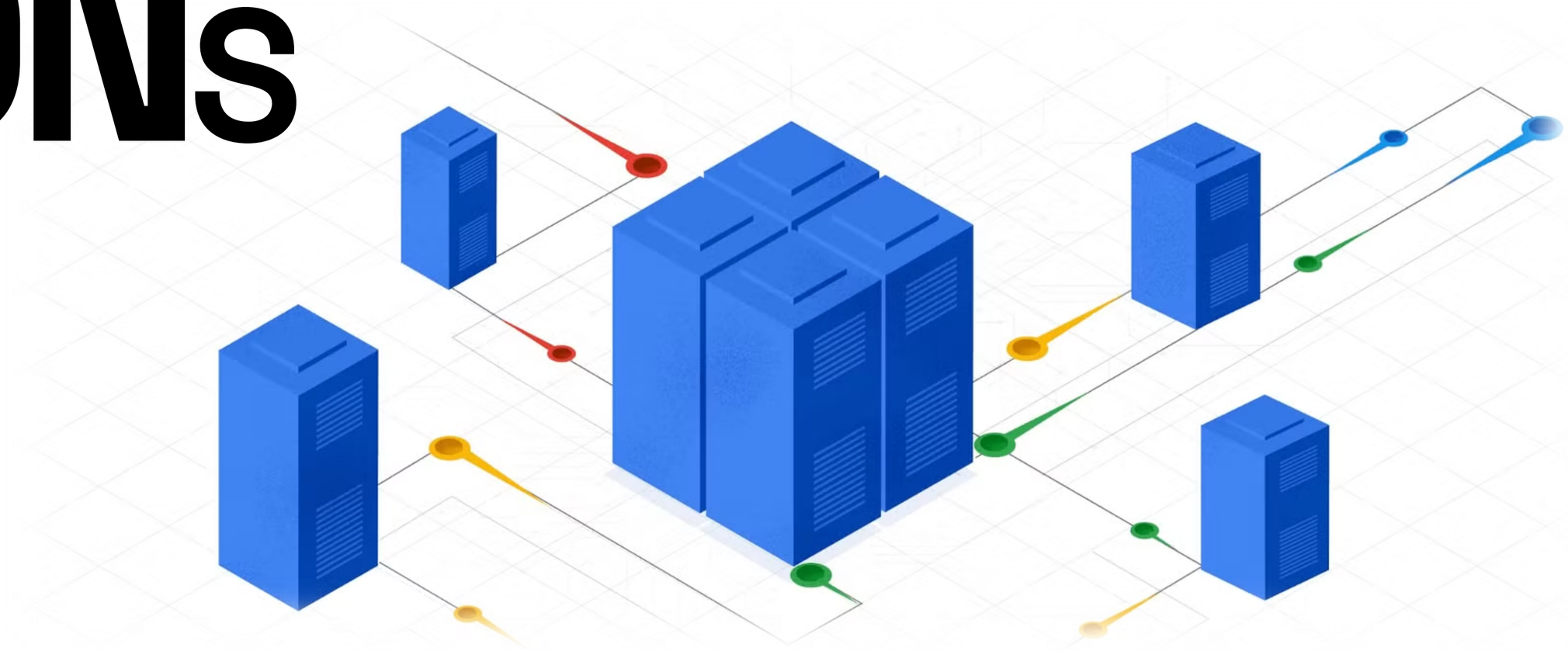
Hannes Koksch (hk058)

ANALYSIS OF CDNs

HOCHSCHULE DER MEDIEN



“29% of websites use a CDN to deliver HTML

— Web Alamanac 2022 [1]

“47% of websites that serve resources from subdomains use a CDN

— Web Alamanac 2022 [1]

Structure

Link to GitHub repository

git.new/cdns

- This presentation
- Source code
- Analysis results

Hannes Koksch (hk058)

GOALS AND RESEARCH QUESTIONS

Goals

- Understanding the CDN market
- What providers are there and how popular are they?
- Improving technical understanding

Research questions

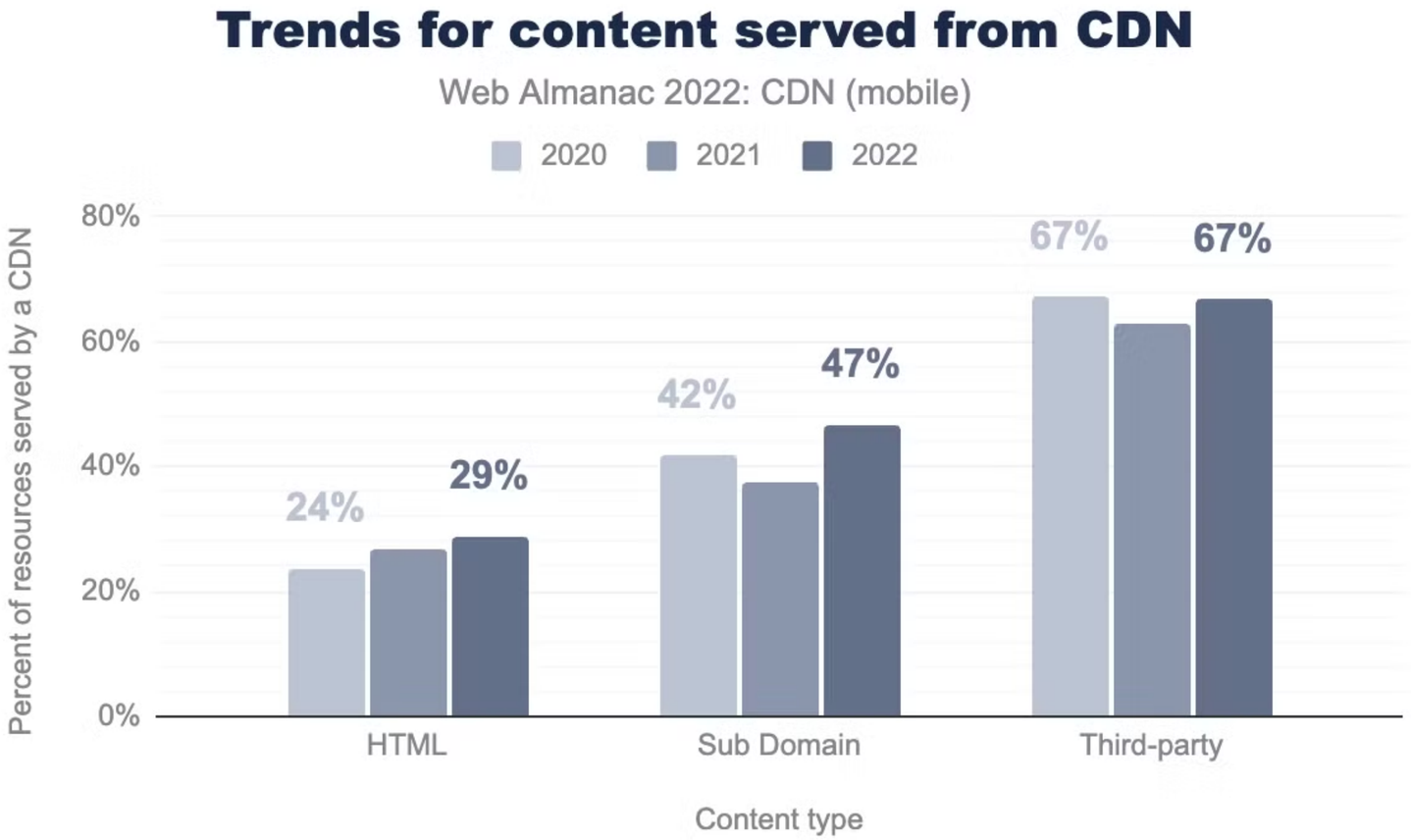
- Which CDNs are used by the top 1000 websites?
- And how often?
- What percentage of the total delivered web site assets are served by CDNs?

