A Stable and Open Method for Ranking Domains

Stéphane Coulondre
University of Lyon
Villeurbanne
France

Blake Sitney
Profound Networks, LLC
Bellevue WA 98004
USA

ABSTRACT

Researchers, advertisers, investors and an increasing number of professions are interested in making business decisions based on the relative rank of top websites. In response to this demand a handful of organizations offer top lists. Recent studies have shed light on the biases and possible manipulations by adversaries. In this paper we propose DomainRank, a well-founded and reproducible method for measuring and ranking domains, and we analyze its pros and cons compared to other top lists. The major differences lies in data gathering performed by *rendering* the Web, and in the ranking algorithm based on a modified version of the formally well-founded PageRank algorithm. It also provides a score for each domain. We have been running the method for 18 months and obtained stable results with low volatility even for long tail domains.

CCS CONCEPTS

• **Networks** → **Network measurement**; • **Information systems** → *Page and site ranking*.

ACM Reference Format:
Stéphane Coulondre and Blake Sitney. 2019. A Stable and Open Method for Ranking Domains. In *Proceedings of Internet Measurement Conference (IMC'19).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A broad range of businesses rely on ranked domain lists to make business decisions. Recent studies (Le Pochat et al. [2018]; Scheitle et al. [2018]) show that three lists are primarily used: Alexa Global [Alexa 2019c], Cisco Umbrella [Cisco 2019] and Majestic Million [Majestic 2019]. As revealed in these studies, domain ranking lists continue to have major shortcomings that impact the reliability of scientific research results. These shortcomings introduce biases into numerous business and security applications. It has also been shown recently that these lists can be manipulated (Le Pochat et al. [2018]). Additionally, these domain ranking lists provide only ordinal stack ranking values. This conveys a lack of formal knowledge that precludes further use and leverage by end-users, especially when it comes to security applications.

These studies reveal the fact that much improvement can be made regarding *consistency*, *transparency* and *stability*. With these three properties in mind, we have designed a domain ranking method based on formal specifications and public data only. Our method differs from others in several major ways: data gathering is performed by *rendering* the Web, the ranking algorithm is based on an adapted version of the widely-understood yet formally well-founded PageRank algorithm (Page et al. [1998]), and the ranking is induced by the relative score that each domain gets. In this paper we'll present the method and the resulting ranking, and will try to compare it to the other lists, according to the approach followed in Scheitle et al. [2018].

The gathering method we use goes beyond simply crawling and parsing of static Web pages, and *renders* each Web page while dynamically monitoring browser traffic. We therefore capture the requests generated by JavaScript execution, including hidden advertising and analytics snippets, JavaScript injections, dynamic page loading, etc. We have been running the method for 18 months and obtained stable results with low volatility even for long tail domains. DomainRank computes not only the ordinal rank value, but also the score for each of the top 5 million domains.

To ensure reproducibility and further usage, we provide to the research community 18 months of historical ranking data used in this paper as well as future releases on *domainrank.io*

2 RELATED RANKING METHODS

Ranking the Web is exemplified by top domain lists. The three top lists presented above are created by different methods and from different data sources, resulting in different sets of domains.

Alexa is probably the most popular top domain list. It is based on an undisclosed number of users who have installed a Web browser extension that tracks the sites they visit. Website popularity is then calculated based on user behavior.

IMC'19, October 21-23, 2019, Amsterdam, Netherlands

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

According to Scheitle et al. [2018] who dissected the Alexa toolbar, each visited page tranmits the following items: the entire URL including all GET parameters, screen/page size, referer, window IDs, tab IDs, and load time metrics. Alexa therefore tracks the browsing profile of its users which is prone to sociocultural and socioeconomic biases. It is worth stressing that although many studies rely on this list, Alexa however clearly states that "*Sites with relatively low measured traffic will not be accurately ranked by Alexa. We do not receive enough data from our sources to make rankings beyond 100,000 statistically meaningful.*" [Alexa 2019a]. To our knowledge, the Alexa browser extension is not available for smart phones - only for desktop and laptop computers - and the software and website are currently only available in English. These biases, among others, influence the ability to properly extrapolate global Internet traffic rankings beyond the demographic of technically savvy English speakers with laptops, mainly from 6 western countries [Alexa 2019b].

