From Web Graphs to Prioritizing Web Crawls

Sebastian Nagel
sebastian@commoncrawl.org

3rd International Symposium on Open Search Technology, OSSYM 2021,
hosted by CERN, Geneva, Switzerland, Oct 11–13, 2021

About Common Crawl

- we're a non-profit that makes web data accessible to programmers and data scientists
- hosted as Open Data set on Amazon Web Services [1]
- for natural language processing, web science, semantic web, internet security research, ...
- 220 billion web pages (HTML) captured 2008 – 2021
- sample crawls, not a comprehensive crawl

Common Crawl Crawler 2008 – 2012

- in-house development by Ahad Rana [2, 3]
- batch-based (Hadoop MapReduce)
- yearly data releases
- pagerank calculation
- deduplication as post-processing step

Common Crawl Crawler 2013 – now

- Apache Nutch [4,5] – community-based open source crawler
- with few in-house modifications and extensions [6]
- batch-based (Hadoop MapReduce)
- (bi)monthly data releases

- (since 2016) Stormcrawler [7] used to crawl the continuously released news dataset

Prioritization – Which Pages or Sites to Crawl

Why prioritization is necessary? Why not just follow links?

- an average monthly crawl includes 3 billion page captures with
  - 500+ billion links
  - 25+ billion unique URLs linked
- need to select a representative sample of web pages
- given limited resources and the need for crawler politeness

Prioritization – Looking Back

2008—2012 in-house pagerank implementation

2012—2015 web search engine blekko [8] donates ranking and metadata of 140 million web sites and 22 billion pages

> *Common Crawl will use blekko's metadata to improve its crawl quality, while avoiding webspam, porn, and the influence of excessive SEO* [9]

- focus on efficiently and politely fetching web pages
- no need to maintain a large URL frontier and to "steer" the crawl

2016—2018 occasionally received seed donations

- up to 400 million URLs
- not enough to "feed" the crawler

2016 Alexa and Common Search rankings

2017—now rankings based on in-house webgraphs

Web Graphs based on Common Crawl

2013—2015 Web Data Commons, University of Mannheim: hyperlink graphs and rankings [10,11,12]

- page/host/domain-level hyper-link graphs
- host-level site ranking by harmonic centrality, pagerank, indegree centrality, Katz's index

2016 Common Search: host-level webgraph and pagerank [13,14]

2017—now host/domain-level webgraphs and rankings (harmonic centrality and pagerank) based on 3 monthly crawls

- publicly released webgraph datasets
- used to "steer" the crawler for the next 3 months
- harmonic centrality more robust against link spam than page rank [15]

Comparison of Web Site Rankings i

Are CC's graph-based rankings comparable with other web site rankings?

- Alexa top-1-million sites [16,17]
- Cisco Umbrella Popularity list [18]
- Majestic Million [19,20]
- Tranco list [21]
- latest CC domain-level harmonic centrality rankings [22]

… the comparison is inspired by the Tranco research paper [23], calculation and more detailed numbers in [24]

Comparison of Web Site Rankings ii

Caveats: lists are not entirely comparable because of different

- notion of a site (host, subdomain, registered domain)
- ranking method
  - Alexa: traffic / visitors
  - Cisco Umbrella: DNS traffic
  - Majestic: backlinks aggregated by IPv4 /24 subnets
  - Tranco: weighted combination of the above (plus Quantcast)
  - Common Crawl: harmonic centrality of hyperlink graph
- data aggregation period: daily updates vs. CC domain rankings released 3–4 times per year

Comparison of Web Site Rankings iii

CC's domain ranks are most similar to Majestic and Tranco

37% resp. 33% overlap, .32 and .35 rank-Biased overlap (RBO) [25]