Inspiration & orientation

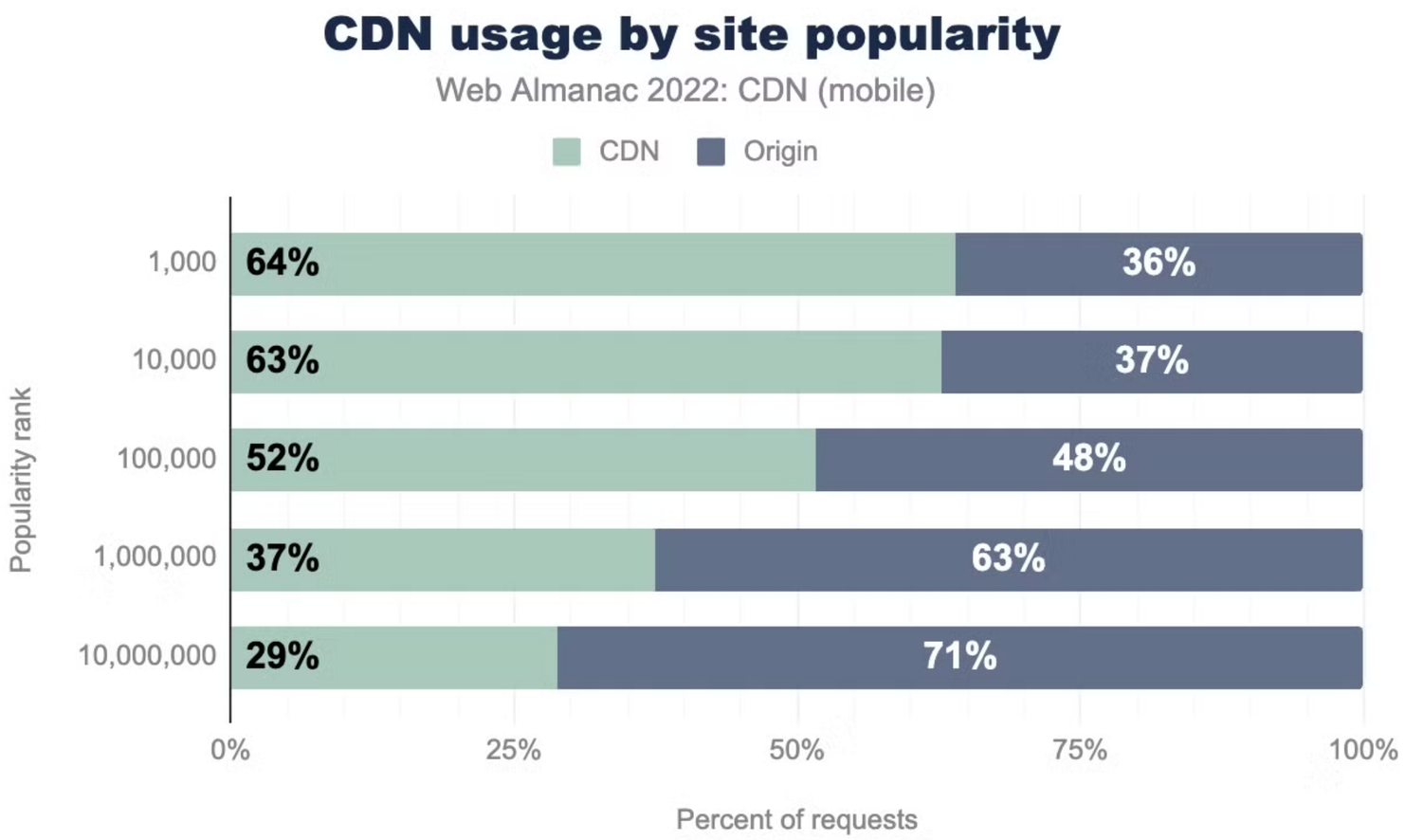
Web Almanac

By HTTP Archive

Web Almanac 2022 [1]



Web Almanac 2022 [1]



Web Almanac 2022 [1]

ABOUT THE DATASET

Choosing the Dataset

What are the ”top 1000” websites?

- Alexa top 1000 most visited websites (discontinued)
- Top 1000 ranked pages from SEO companies
- Top 1000 domains from a DNS server provider
- Chrome UX Report (CrUX)

Going with CrUX

”Overall, we find that top lists (including Alexa) capture the set of top websites relatively poorly across all of our metrics with one exception: Google’s Chrome User Experience Report (CrUX), which has recently begun publishing [...] the most popular websites as seen by Google Chrome.”

Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists

Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric

2022

[2]

METHODOLOGY

Methodology

Recording all requests

Recording all network requests and responses when iterating through the CrUX dataset.

Selenium Web Driver

Headless Chrome

Identifying CDNs

Identifying CDNs using the response headers, the URL, the CNAME and the autonomous system number (ASN).

Python script

Analysis

Use the collected data and aggregate it to create plots that answer the research questions.

Jupyter Notebook

Pandas

Seaborn

Methodology

How to get all assets?

Inspecting the HTML file and all its assets

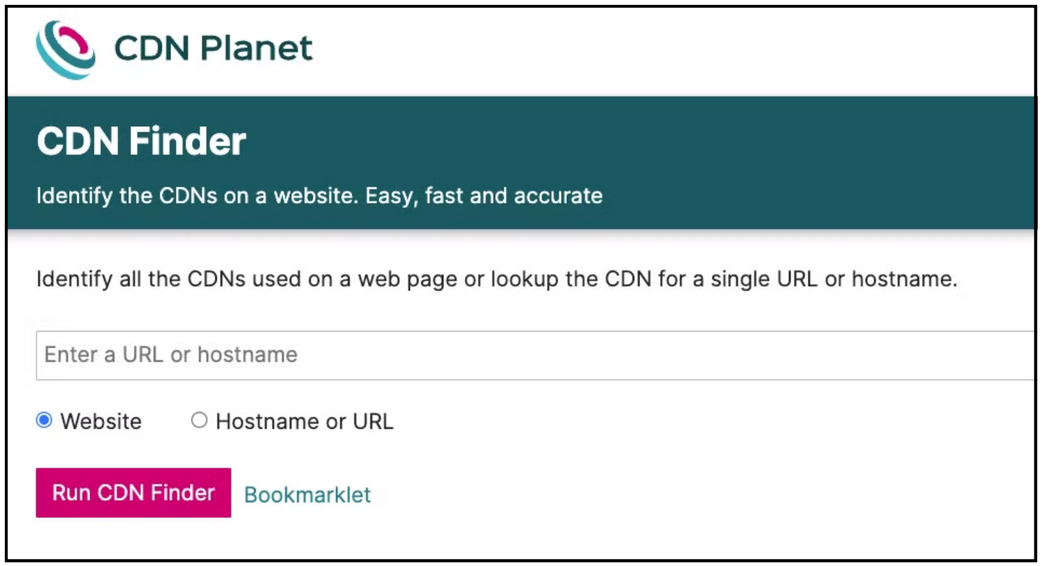
- Fast and easy
- No dynamic content,
  only in HTML embedded content
- No API requests, no fetch calls

Using a headless browser

- Supports loading and execution of Javascript
- Simulates real website visit
- Slower

Methodology

Identifying CDNs



https://www.cdnplanet.com/tools/cdnfinder

- via response headers
- via URL
- via CNAME
- via ASN

Using the Response Headers

https://st.deviantart.net/eclipse/browser-support.min.js?20231214

```
{'access-control-allow-origin': '*',
 'age': '16887940',
 'cache-control': 'max-age=31536000', #
 'content-encoding': 'gzip',
 'content-type': 'application/javascript',
 'date': 'Thu, 14 Dec 2023 09:22:17 GMT',
 'expires': 'Fri, 13 Dec 2024 09:22:17 GMT',
 'last-modified': 'Thu, 14 Dec 2023 07:36:22 GMT',
 'server': 'nginx', 'via': '1.1 badff53d2116a4b3d32a2dd1eb918a48.cloudfront.net (CloudFront)',
 'x-amz-cf-id': 'ygklJjELs0D16EZCZlaSRW4VA5WmoPLDrTaBnWDZMM8iXB2Kcd4ecw==',
 'x-amz-cf-pop': 'MUC50-P1',
 'x-cache': 'Hit from cloudfront'}
```

Using the URL

https://community.akamai.steamstatic.com/public/shared/images/header/logo_steam.svg?t=962016

