Common Crawl DNSSEC Analysis

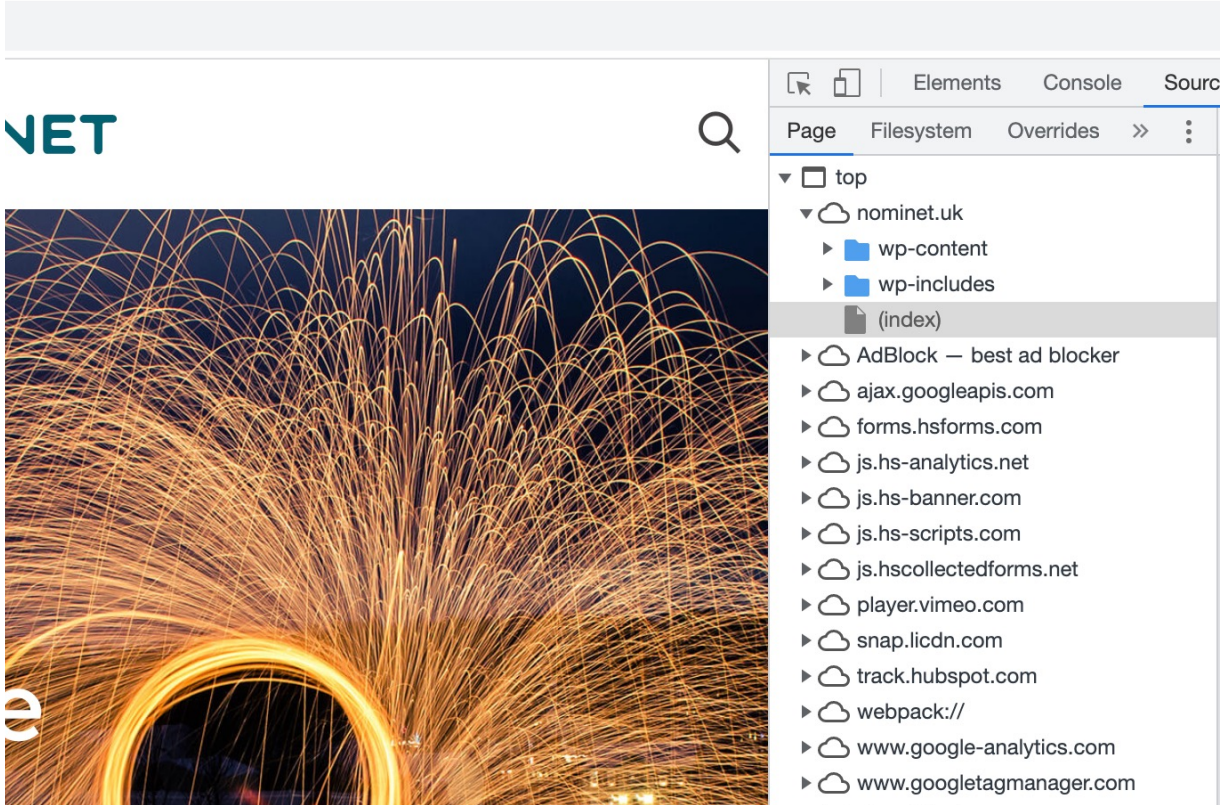
James Richards, Researcher

Nominet

What are we studying?

- Many DNSSEC deployment studies focus on lists of domains such as those occurring in popular lists or TLDs.