Alexa	Umbrella	Tranco	Majestic	CC
google.com	google.com	google.com	google.com	googleapis.com
youtube.com	netflix.com	youtube.com	facebook.com	facebook.com
tmall.com	www.google.com	facebook.com	youtube.com	google.com
qq.com	microsoft.com	netflix.com	twitter.com	w.org
baidu.com	ftl.netflix.com	microsoft.com	instagram.com	twitter.com
sohu.com	prod.ftl.netflix.com	twitter.com	linkedin.com	youtube.com
facebook.com	api-global.netflix.com	instagram.com	microsoft.com	instagram.com
taobao.com	data.microsoft.com	tmall.com	apple.com	googletagmanager.com
360.cn	nrdp.prod.ftl.netflix.com	linkedin.com	wikipedia.org	gmpg.org
jd.com	ichnaea.netflix.com	apple.com	wordpress.org	linkedin.com
amazon.com	events.data.microsoft.com	qq.com	googletagmanager.com	gstatic.com
yahoo.com	netflix.net	wikipedia.org	youtu.be	cloudflare.com
wikipedia.org	partner.netflix.net	baidu.com	en.wikipedia.org	gravatar.com
weibo.com	prod.partner.netflix.net	sohu.com	pinterest.com	wordpress.org
zoom.us	preapp.prod.partner.netflix.net	googletagmanager.com	plus.google.com	pinterest.com
sina.com.cn	safebrowsing.googleapis.com	live.com	play.google.com	wikipedia.org
xinhuanet.com	windowsupdate.com	yahoo.com	vimeo.com	apple.com
live.com	live.com	amazon.com	maps.google.com	wordpress.com
microsoft.com	ctldl.windowsupdate.com	wordpress.org	goo.gl	vimeo.com
netflix.com	clientservices.googleapis.com	youtu.be	adobe.com	youtu.be

Comparison of Web Site Rankings iv

Geographic distribution (by country code top-level domain):
CC's top-million includes more domains from Europe, less generic domains (.com, .org, ...)

	Alexa	Umbrella	Tranco	Majestic	CC
(generic)	70.1845	84.5638	69.0205	68.7303	64.6606
Europe	10.4595	6.3905	16.3954	18.8020	23.5802
Asia	12.7613	5.3820	8.3497	6.8734	5.1279
South America	1.7806	1.2339	2.0610	1.6045	2.3730
Oceania	1.9455	0.7940	1.5058	1.5049	1.8515
North America	1.7606	1.3711	1.7441	1.7052	1.8062
Africa	1.1080	0.2647	0.9231	0.7793	0.5990
Antarctica	0.0002	0.0000	0.0004	0.0004	0.0016

Prioritization – A Deeper Look into the Current Implementation i

CC's crawler uses knowledge from previous crawls to identify relevant URLs. Rather than continuously queuing links (as a "traditional" web crawler does) the queues are filled shortly before the crawl using

- links sampled from the preceding crawl
- URL discovery via sitemaps [26]
- and a shallow crawl of sampled home pages

Domain-level harmonic centrality ranks are used

- to define a "budget" for every domain how many URLs/pages are sampled or fetched

Prioritization – A Deeper Look into the Current Implementation ii

- to sample sitemaps or home pages for URL discovery (always for top-ranking domains, sometimes for lower ranks)
- as domain-level scores "projected" to the page-level by OPIC [27] or inlink counts

Per-domain limits (summer 2021)

- top domains: 25 million URLs, 150k URLs per host, 500k subdomains
- long tail (below rank 24M): 3k URLs, 2.5k per host, 12 subdomains
- log distribution between top and tail

Link Spam – Challenging Centrality Measures

- spam is part of the web, it's ok if some is contained in the data
- October 2017: the crawler hit a spam cluster
  - crawled: 56 million pages (1.5% of the crawl), 70,000 domains
  - known from links: 320,000 domains, 2.5 billion subdomains
- highly branching spam clusters expensive for a crawler: every subdomain requires DNS look-up and robots.txt fetch/caching
- measures: set a strict limit of crawled subdomains per domain and try to detect and block the worst link spam clusters

Link Spam Detection i

- spam clusters are volatile
- must detect spam with (almost) no training data
- need binary rule (is a spam domain or not)
- simple heuristics proved to work with little supervision based on imbalances between
  - centrality score
  - outgoing links
  - number of subdomains

  low-ranking domains with too many outlinks or subdomains are suspicious
- once some nodes of a spam cluster are identified, other nodes are easily found by looking for a strongly connected subcluster in the graph

Link Spam Detection ii

Example based on the Jun/Jul/Sep 2021 domain-level graph taking as spam indicator an exceptionally high product of harmonic centrality rank and number of subdomains

sort	$log_2(r \cdot n)$	hc rank r	n subdomains	domain
1	44.93	33827380	993576	6suqmv2.site
2	44.34	50162956	445037	1st-muscle-guide.com
3	44.25	34012364	616681	wcpeoxt.icu
4	44.16	60683905	323917	ehime-di.com
5	44.09	36323509	515162	7ikoqnp.site
6	44.06	34195171	536038	m85g3vs.site
7	44.04	33824385	536460	5esg5j6.site
8	44.02	34925230	509545	mqv4s31.icu
9	43.90	33701024	487860	dcw7v3.xyz
10	43.81	35522472	433766	8s60fy.xyz
11	43.81	36357152	423051	76m30o.xyz
12	43.80	34202757	448281	x80u6n.xyz
				...
2371148	26.89	24	5176495	blogspot.com
				...
2767418	26.67	18	5913686	wordpress.com
				...