STEAM®

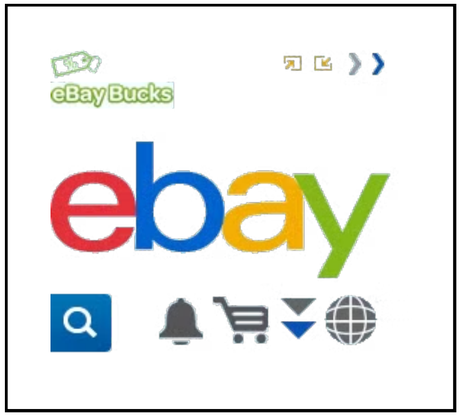
Using the CNAME

https://ir.ebaystatic.com/rs/v/fxxj3ttftm5ltcqnto1o4baovyl.png

↓

```
nslookup ir.ebaystatic.com
Server:     192.168.178.1
Address:    192.168.178.1#53

Non-authoritative answer:
ir.ebaystatic.com   canonical name = ir.ebaycdn.net.
ir.ebaycdn.net  canonical name = ebaystatic.ebay.map.fastly.net.
Name:   ebaystatic.ebay.map.fastly.net
Address: 199.232.190.206
```



Using the ASN

https://apkpure.com/favicon.ico

```
nslookup apkpure.com
Server:     192.168.178.1
Address:    192.168.178.1#53

Non-authoritative answer:
Name:   apkpure.com
Address: 104.22.5.119
Name:   apkpure.com
Address: 172.67.8.127
Name:   apkpure.com
Address: 104.22.4.119
```

```
curl 127.0.0.1:80/v1/as/ip/104.22.5.119
{
  "announced": true,
  "as_country_code": "US",
  "as_description": "CLOUDFLARENET",
  "as_number": 13335,
  "first_ip": "104.16.0.0",
  "ip": "104.22.5.119",
  "last_ip": "104.22.79.255"
}
```

Methodology

Analysis

Imports

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import tldextract
import numpy as np
```

Reading the collected data

In [2]:
```
# read the csv file
df = pd.read_csv('requests.csv')

# show 10 random entries
df.sample(10)
```

Out[2]:

	Page	Full URL	Subdomain	Domain	Suffix	IP	CNAME	ASN	ASN Description	Response Headers
42350	https://dantri.com.vn	https://fonts.googleapis.com/css?family=Archiv...	fonts	googleapis	com	142.251.36.228	NaN	15169.0	GOOGLE	{'access-control-allow-origin': '*', 'alt-svc'...
51486	https://as.com	https://pagead2.googlesyndication.com/pagead/g...	pagead2	googlesyndication	com	142.251.36.164	NaN	15169.0	GOOGLE	{'alt-svc': 'h3=":443"; ma=2592000,h3-29=":443...
35961	https://www.amazon.com.tr	https://m.media-amazon.com/images/I/61AES6+pEG...	m	media-amazon	com	NaN	c.media-amazon.com	NaN	NaN	{'accept-ranges': 'bytes', 'access-control-all...
12396	https://www.fmkorea.com	https://static.fmkorea.com/classes/lazy/js/scr...	static	fmkorea	com	93.184.223.182	a760.w39.akamai.net	15133.0	EDGECAST	{'access-control-allow-origin': '*', 'cache-co...
45439	https://jp.pornhub.com	https://ei.phncdn.com/videos/202212/08/4210515...	ei	phncdn	com	NaN	ei.phncdn.com.sds.rncdn7.com	NaN	NaN	{'access-control-allow-origin': '*', 'cache-co...
19880	https://www.iltalehti.fi	https://img.ilcdn.fi/_rKIIZP6x0KTZpZwv50DHTeg6...	img	ilcdn	fi	NaN	d3cdjjarcvj45p.cloudfront.net	NaN	NaN	{'age': '4175', 'cache-control': 'max-age=3153...
40638	https://snaptik.app	https://adsystem.pocpoc.io/js/v1/adtag.js	adsystem	pocpoc	io	104.26.14.167	NaN	13335.0	CLOUDFLARENET	{'age': '45', 'alt-svc': 'h3=":443"; ma=86400'...
40354	https://www.elnacional.cat	https://www.tradingview-widget.com/static/bund...	www	tradingview-widget	com	NaN	tradingview-widget.b-cdn.net	NaN	NaN	{'access-control-allow-origin': '*', 'cache-co...
48726	https://www.abozeb.com	https://www.abozeb.com/	www	abozeb	com	104.21.83.142	NaN	13335.0	CLOUDFLARENET	{'age': '134573', 'alt-svc': 'h3=":443"; ma=86...
25769	https://www.xvideos91.com	https://www.xvideos91.com/static-files/v-02405...	www	xvideos91	com	185.88.181.3	www.xvideos.com.cdn.cloudflare.net	46652.0	SERVERSTACK-ASN	{'accept-ranges': 'bytes', 'access-control-all...

Own analysis [3]

Imports

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import tldextract
import numpy as np
```

Reading the collected data

In [2]:
```
# read the csv file
df = pd.read_csv('requests.csv')

# show 10 random entries
df.sample(10)
```

Out[2]:

	Page	Full URL	Subdomain	Domain	Suffix	IP	CNAME	ASN	ASN Description	
42350	https://dantri.com.vn	https://fonts.googleapis.com/css?family=Archiv...	fonts	googleapis	com	142.251.36.228	NaN	15169.0	GOOGLE	
51486	https://as.com	https://pagead2.googlesyndication.com/pagead/g...	pagead2	googlesyndication	com	142.251.36.164	NaN	15169.0	GOOGLE	{'alt-s
35961	https://www.amazon.com.tr	https://m.media-amazon.com/images/I/61AES6+pEG...	m	media-amazon	com	NaN	c.media-amazon.com	NaN	NaN	{
12396	https://www.fmkorea.com	https://static.fmkorea.com/classes/lazy/js/scr...	static	fmkorea	com	93.184.223.182	a760.w39.akamai.net	15133.0	EDGECAST	{
45439	https://jp.pornhub.com	https://ei.phncdn.com/videos/202212/08/4210515...	ei	phncdn	com	NaN	ei.phncdn.com.sds.rncdn7.com	NaN	NaN	{
19880	https://www.iltalehti.fi	https://img.ilcdn.fi/_rKIIZP6x0KTZpZwv50DHTeg6...	img	ilcdn	fi	NaN	d3cdjjarcvj45p.cloudfront.net	NaN	NaN	{'a
40638	https://snaptik.app	https://adsystem.pocpoc.io/js/v1/adtag.js	adsystem	pocpoc	io	104.26.14.167	NaN	13335.0	CLOUDFLARENET	{
40354	https://www.elnacional.cat	https://www.tradingview-widget.com/static/bund...	www	tradingview-widget	com	NaN	tradingview-widget.b-cdn.net	NaN	NaN	{
48726	https://www.abozeb.com	https://www.abozeb.com/	www	abozeb	com	104.21.83.142	NaN	13335.0	CLOUDFLARENET	{'
25769	https://www.xvideos91.com	https://www.xvideos91.com/static-files/v-02405...	www	xvideos91	com	185.88.181.3	www.xvideos.com.cdn.cloudflare.net	46652.0	SERVERSTACK-ASN	{

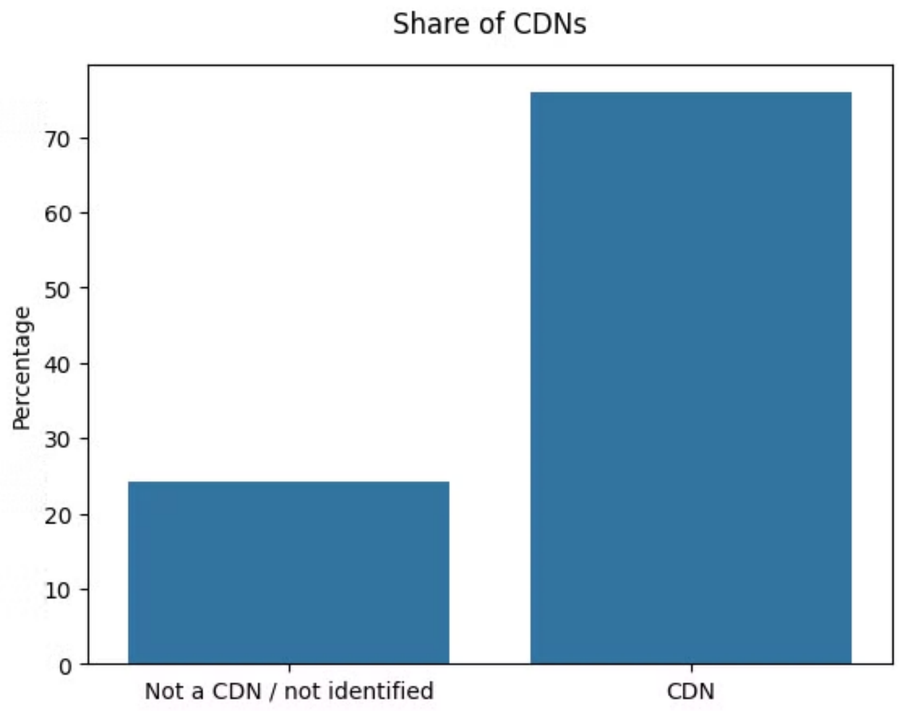
Own analysis [3]