Cisco Umbrella relies on OpenDNS servers that can be used by anyone who does not want to rely on a particular Internet provider for their DNS. As such, it acts in a passive way by monitoring DNS requests, while gathering all the domains accessed in a Web browsing session (including all of the external domains' hosting resources – images, scripts, videos, etc. - linked to by the pages). If also gathers domains accessed by other IP protocols (ftp, ssh, etc.), and more generally all the domains submitted for resolution, including invalid ones. However, as the DNS protocol uses a cache to optimize traffic, the number and frequency of access requests is not gathered until the cache is refreshed. In short, although the domains accessed are known, there is a strong bias towards the number of users surfing from distinct DNS servers compared to the target website access frequency.

Majestic Million relies on active crawling, and they have set up their own engine. Their method is based, for each included domain, on backlink counts grouped by /24 IPv4 subnet. Since this method does not take into account the quality of these backlinks, it is prone to manipulation, as shown in Le Pochat et al. [2018], who also state that the completeness of their data is affected by how their crawler discovers websites. Majestic also sometimes includes both a domain and some of its subdomains. In such a setting, it is not clear how the rank calculation is made, if the impact of a subdomain also counts for its domain, or if it is removed from it prior to ranking. Moreover there is no straightforward relation between backlinks and the count of IPv4 /24 networks on which they are hosted. Content-Delivery Networks are extensively used for caching pages all around the world, so that hundreds of /24 subnets could host a single page. It is not clear how Majestic counts these networks when weighting domains.

As a consequence of how they are created, these lists are limited in how they can be utilized. For instance, their rankings are ordinal so there is no way to quantify other than empirically the rank differences between domains. This precludes placing a justification "value" on each position for the sake of comparison. Umbrella ranks DNS queries without knowing the real usage and traffic implied by these queries. Alexa ranks websites passively without taking into account the links injected by on-page scripting and, as far as we know, today Majestic does not either. If the quality and quantity of on-page links is to have value, surely the links which appear after rendering have some value. It's an aspect these other ranking approaches completely overlook.

But Web pages are dynamic. Users interact with Web pages that have been rendered from remote resources, often in response to user behavior. Any ranking system based exclusively on static HTML is inherently incomplete.

3 DOMAINRANK

DomainRank is a data modeling, gathering and rendering methodology used to measure and rank root level domains on the Web only. We never mix domains with their subdomains. While the approach is also applicable to subdomains, it's always wise to compare distinct items, i.e. TLDs against TLDs, root level domains against root level domains, subdomains against subdomains, and so on. If an inclusion relationship exists between any 2 of these distinct items, and unless this inclusion is formally quantified and appropriately dealt with, the results are biased.

Beyond research reproducibility which is required for a domain list to be considered in various applications, it also means for us, in the context of domain ranking, that the data and algorithms used to generate the scores and rankings are publicly available. There are no proprietary data or algorithms used in the DomainRank method. Two different parties making independent measurements from different Internet locations should produce the same or similar results. Reproducibility is a paramount constraint for any global measure of domain rankings. This can not be reached with passive monitoring from a proprietary or priviledged location, or by any "secret sauce" that prevents anyone to understand how the data is generated.

We also focus on stability, so that time has only a proportional impact on results. A stable measurement method should minimize global changes over time to both the individual domain scores and their relative position in the rankings, even for the long tail.

3.1 Data Gathering

We model the domain relationships as a simple directed graph, where each node is a domain and the outbound edges

are all the relationships gathered from hosted pages. There is no more than one edge between any two domains. We use the term *relationship* instead of the traditional Web term *link* which has many different meanings.