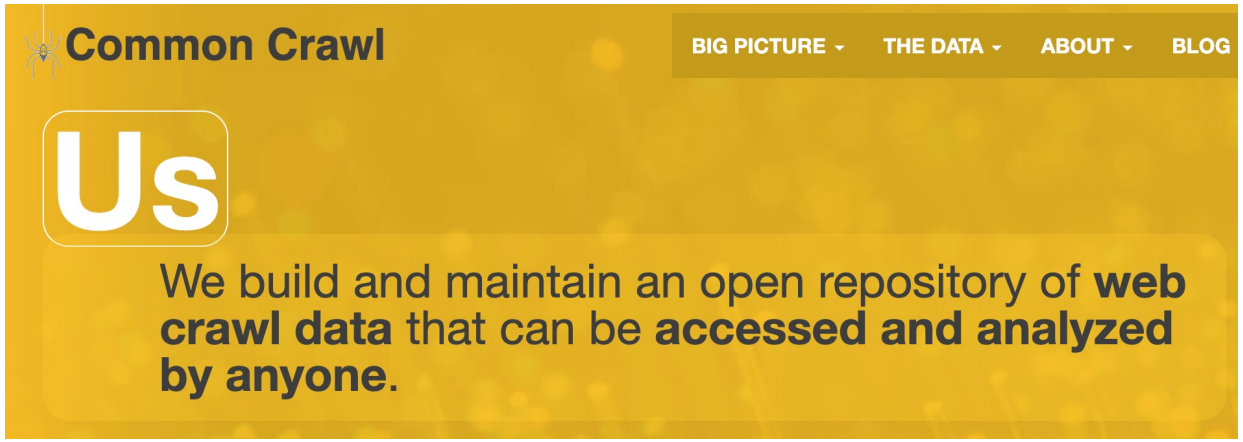
- Users are rarely delivered website content from one single domain name – often multiple domains are utilized



- How do the DNSSSEC prevalence statistics change when all domains used to load content are considered?

Dataset

- Common Crawl
- Amazon Public Datasets program
- July/August 2021 dataset




```
<!doctype html public "-//W3C//DTD HTML 4.0 Tran
<html>
<head>
<title>
        BBC NEWS | Africa | Namibia braces for N
</title>

<meta name="keywords" content="BBC, News, BBC Ne
<meta name="OriginalPublicationDate" content="20
```