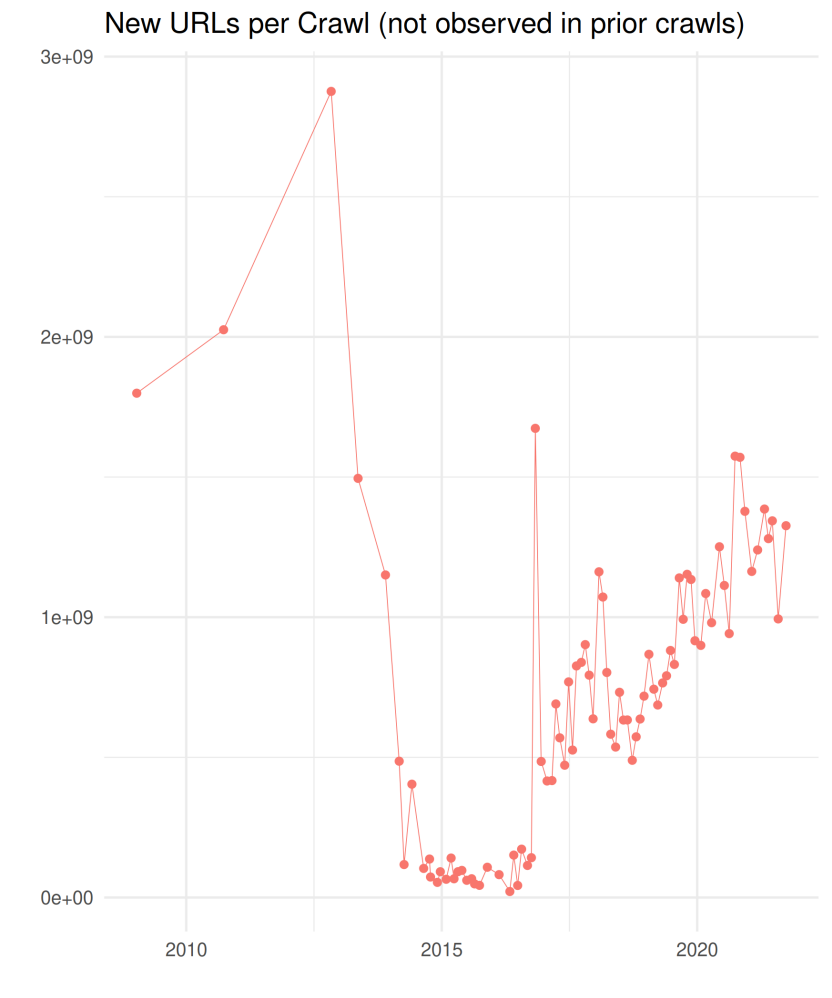
Are the Common Crawls Representative?