3.1.1 Rendering the Web. To truly measure and rank domains, it is essential to understand how one domain refers to another. The URL is just a seed for a lot of traffic to a large set of external domains (eCommerce, SaaS, marketing companies, cloud providers, CDNs, traffic monitoring, etc.). Web pages have become real pieces of software. *"To allow users to continue interacting with the page, communications such as data requests going to the server are separated from data coming back to the page"* [Wikipedia 2019]. This makes static HTML parsing insufficient when it comes to understanding comprehensive interactions that take place when a human user is browsing the Web. Common Crawl [Crawl 2019] e.g. indexes static HTML only. Therefore, results drawn from this dataset are incomplete.

In order to monitor this traffic, we need to render each webpage while dynamically monitoring the browser and its DOM. We capture the requests and updates to the DOM, including clickable links, images, forms, script injections, etc. generated by JavaScript execution, including obfuscated code. We therefore capture relationships generated by advertising and analytics snippets, pingbacks, CDNs, invisible pixels [Ruohonen and Leppänen 2018], etc. DomainRank has been able to use this technology for 18 months, in its own crawler engine called LinkExtract, generating history on over 150 million unique domains.

Rendering all the pages is of course extremely resource consuming. To keep things tractable, we use a statistical optimization model capable of allocating rendering resources according to the probability of domains hosting javascript-based relationships. We target these domains in priority, and ensure that an optimal tradeoff between rendering time, resources, and coverage is reached. Relationships taken into account by DomainRank in rendered pages are primarily those embedded in clickable Web links which are followed by website visitors, as well as a carefully chosen set of other tags invisible to the user. They are used by website owners to provide the resources required for rendering and controlling the behavior of the website.

3.1.2 Representativity. The score of each domain is computed from its inbound relationships. These relationships are not found on the domain itself or its subdomains, but rather come from other domains that point to the target domain. These influencing domains can be spread all over the Web. Rendering only the target domain is insufficient. We need to render the whole Web, or at least a representative subset. To reach this goal, we start from the same input seeds as those used by the Common Crawl engine in its monthly releases, which fetches "*a sample of the web and tries to make every monthly snapshot a representative and diverse sample on its own*". As the Web is like a living organism, new links are constantly added and old links are removed. In order to eliminate old links while keeping a complete view of domains, we derive rolling cumulative datasets which, according to Common Crawl published statistics, seems an optimal tradeoff between monthly coverage and the history of already seen URLs.

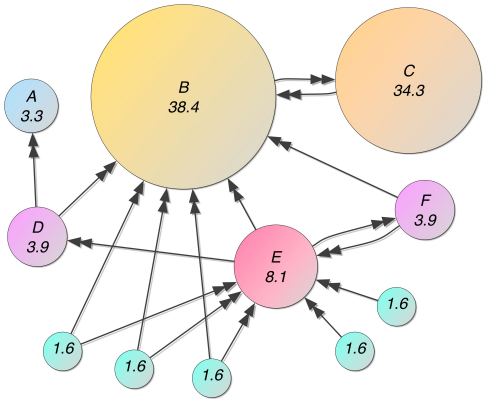

Figure 1: PageRank scores for a simple network, expressed as percentages. *C* has a higher score than *E*, even though there are fewer links to *C*; the one link to *C* comes from an important node and hence is of high value (source: Wikipedia)

3.2 Domain Scoring and Ranking

Once the relationships graph has been built, we use the PageRank algorithm as a centrality measure - as illustrated in figure 1. However, we do not keep the usual damping factor of 0.85 and use a proprietary value that does not impact the validity of the algorithm. According to several studies the "*folklore value of 0.85 is ideal for search engines and other applications where it is far more important to avoid false negatives then false positives*" (Avrachenkov et al. [2007]; Bressan and Peserico [2009]). The algorithm calculates a score for each domain. By construction, the sum of scores is 1. For any domain d, the semantics of its score can be thought as follows: "*Pick a random domain. The pages hosted on it refer to a unique set of external domains. Pick a domain from it, then the odds of finding d is the score*".

