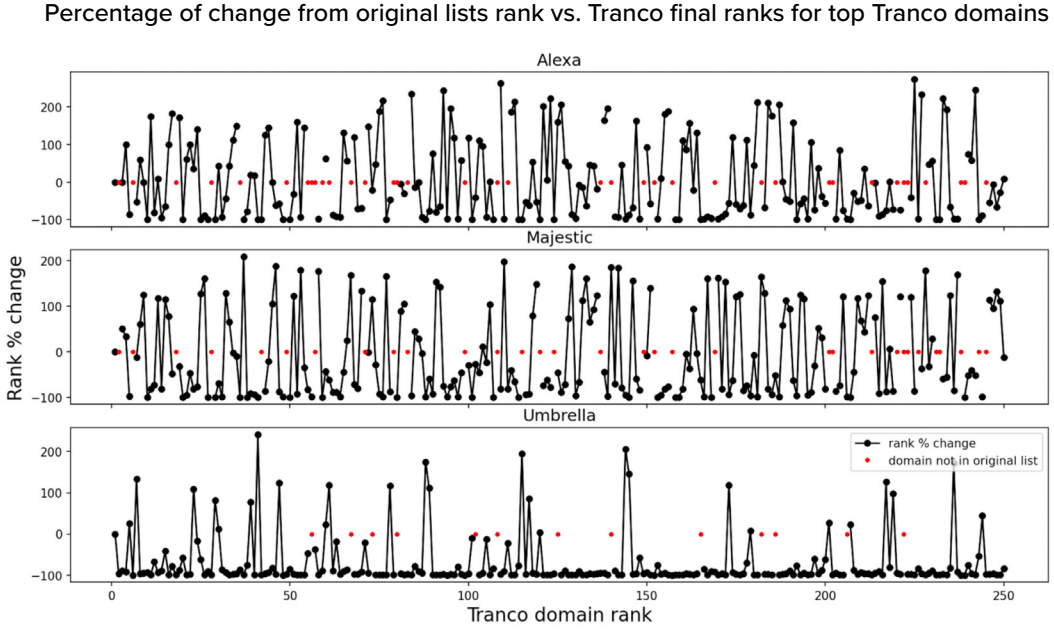


Figure 2: The percentage change in rank of the top 250 domains in Alexa, Majestic, and Umbrella to the corresponding rank in Tranco on May 27th, 2022. Red dots for each contributing list indicate a domain in the Tranco list that is not present in the respective contributor.

We calculate the percentage (%) change of the ranks for the individual lists (Alexa, Majestic and Umbrella), that we refer to as “original” lists in the plot, when compared to the final rank observed in Tranco. This is calculated by dividing the difference of the new rank (Tranco) by the original rank in the list, and multiplied by 100 to obtain the percentage. A positive change means that rank went down in popularity, that the domain in Tranco became less popular than in the original individual list. Similarly, a negative change means that domain ended up being more popular in Tranco than in the original list.

From the figure we can observe that in general, domains in the Umbrella list became more popular when processed into Tranco, when compared to their original rank in their own list. As observed before, Umbrella seems to have the highest number of domains that are in the top 250 Tranco and in their top 1M, which could be the explanation of why Umbrella ranks go higher in popularity since they may receive higher weight because they are present in their list. For Majestic and Alexa, we can observe that their domains go up or down in popularity, approximately equally likely, i.e. they have a similar amount of domains being more popular and less popular in Tranco.

Normalizing Ranks Across Subdomains

Another important factor we observed while evaluating various popular domain lists was the lack of consistency on how second level domains (SLDs) are extracted, and this impacts the way the final ranks get assigned to each of them. Extracting an SLD from the fully qualified domain name (FQDN) is not as straightforward as one may think. For example, the SLD for outlook[.]office365[.]com is easy to determine and is simply office365[.]com — obtained by extracting the two last labels of the domain. However, this rule is not applicable for all of the cases, if we follow the same logic for click[.]uol[.]com[.]br the SLD becomes com[.]br, which is not correct.

Conventionally, the Mozilla Public Suffix List (PSL) is used to obtain the SLD, sometimes referred to as the extended SLD, eSLD, or base domain.[26] While a standardized list, there are different ways to utilize the PSL. For example, the PSL contains both ICANN top level domains and private domains, which are maintained by private companies but may be used by a multitude of users. The list also contains modifiers such as wildcards.