Full URL	Subdomain	Domain	Suffix	IP	CNAME	ASN	ASN Description	Response Headers
googleapis.com/css?family=Archiv...	fonts	googleapis	com	142.251.36.228	NaN	15169.0	GOOGLE	{'access-control-allow-origin': '*', 'alt-svc'...
.googlesyndication.com/pagead/g...	pagead2	googlesyndication	com	142.251.36.164	NaN	15169.0	GOOGLE	{'alt-svc': 'h3=":443"; ma=2592000,h3-29=":443...
mazon.com/images/I/61AES6+pEG...	m	media-amazon	com	NaN	c.media-amazon.com	NaN	NaN	{'accept-ranges': 'bytes', 'access-control-all...
ic.fmkorea.com/classes/lazy/js/scr...	static	fmkorea	com	93.184.223.182	a760.w39.akamai.net	15133.0	EDGECAST	{'access-control-allow-origin': '*', 'cache-co...
n.com/videos/202212/08/4210515...	ei	phncdn	com	NaN	ei.phncdn.com.sds.rncdn7.com	NaN	NaN	{'access-control-allow-origin': '*', 'cache-co...
.fi/_rKIIZP6x0KTZpZwv50DHTeg6...	img	ilcdn	fi	NaN	d3cdjjarcvj45p.cloudfront.net	NaN	NaN	{'age': '4175', 'cache-control': 'max-age=3153...
://adsystem.pocpoc.io/js/v1/adtag.js	adsystem	pocpoc	io	104.26.14.167	NaN	13335.0	CLOUDFLARENET	{'age': '45', 'alt-svc': 'h3=":443"; ma=86400'...
adingview-widget.com/static/bund...	www	tradingview-widget	com	NaN	tradingview-widget.b-cdn.net	NaN	NaN	{'access-control-allow-origin': '*', 'cache-co...
https://www.abozeb.com/	www	abozeb	com	104.21.83.142	NaN	13335.0	CLOUDFLARENET	{'age': '134573', 'alt-svc': 'h3=":443"; ma=86...
videos91.com/static-files/v-02405...	www	xvideos91	com	185.88.181.3	www.xvideos.com.cdn.cloudflare.net	46652.0	SERVERSTACK-ASN	{'accept-ranges': 'bytes', 'access-control-all...

Own analysis [3]