3.3 Results

Each DomainRank release is calculated from a graph of 150 million domains and 2.5 billion domain relationships. This graph is itself obtained by reducing a graph of 1.1 billion subdomains and 6 billion subdomains relationships.

To build this graph, we render 3 billion pages per month. Unlike Majestic for example which is SEO-oriented and as such needs to gather each and every possible backlink, we target the root domain level, and being exhaustive is not as important as being representative. Each month we build a graph based on 6-month rolling rendered data, keeping the most recent version of all pages. Therefore each release is based on 18 billion rendered pages.

However each of these 18 billion pages contain also relationships to unrendered pages, and basically the count of all gathered URLs is much higher. For ethical reasons we don't count them here because this information is of no value. It's not uncommon in the SEO market to see a questionable inflation of statistics stating dozens of billions to trillions of URLs. Rendering 18 billion pages, given the average size of a page (globally) and all its resources (images, scripts, etc.) is from experience nearly 1 megabyte in size, and requires nearly 3 seconds of full CPU time (up to 10 seconds user-time). On a machine with 128 processors, we can only render 128/3 = 42 pages per second. This requires a network bandwith of 42 megabytes per second, which is a sustainable performance for a gigabit connection. Therefore, for 3 billion pages each month, it takes 826 days for one machine, or nearly 9 days for 100 machines, with an approximate cost of $30,000. As detailed above, we don't render all the pages, but carefully select the ones with the best probability to host dynamically javascript-generated links, for the global cost to stay below a threshold.

4 DISCUSSION

Unlike other top lists, DomainRank is not released daily, but monthly. Daily releases carry so much noise that some averaging is necessary to reach a usable dataset [Le Pochat et al. 2018]. Historical DomainRank lists are available going back to September 2017. Therefore, in order to compare with other top domain lists, we build a historical dataset of 18 months from September 2017 to February 2019, reusing data made available by Scheitle et al. [2018]. As we don't want to introduce any bias, we don't preprocess the data or average it. We just keep one instance of each list each month, for example those available on the 28th day, reflecting the way a random user may just download them periodically.

4.1 TLD Coverage

We use the latest IANA TLDs list. On the reference period, the count of valid and invalid TLDs is shown in Table 2.

DomainRank covers a great set of TLDs, although smaller than Alexa on the top 1k. Being based both on Web links and the quality of their source, it is by default resistant to invalid TLDs, as each domain needs to have enough backlinks from other domains with sufficient influence to raise its score.

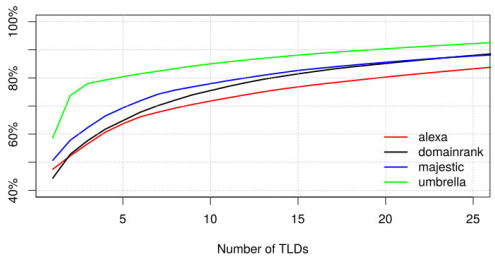

Figure 2: The cumulative distribution function of TLD usage across the lists

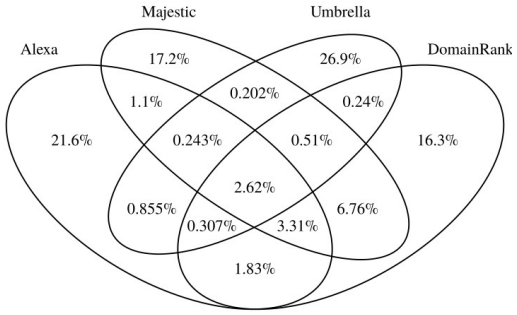

Figure 3: Intersection of the list in the last month release

However, some typical syntax errors may appear on the Web, thus perhaps even amplifying an invalid domain if it appears many times. Notice that this can be avoided by an appropriate domain checking process. We display the distribution of TLDs across the lists in Figure 2, and find the same results as Le Pochat et al. [2018], adding DomainRank to the chart. The *.com* TLD is still the most popular on every list, however its impact is the lowest on DomainRank where it accounts for 45% of the domains.