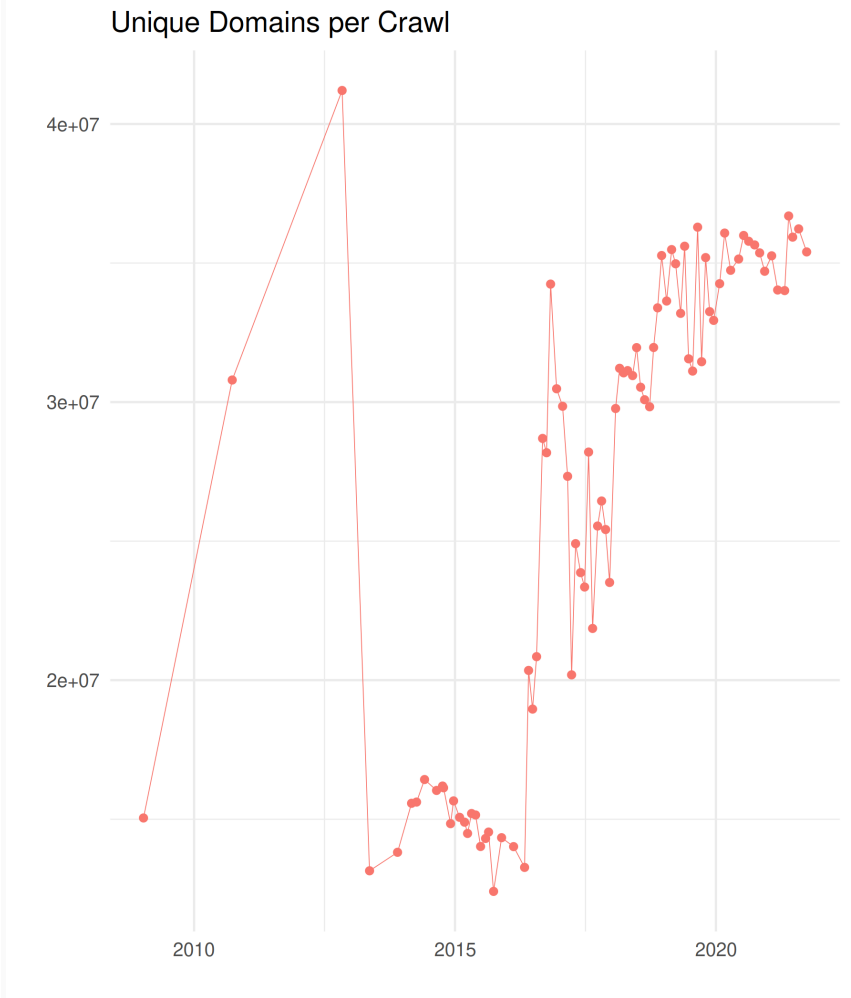
… and do the in-house rankings help do “steer” the crawler?

Aspects of representativity:

- breadth: coverage of unique domains (web sites)
- depth: per-site coverage
- regional coverage (top-level domains, content languages)
- amount of (near-)duplicates (both per crawl and over multiple crawl datasets)