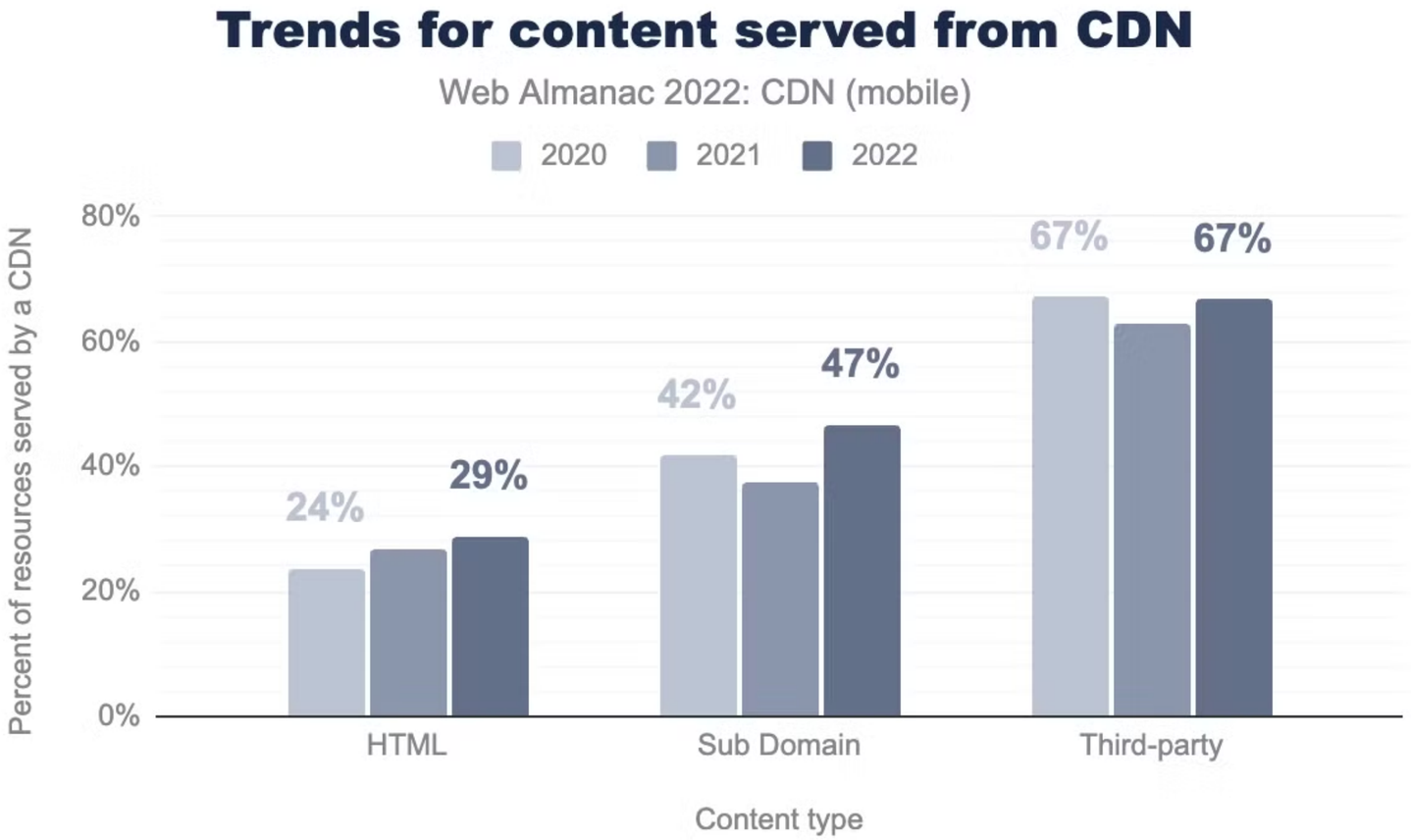
FINDINGS

Findings

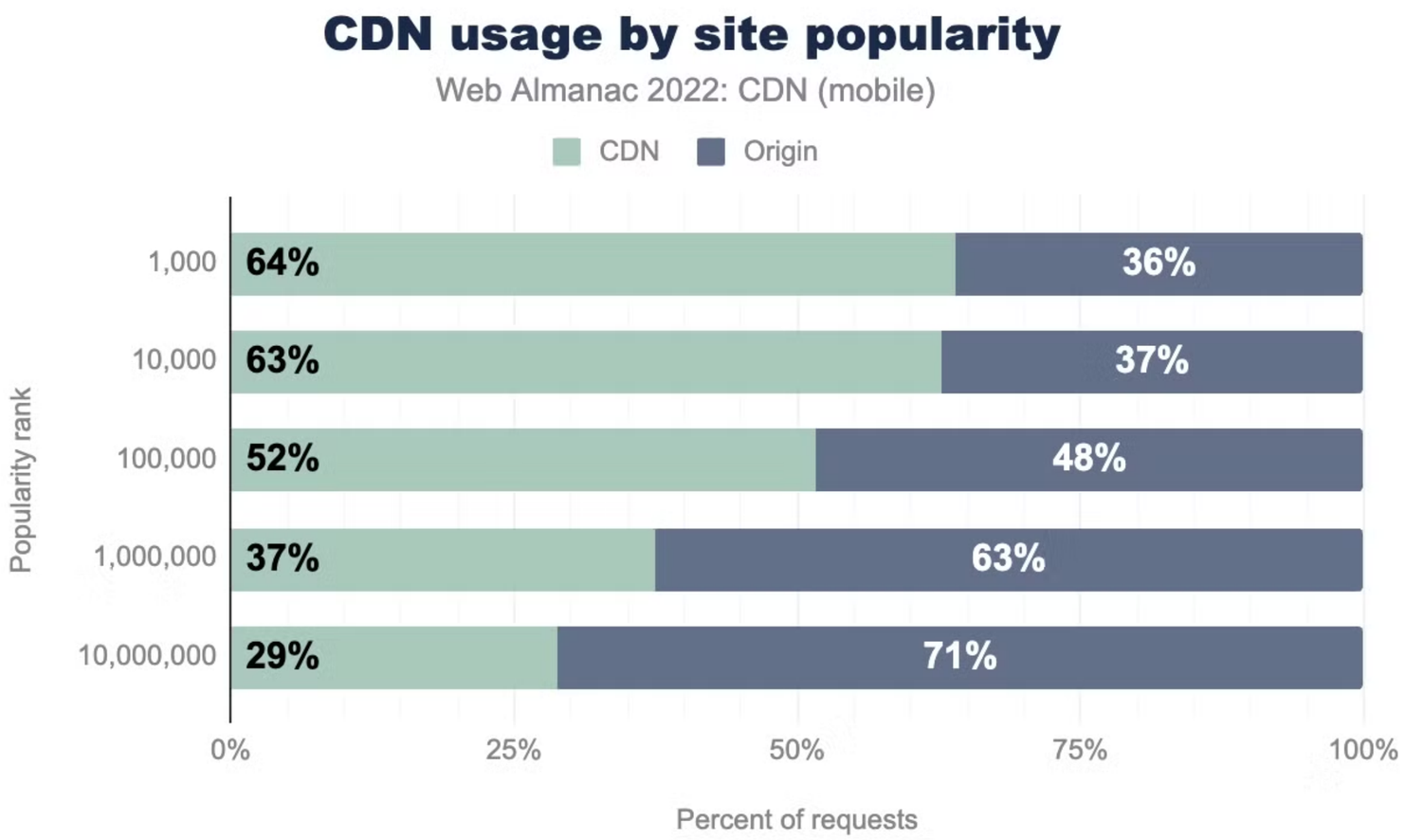


About 75 % of all requests served by CDNs

Own analysis [3]

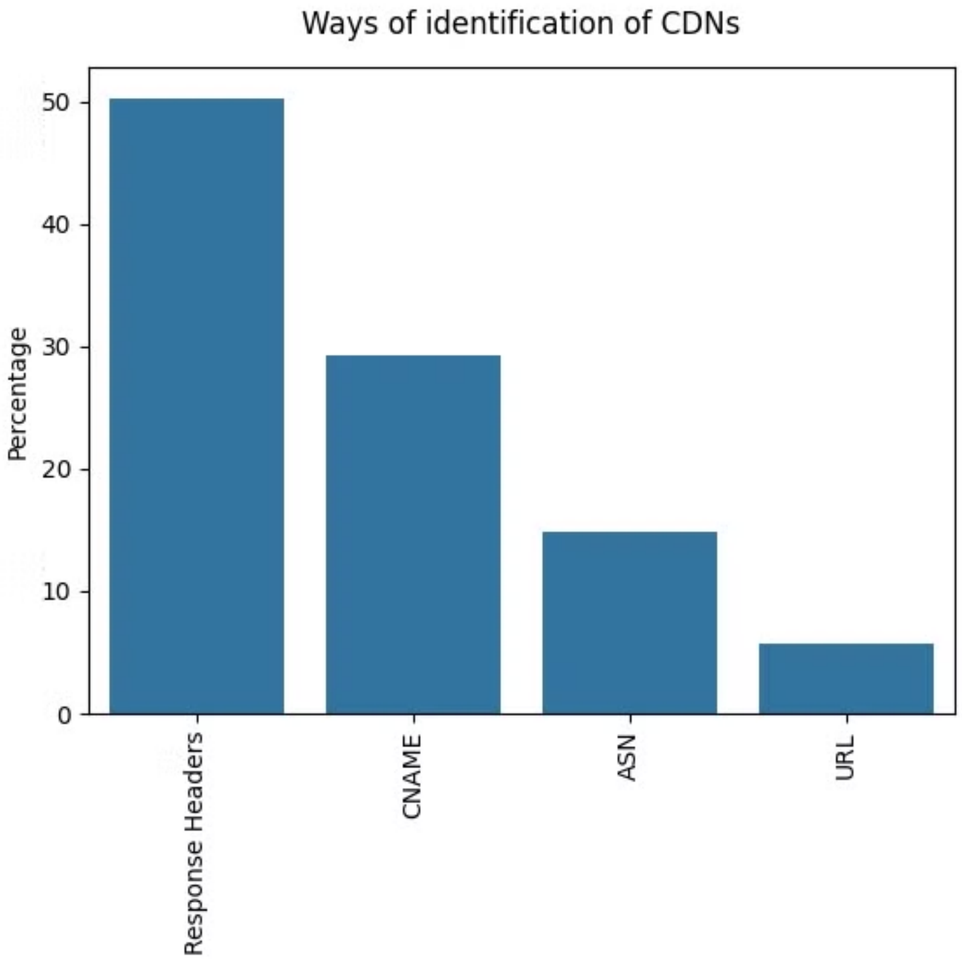


Web Almanac 2022 [1]



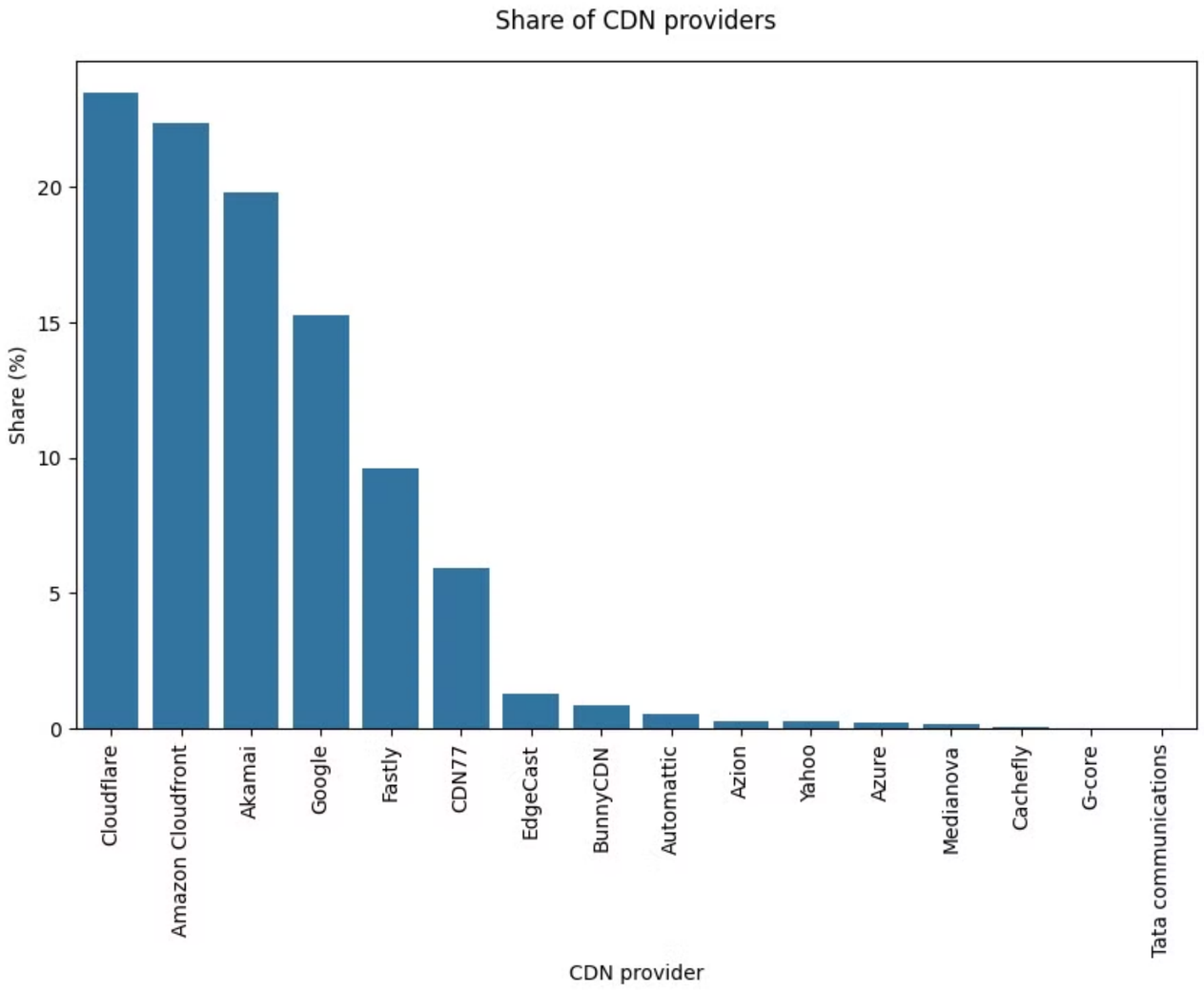
Web Almanac 2022 [1]