4.2 Similarity

As shown in Figure 3, DomainRank has the smallest exclusive set of domains with 16.3%, meaning that its content has the best coverage of other lists, and therefore it disagrees the least with the other rankings. It is closely followed by Majestic. This is confirmed by the Rank Biased Overlap score (Webber et al. [2010]) shown in Table 3. Umbrella is still the least similar, while DomainRank and Majestic are the most similar with an RBO of 50%. This is not a surprise however as they are both based on the graph structure of the Web, although they don't exploit it the same way.

4.3 Stability

We investigate the stability of the lists, by looking at how long a domain stays in a subset of the top. In Figure 4, we

Domain	Highest rank				Median rank				Lowest rank			
	Alexa	Umbrella	Majestic	DomainRank	Alexa	Umbrella	Majestic	DomainRank	Alexa	Umbrella	Majestic	DomainRank
google.com	1	1	1	2	1	4	1	2	2	6	1	2
facebook.com	3	5	2	1	3	8	2	1	3	12	2	1
netflix.com	21	1	429	991	28	1	459	1051	33	2	519	1188
google-analytics.com	41102	9	14934	5698	86285	13	19781	8064	164355	32	24586	11117
jetblue.com	2323	18351	4855	7270	3207	25128	5140	9387	4813	38975	5358	10925
mdc.edu	34401	211600	19200	37307	40411	282678	25213	40507	122309	369996	26356	48493
puresight.com	341186	831150	609160	206454	568275	-	681688	261833	981407	-	786813	372351

Table 1: Rank variation for a few domains across 18 months

	top 1k		top 1M	
	Valid	Invalid	Valid	Invalid
Alexa	162	0	1001	1
DomainRank	80	0	964	4
Majestic	57	0	810	12
Umbrella	22	0	778	2536

Table 2: TLDs validity

RBO(p=0.98)	Alexa	DomainRank	Majestic	Umbrella
Alexa	-	31.7 %	33.9 %	13.9 %
DomainRank	31.7 %	-	50.2 %	12.7 %
Majestic	33.9 %	50.2 %	-	16 %
Umbrella	13.9 %	12.7 %	16 %	-

Table 3: Rank Biased Overlap score

display for incremental top sublist sizes, the percentage of domains that remain in this sublist after 1, 3 and 12 months. For example, 74% of domains listed in Alexa top 50k remain in the top 50k the next month, and 60% of domains listed in the top 300k remain in the top 300k after 3 months. DomainRank has the best stability overall, while Alexa shows a very low stability even across short periods. Majestic and Umbrella show a low between 150k and roughly 350k, revealing greater instability in this area. As an illustration of long tail stability, we rebuilt in Table 1 the same table as in Scheitle et al. [2018]. We add *google-analytics.com* to this table, a traffic monitoring tool used my many by website owners. Tens of millions of websites, and their corresponding organizations and businesses, leverage Google Analytics. It is an example of domains that few would ever enter into a browser, yet these domains can be enormously influential because potentially millions of high quality domains link to them. Without surprise, Umbrella ranks it very high, although it's a purely "technical" domain, and on the other side of the ladder, it is not a popular Web destination for Alexa users or the general public. Majestic and DomainRank rank it fairly, although it is difficult to decide which rank such domains should get compared to human-browsable domains. DomainRank can easily calculate another ranking for technical domains by filtering on specific link types.