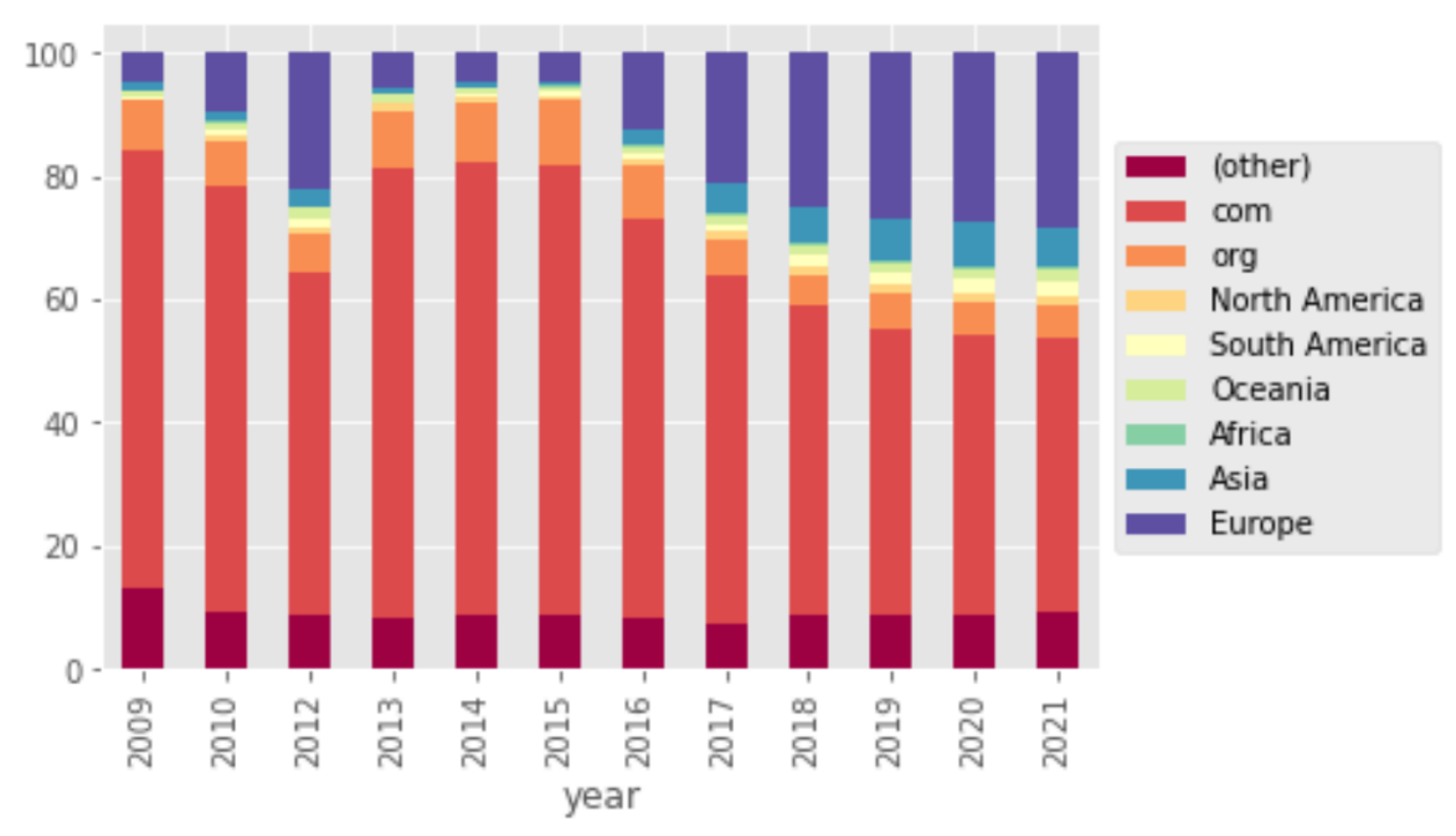
New URLs and Domain Coverage





Geographical Coverage

… by country-code top-level domain (percent of pages)



Lessons Learned, Outlook and Questions?

- relevance isn’t easy to achieve and measure
- we’re learning permanently and trying to get better
- data and source code is open to anybody – share your findings!

References i

1. Registry of Open Data on AWS https://registry.opendata.aws/
2. Ahad Rana 2010: Common Crawl – Building an open web-scale crawl using Hadoop. https://www.slideshare.net/hadoopusergroup/common-crawlpresentation
3. https://github.com/commoncrawl/commoncrawl-crawler
4. https://nutch.apache.org/
5. Jordan Mendelson 2014: Common Crawl's Move to Nutch.

   https://commoncrawl.org/2014/02/common-crawl-move-to-nutch/
6. https://github.com/commoncrawl/nutch
7. https://stormcrawler.net/
8. https://en.wikipedia.org/wiki/Blekko
9. https://commoncrawl.org/2012/12/blekko-donates-search-data-to-common-crawl/

References ii

10. Web Data Commons - Hyperlink Graphs, 2013
    http://webdatacommons.org/hyperlinkgraph/index.html
11. Robert Meusel, Sebastiano Vigna, Oliver Lehmberg, Christian Bizer, 2015: The Graph Structure in the Web – Analyzed on Different Aggregation Levels
    https://doi.org/10.1561/106.00000003
12. The Common Crawl WWW Ranking http://wwwranking.webdatacommons.org/
13. Common Search: Our first public datasets: Host-level WebGraph and PageRank https://web.archive.org/web/20170729110709/https://about.commonsearch.org/2016/07/our-first-public-datasets-host-level-webgraph-and-pagerank/
14. https://github.com/commonsearch/cosr-back/blob/master/spark/jobs/pagerank.py
15. Boldi, Paolo 2013: A modern view of centrality measures
    https://events.yandex.ru/events/science-seminars/boldi-23sep
16. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip

References iii

17. https://support.alexa.com/hc/en-us/sections/200063274-Top-Sites
18. http://s3-us-west-1.amazonaws.com/umbrella-static/index.html
19. http://downloads.majestic.com/majestic_million.csv
20. https://blog.majestic.com/development/majestic-million-csv-daily/
21. https://tranco-list.eu/
22. Common Crawls Jun/Jul/Sep 2021 webgraph dataset
    https://commoncrawl.org/2021/10/host-and-domain-level-web-graphs-jun-jul-sep-2021/
23. Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, Wouter Joosen 2019: Tranco: A research-oriented top sites ranking hardened against manipulation
    https://tranco-list.eu/assets/tranco-ndss19.pdf
24. https://github.com/commoncrawl/cc-notebooks/blob/master/cc-webgraph-statistics/comparison_domain_ranks.ipynb

References iv

25. Webber, Moffat, Zobel 2010: A similarity measure for indefinite rankings
    http://codalism.com/research/papers/wmz10_tois.pdf
26. https://sitemaps.org/
27. Serge Abiteboul, Mihai Preda, Gregory Cobena 2003: Adaptive on-line page importance computation https://dx.doi.org/10.1145/775152.775192