Findings



Own analysis [3]

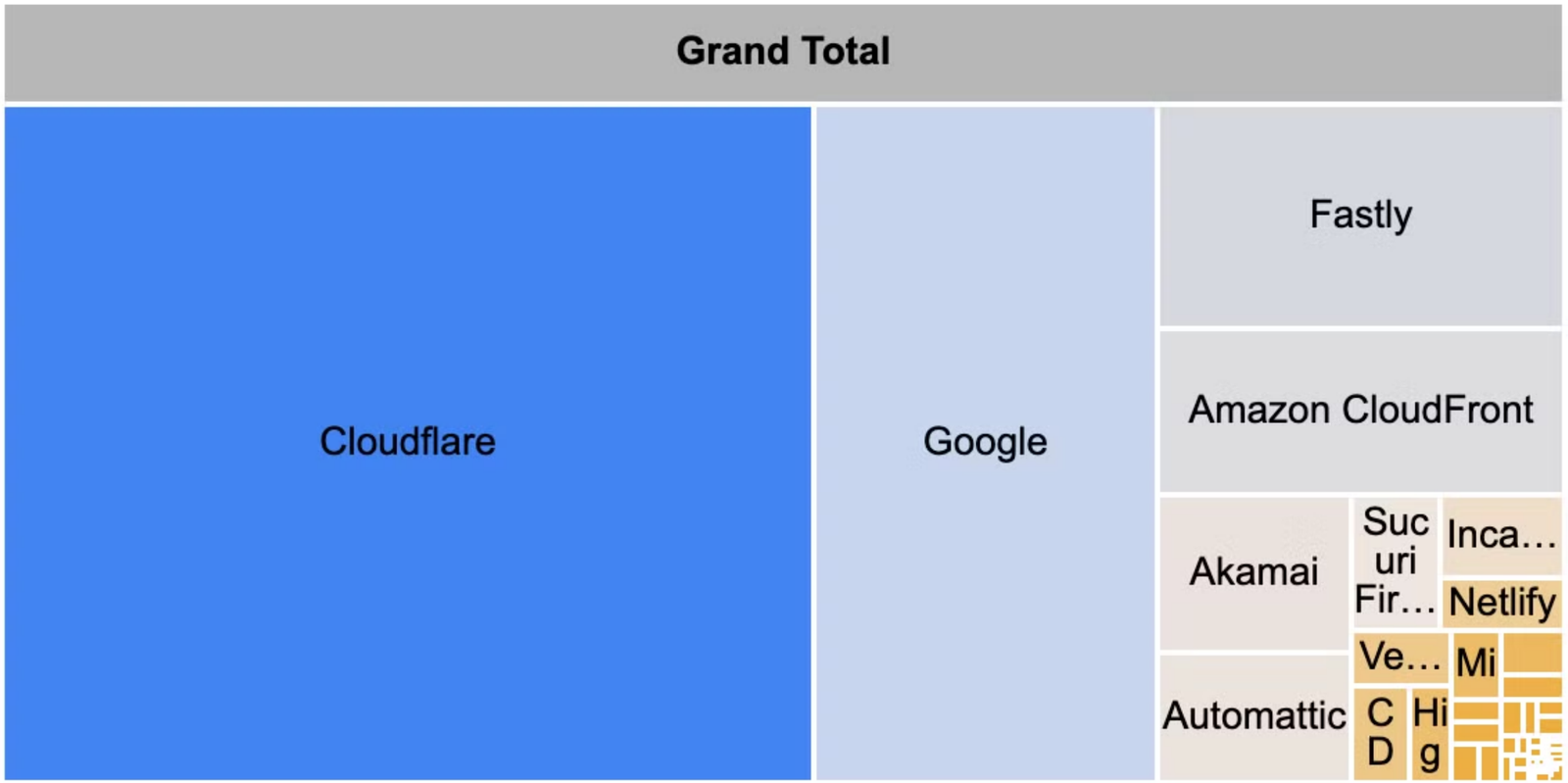
Findings



Own analysis [3]

Top CDNs for HTML requests

Web Almanac 2022: CDN (mobile)



Hannes Koksch (hk058)

CONCLUSION AND CRITICAL REVIEW

Conclusion

- Identified ”big players” in the field
- Confirming continuous trend towards CDNs
- Gained understanding of data collection, DNS and CNAME

Critical review

- The results are only an approximation of the real situation
- Not all CDNs may have been detected
- Not every CDN may have been detected correctly
- Some requests may have been misidentified as a CDN
- Some site use multiple CDNs
- Technical struggles

THANK YOU

Sources

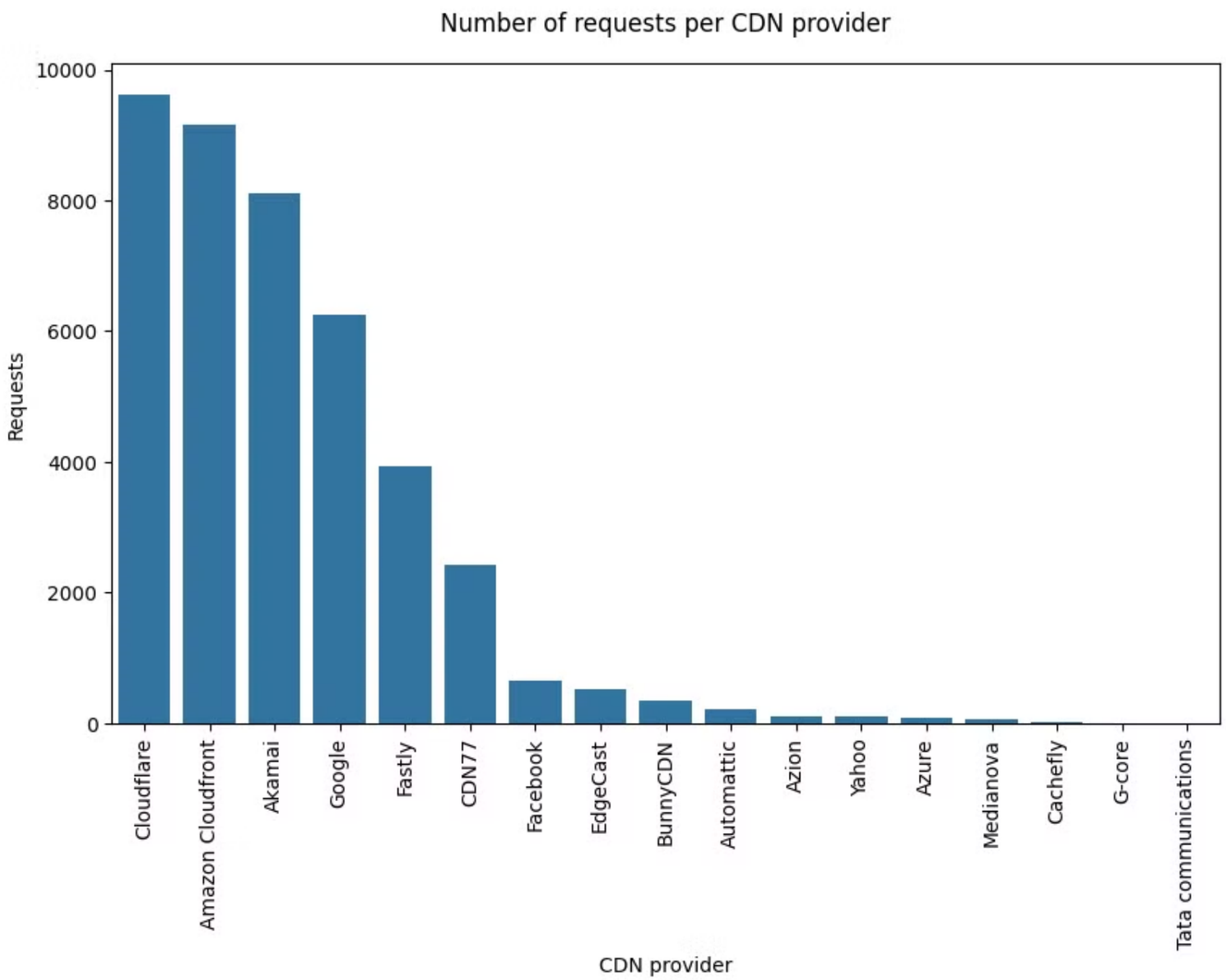
[1]
HTTP Archive. Web Almanac 2022. Chapter 22
https://almanac.httparchive.org/en/2022/cdn

[2]
Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric.
Toppling top lists: evaluating the accuracy of popular website
lists. Proceedings of the 22nd ACM Internet Measurement Conference, 2022.