4.4 Volatility

Too much volatility can lower trust in a model, because there is less confidence in the ranking. Each variation can either be caused by a real gain or loss in the score, or caused simply by noise. We use the coefficient of variation, also called relative standard deviation, as a measure of dispersion. It is defined as the ratio of the standard deviation of rank to the mean across time. In Figure 5, we display the coefficient of variation on a log scale, against the cumulative distribution of domains, across a period of 1, 3 and 12 months. DomainRank is very close to Majestic compared to other lists, and their volatility is much lower. On the long tail however, nearly >70%-80%, DomainRank becomes less volatile. On periods longer than 1 month, we notice a threshold curves around 65%-70% for Majestic. The curves show a clear volatility acceleration compared to all other lists for which the whole cumulative distribution is smooth.

4.5 Manipulation

A very thorough analysis of manipulation impact and complexity is carried out in Le Pochat et al. [2018]. Here we try to figure out how resistant DomainRank is to manipulation. First of all, DomainRank is based on a structural analysis of the domain relationships, and as such does not carry the drawbacks of rankings based on live traffic-monitoring such as Umbrella and Alexa. The closest threat model is that of Majestic, which is also based on the domain relationships graph model. It has been shown in Le Pochat et al. [2018] that Majestic does not take the quality of backlinks into account when ranking domains and that purchasing large quantities of backlinks, especially those hosted on separate

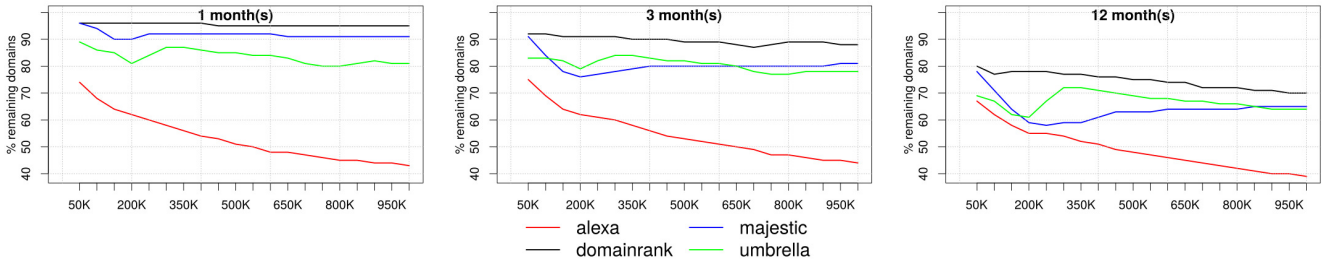

Figure 4: Stability of rankings against the size of the top sublist after different periods

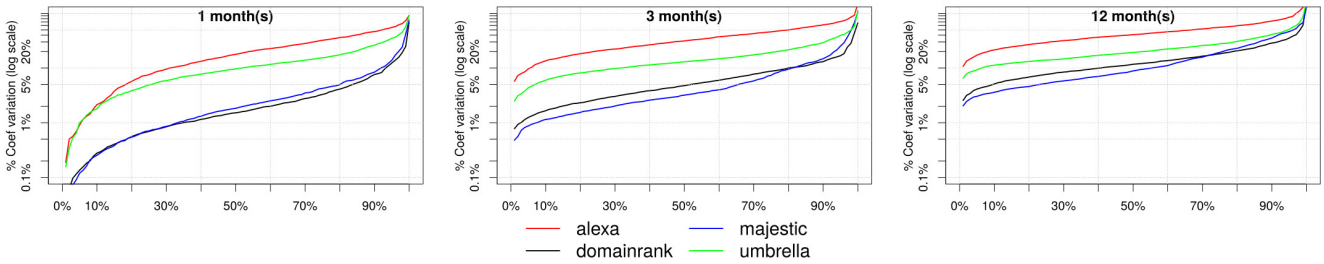

Figure 5: Volatility of rankings against the cumulative distribution of domains, after different periods

/24 IPv4 networks, can have a noticeable impact. However, DomainRank does take into account the quality of backlinks, which significantly raises the cost of purchasing backlinks. This naturally acts as an economic deterence factor against DomainRank manipulation.