As an example of the potential impact of subdomains and the PSL on rankings, consider blogspot[.]com. Blogspot is a private domain controlled by Google that contains user content in the subdomains and resource records.[27] As shown in the table below, the Alexa, Majestic, and Umbrella lists contain many resource records and subdomains of Blogspot at different ranks. Due to the different ways that each list considers domains, the ranks for a given domain can be absent altogether, and the many resource records and subdomains have a wide range of rankings within the lists. The fully qualified domain name bp[.]blogspot[.]com has a rank in each source list, but this is a subdomain that does not have its own resource records; that is, there is no IP address or other data directly assigned to the domain name. In the Tranco list, only blogspot[.]com is present, and it has the rank 68.

Alexa	Alexa Rank	Majestic	Majestic Rank	Umbrella	Umbrella Rank
blogspot[.]com	N/A	blogspot[.]com	25	blogspot[.]com	15
www[.]blogspot[.]com	35178	www[.]blogspot[.]com	N/A	www[.]blogspot[.]com	364090
bp[.]blogspot[.]com	698	bp[.]blogspot[.]com	149	bp[.]blogspot[.]com	6254
1[.]bp[.]blogspot[.]com	N/A	1[.]bp[.]blogspot[.]com	240	1[.]bp[.]blogspot[.]com	8911

Table 5: Domain ranks sample for the different variations of blogspot[.]com subdomains for the Alexa, Majestic and Umbrella lists.

26. https://www.publicsuffix.org

27. A subdomain is a domain that is part of another domain; it may not have resource records of its own. Resource records contain the actual data stored in the DNS. For example, bp.blogspot.com is a subdomain of blogspot.com, but has ho resource records. It contains 1.bp.blogspot.com which does have associated resource records and an IP address.

Replacing Alexa Rankings in Your Workflow

Top domain lists, regardless of the source, are inherently limited. They provide a specific perspective on a given network, or set of networks, and the tendency to try and normalize data can obscure the underlying truth. Malicious domains can, and do, occur with frequency for various reasons and can make their way into any popularity list. The variance in how popularity is calculated by different sources can reduce their relevance to security use cases.

So what should you do to replace your Alexa top 1M list in your workflows? Most importantly, use data sources that are relevant to your environment and use cases. For most security use cases, the best list of top domains is one generated from your own network traffic, or one containing similar traffic to your own. Like every bookseller on the planet, your network is unique and what is normal is best seen by monitoring your own traffic over time. If you choose to use one or several of the publicly available lists, let them inform, rather than dictate, decisions in your workflow.

Avoid combining lists through addition or a system like positional voting. These approaches bring together an "apples and oranges" mix of inherently different data sources and do not ultimately represent domain rankings. As we've seen, they can actually increase the inclusion of malicious domains. Instead, consider using an intersection or union of data normalized for your needs. Our allowlist algorithm incorporates multiple sources, but does not try to weight them in the process. It also considers known threats to reduce the likelihood of allowing bad behavior to pass through the network undetected. Using popularity as a way to create allow lists contains risks, so be mindful of the potential for accidentally allowing malicious domains to pass through your network.

While it is tempting to assume that domains which don't occur in a top 1 million list are suspicious, the unique nature of network activity will ensure that most of the domains in any given network are not part of any top 1 million domain list. By themselves, domain rankings do not promise legitimate behavior or identify suspicious activity. To utilize domain rankings to surface potentially suspicious behavior, combine the rankings with other features, such as domain age or name servers, to avoid overwhelming the security operations team.

In the end, your best bet to identify normal domains is to compare your own network with itself over time. While domains with a low volume of queries make up the vast majority of all queries, domains that are critical to the functioning of your network will surface due to their consistent popularity. You don't need to capture a million domains for this purpose. The majority of your important domains will occur on a much shorter list, though the exact length will vary depending on the particulars of your network traffic. We recommend you keep track of the top 1k-20k domains.

Corporate Headquarters
2390 Mission College Blvd, Ste. 501
Santa Clara, CA 95054

+1.408.986.4000
www.infoblox.com