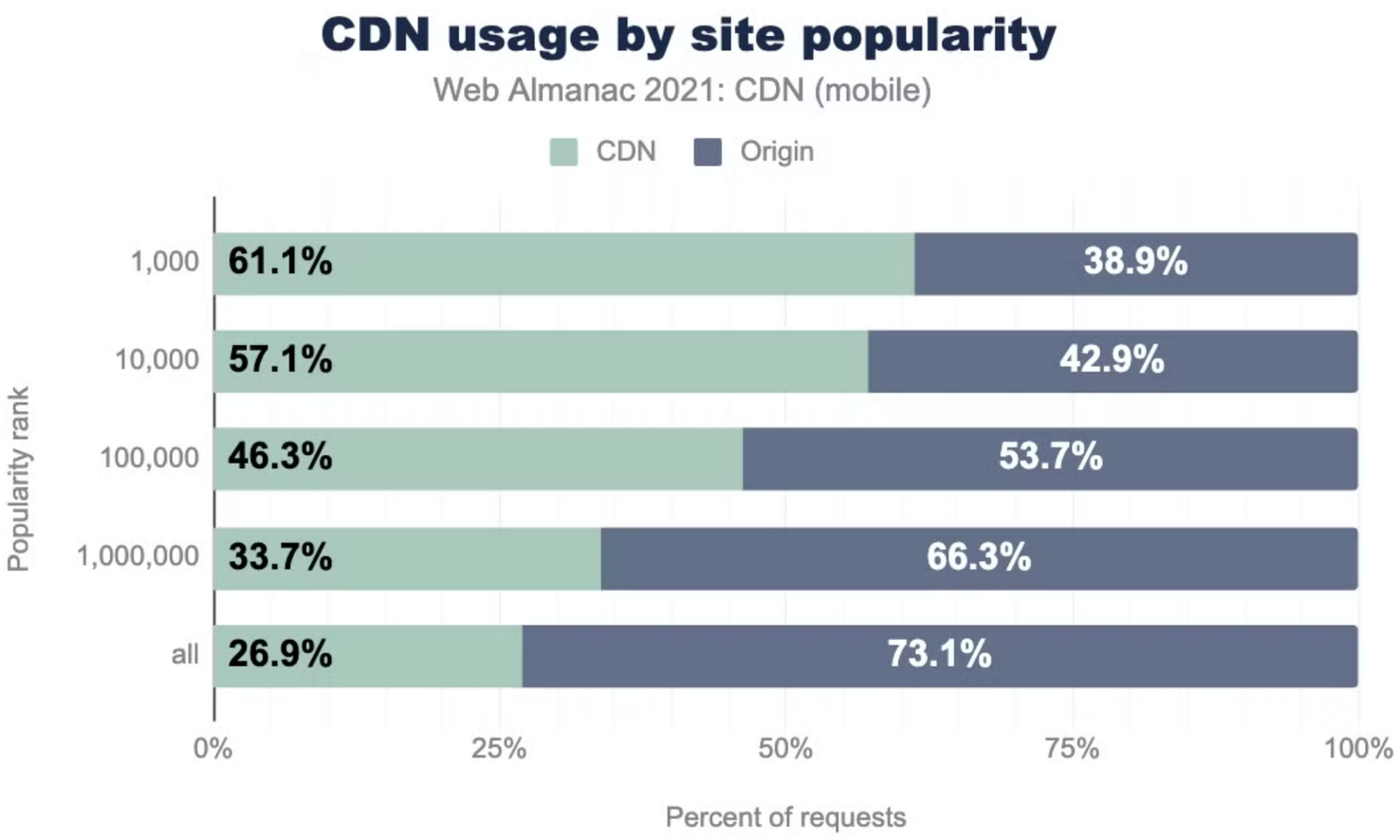
[3]
Own analysis on GitHub. Github/hanneskoksch/Analysis-of-CDNs
https://github.com/hanneskoksch/Analysis-of-CDNs

All links in this presentation were last checked on 27.06.2024

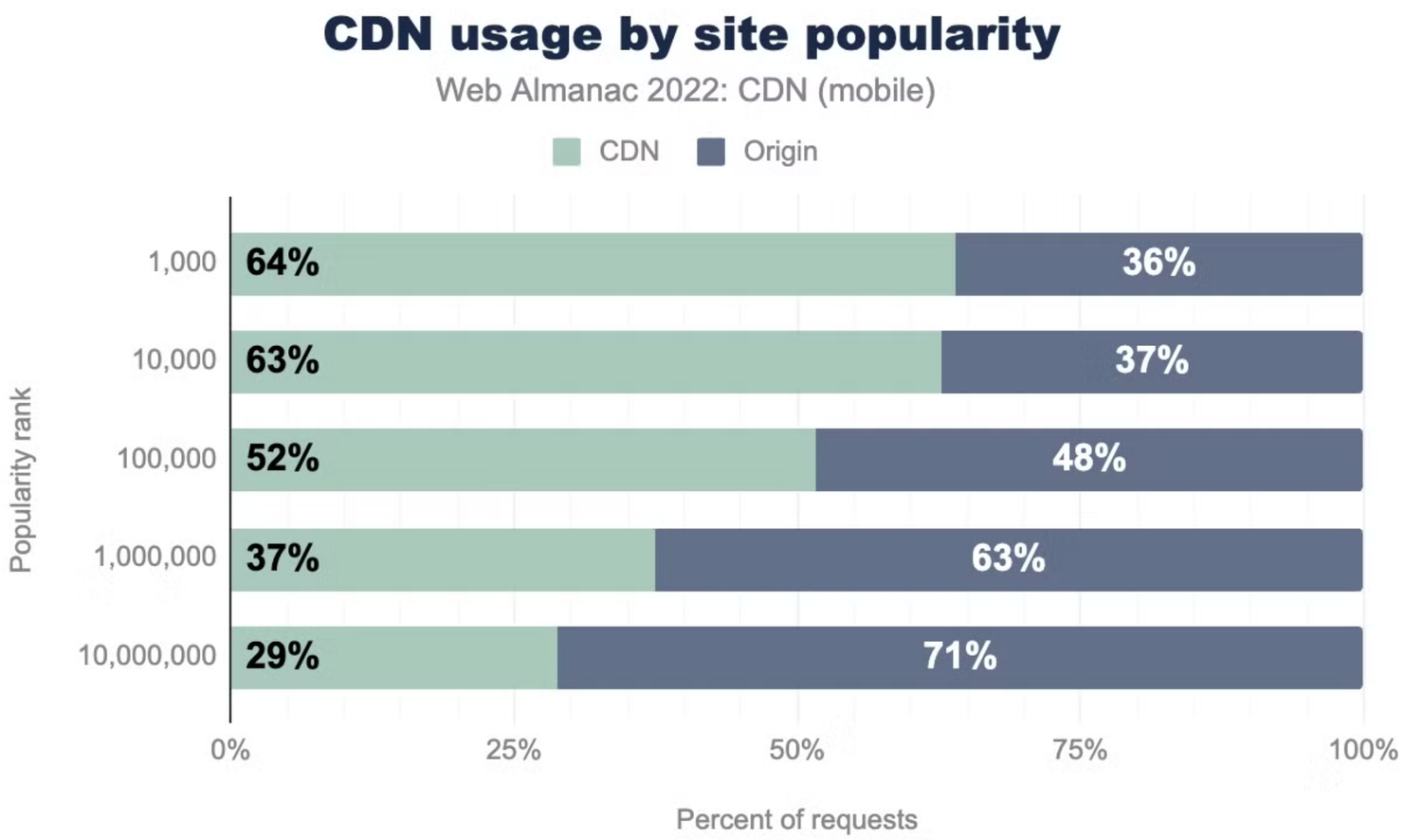
Findings



Own analysis [3]