4.6 Ranking vs Scoring

DomainRank ranking is not the primary result of the domain relationships graph analysis, but it is induced by the scores generated by the modified version of the PageRank algorithm. Therefore the primary results we calculate on domains are the scores. If we refer to the rankings, this is what we learn for February 2019, *facebook.com* is ranked #3 for Alexa, #1 for DomainRank, #2 for Majestic and #9 for Umbrella, and *baidu.com* is ranked #4 for Alexa, #35 for DomainRank, #10486 for Majestic and #1400 for Umbrella.

Clearly all lists agree on the fact than *facebook.com* is more popular than *baidu.com*. However, is *facebook.com* 10% more popular than *baidu.com*? 6 times more popular? We cannot know based on ordinal ranking. As DomainRank produces the scores for each domain, we can calculate the precise difference. In this example, *facebook.com* is scored 15.45e-03 and *baidu.com* is scored 0.53e-03. We can calculate the ratio of *facebook.com* score to the *baidu.com* score: $\frac{score(facebook.com)}{score(baidu.com)} = \frac{15.45}{0.53} = 28.92$. We interpret this to mean that *facebook.com* has 28.92 times greater chance of unique domains leading to it than *baidu.com*. In this context, "leading" means that there are direct links and indirect links on domains providing a pathway to a specific domain. However we can also figure out by further analysis that the month before, this number was 30.16, therefore *baidu.com* gained against *facebook.com*, and that 18 months before this number was 63.99, so there is a clear trend between these two domains.

5 CONCLUSION

In this paper we presented DomainRank, a methodology for modeling, gathering and rendering data to measure and rank domains. The DomainRank methodology produces a score based on the PageRank algorithm from which the ranking is derived. The result is a function of the graph of all website domains in relation to the structural components of Web sites rather than traffic analysis or individual Web browsing user behavior.

The DomainRank methodology produces a ranking that is reproducible, stable with low volatility, representative coverage and good resistance to manipulation. It offers users greater statistical insights into domain ranking and allows for their use in time series analysis and predictive analytics. Perspectives include more frequent releases and different rankings based on different use cases including: browsable domains, technical domains, CDNs, targeted service providers, among others.

REFERENCES

Alexa. *How are Alexa's traffic rankings determined?*, 2019a. URL https://support.alexa.com/hc/en-us/articles/200449744.

Alexa. *What countries does Alexa offer unique visitor, visits, and pageview estimates for?*, 2019b. URL https://support.alexa.com/hc/en-us/articles/204211004.

Alexa. *Alexa Web Information Service*, 2019c. URL https://aws.amazon.com/awis/.

K Avrachenkov, Nelly Litvak, and Son Pham. A singular perturbation approach for choosing pagerank damping factor. *Internet Mathematics*, 5, 01 2007. doi: 10.1080/15427951.2008.10129300.

Marco Bressan and Enoch Peserico. Choose the damping, choose the ranking? In *Proceedings of the 6th International Workshop on Algorithms and Models for the Web-Graph*, WAW '09, pages 76–89, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-540-95994-6. doi: 10.1007/978-3-540-95995-3_7.

Cisco. *Umbrella Popularity List, Top Million Domains*, 2019. URL https://docs.umbrella.com/investigate-api/docs/top-million-domains.

Common Crawl. *Common Crawl*, 2019. URL https:/commoncrawl.org.

Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. volume abs/1806.01156, pages 1–15. Internet Society, 2018. ISBN 189156255X.

Majestic. *The Majestic Million*, 2019. URL https://majestic.com/reports/majestic-million.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

Jukka Ruohonen and Ville Leppänen. Invisible pixels are dead, long live invisible pixels! In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, WPES'18, pages 28–32, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5989-4. doi: 10.1145/3267323.3268950.

Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. A long way to the top: Significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 478–493, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5619-0. doi: 10.1145/3278532.3278574.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, November 2010. ISSN 1046-8188. doi: 10.1145/1852102.1852106.

Wikipedia. *Web 2.0*, 2019. URL https://en.wikipedia.org/wiki/Web_2.0.