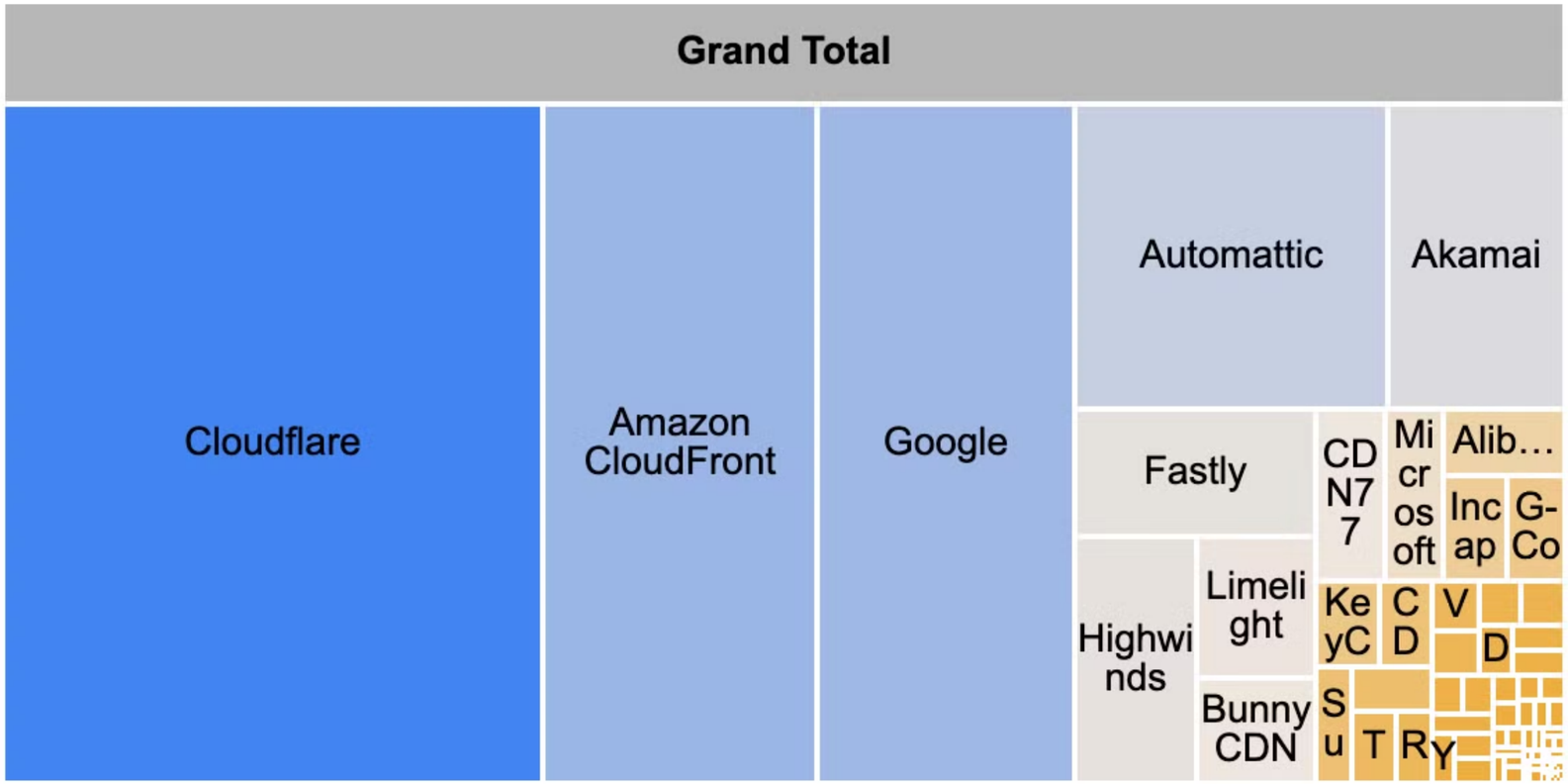
Web Almanac 2021



Web Almanac 2022 [1]

Top CDNs for subdomain requests

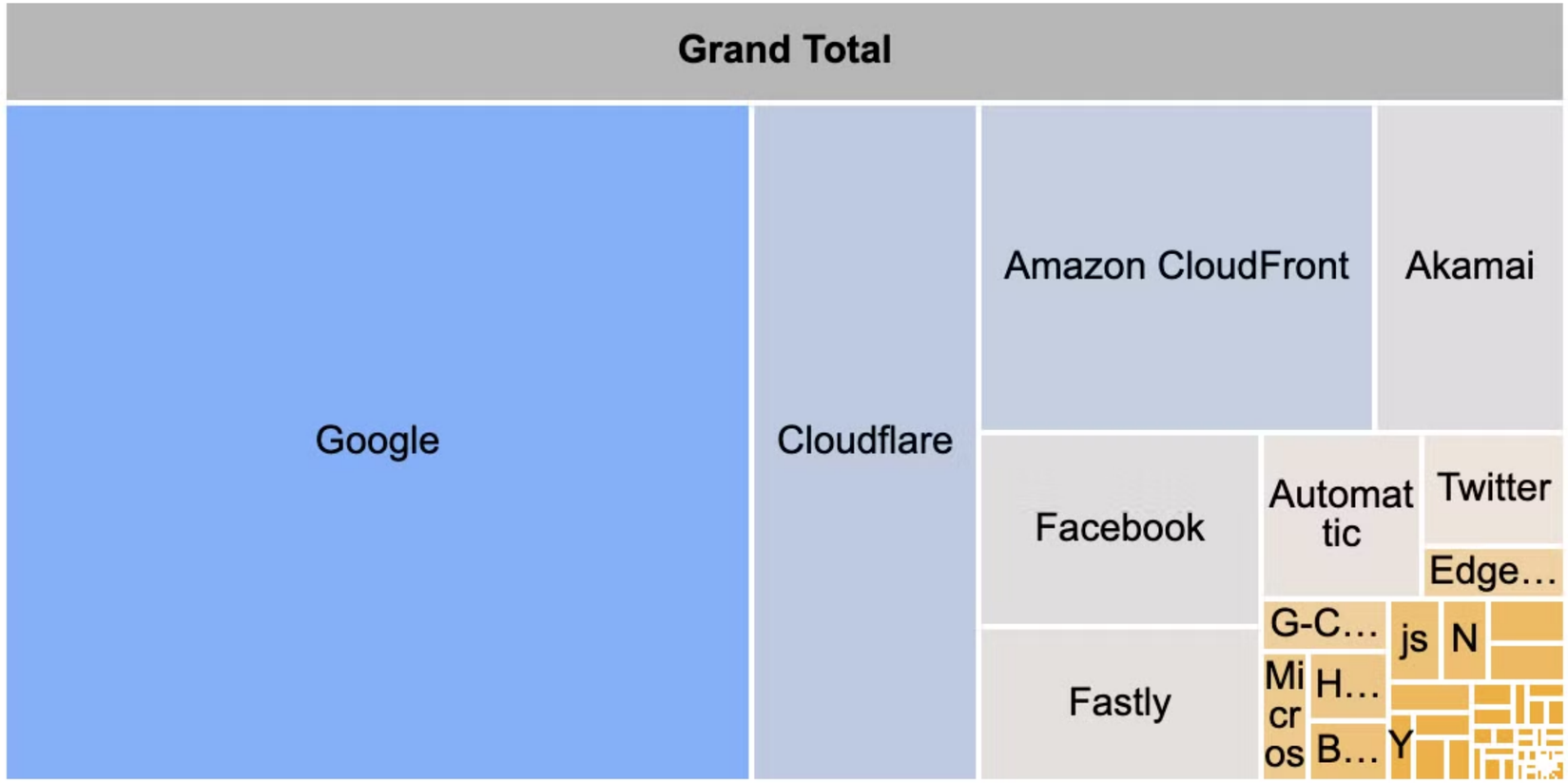
Web Almanac 2022: CDN (mobile)



Web Almanac 2022 [1]

Top CDNs for third-party requests

Web Almanac 2022: CDN (mobile)



Web Almanac 2022 [1]