WARC - Web Archive Format

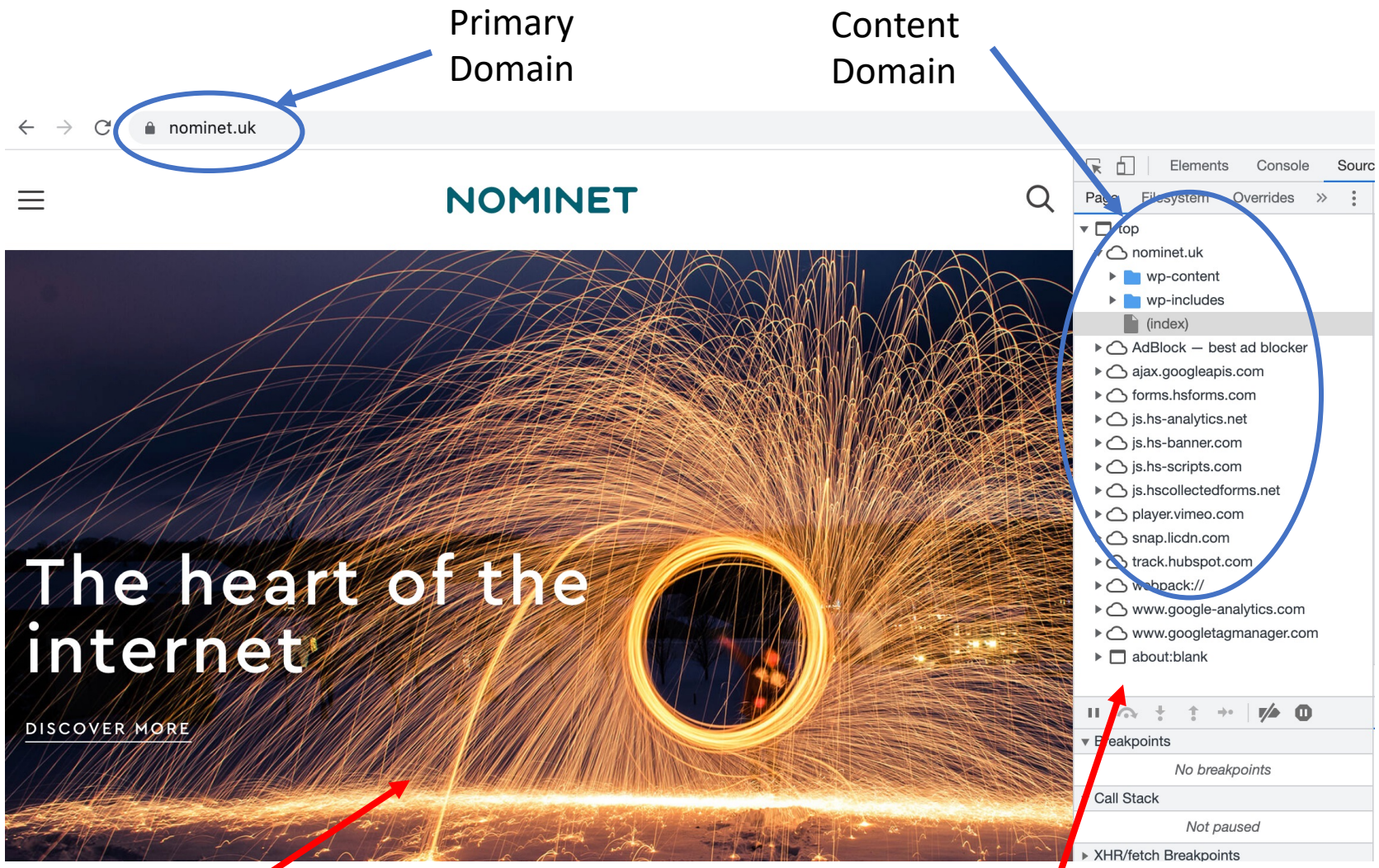
```
"HTML-Metadata" : {
   "Links" : [
      {
         "href" : "/css/screen/shared/styles.css",
         "path" : "STYLE/#text"
      },
      {
         "href" : "/css/screen/shared/toolbar_ifs.css",
         "path" : "STYLE/#text"
```

WAT - WARC computed Metadata

Analysis

We want to analyse domains used to deliver website content.

This requires some judgement



Hyperlink
[not content]

Image
[content]

Script
[content]

Analysis

Grouped into categories using the Mozilla MDN elements reference, for example:

- Content sectioning (article, h1, h2, header, …)
- Image and multimedia (video, img, audio, ….)
- Scripting (script, canvas, …)
- …
- Tags not falling into an MDN category were assigned manually or given the label *esoteric*

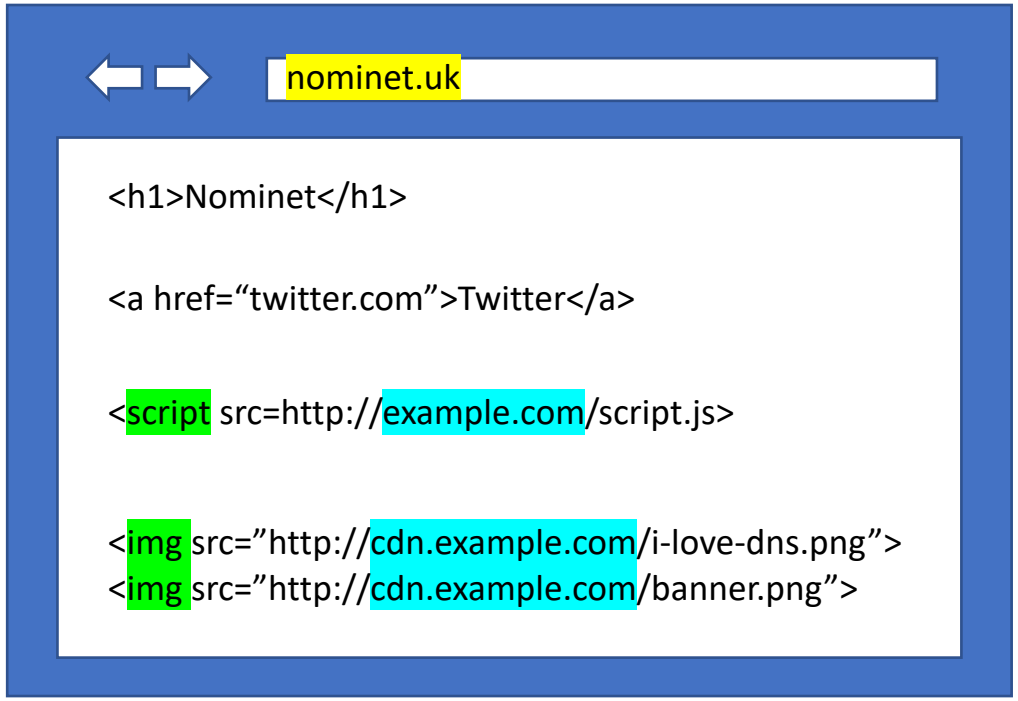
Odd stuff:

- Websites with very high number of unique domains in the page
- Content that doesn’t resolve (NxDomain, ServFail, Timeout)
- Sometimes difference between what is rendered in browser and the web scrape – a challenge we must live with for this study

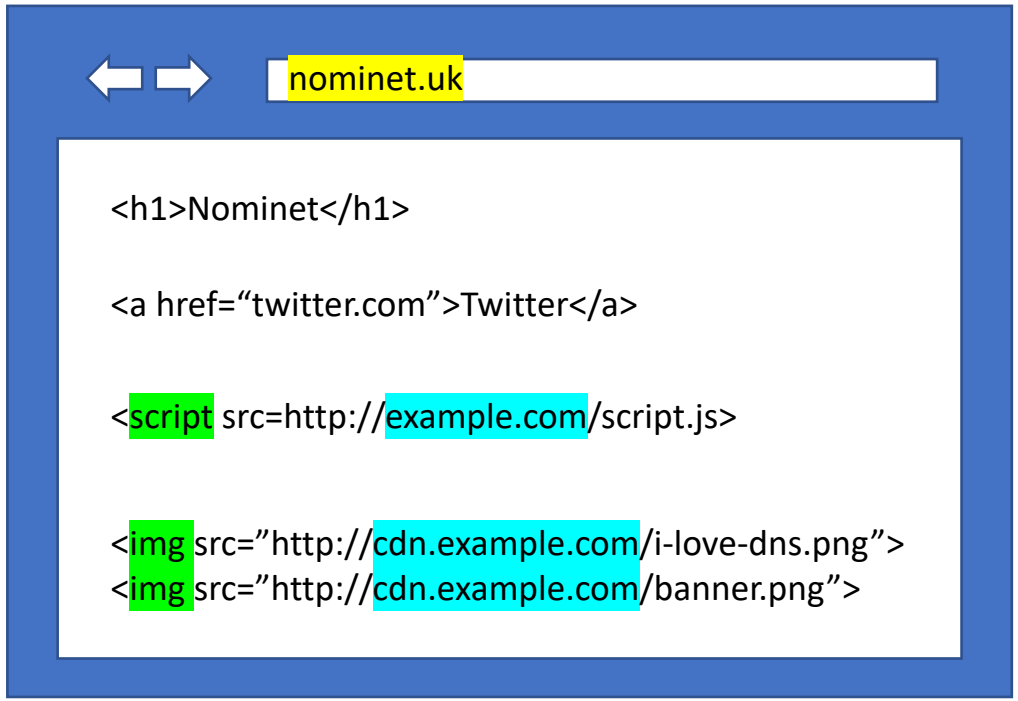
Data Collection

72,000 WAT files (21.67 TB) processed in batches in parallel:

1. Download a file and iterate through it with *warcio*
2. Select domains under an ICANN suffix: nominet.uk
3. Select URLs at the root of a website: nominet.uk/
4. Lookup A record for primary domain with DNSSEC OK bit set
5. Extract domain names and tags from content in website
6. Lookup A record for content domain with DNSSEC OK bit set
7. Throw away data that isn't representative of content (a, area, p, mark, h, div, …)
8. Make assumption that tags without domains are relatively referenced from primary domain e.g. <img src="/picture.jpg">



Taking the right measurements



domain	domain_a	domain_rrsig	domain_algo	element	class	content_domain	count	content_domain_a	content_domain_rrsig	content_domain_algo
nominet.uk	Yes	Yes	13	SCRIPT@/src	Scripting	example.com	1	Yes	No	None
nominet.uk	Yes	Yes	13	IMG@/src	Image and multimedia	cdn.example.com	2	Yes	No	None

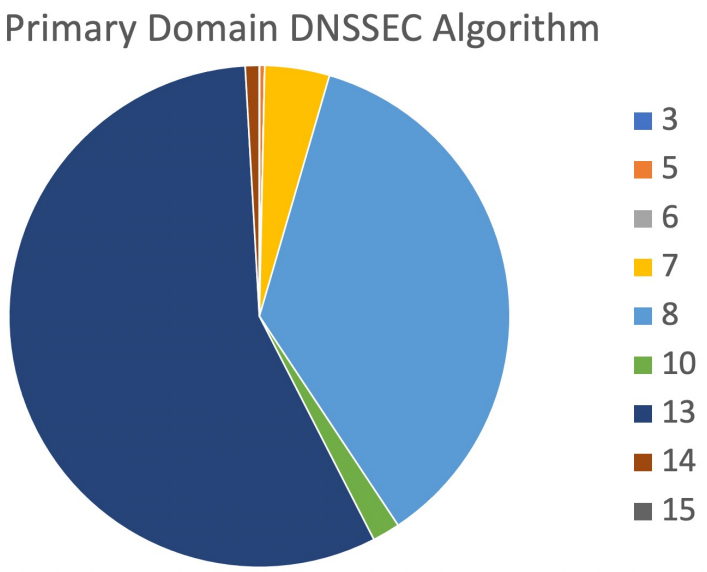
High Level Statistics

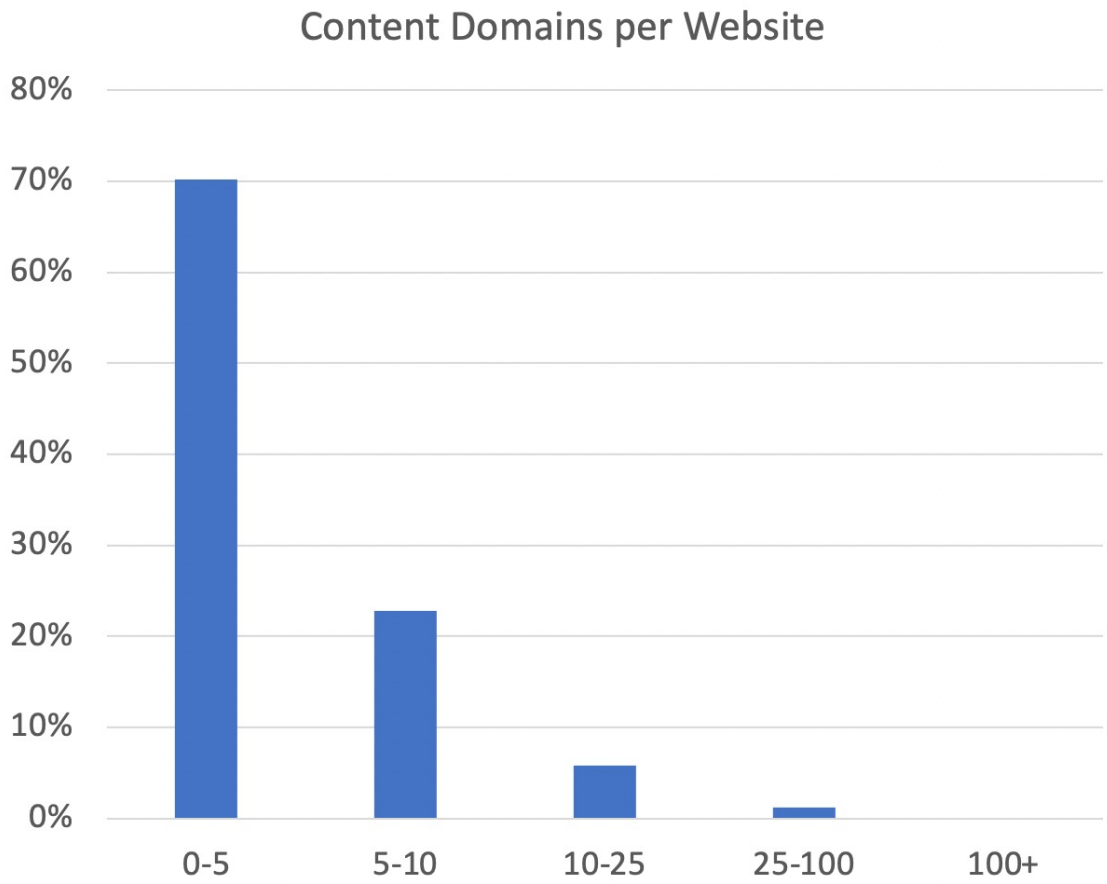
13.5 million websites analysed across 1113 different TLDs:

com / org / net	58%
country code TLDs	37%
gTLDs / others	5%

Contained **16 million** content domains

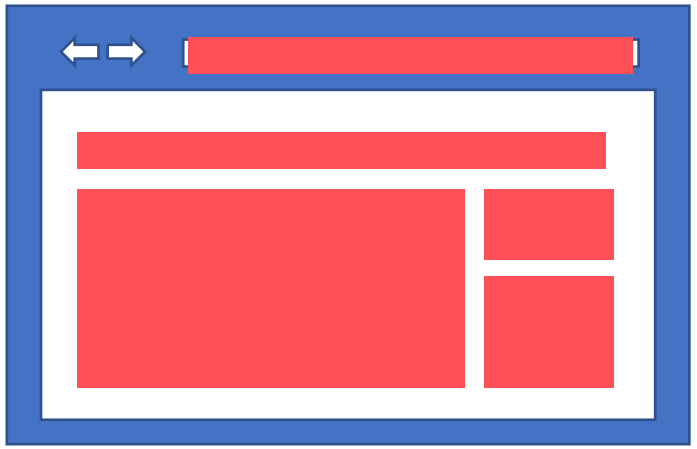
7.7% of the primary domains were signed and most used either algorithm **8 or 13**



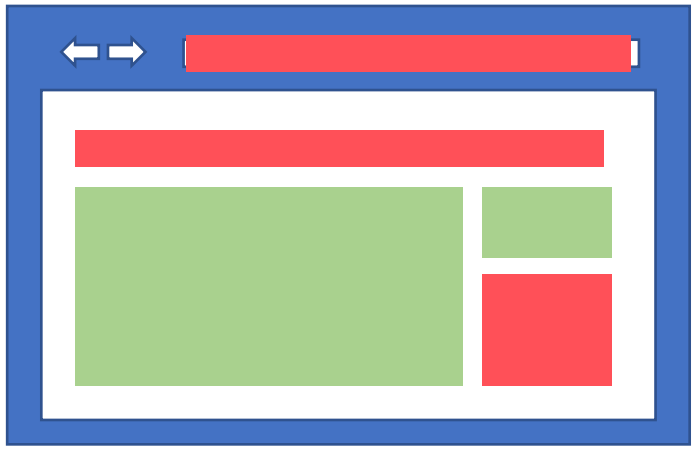


Analysis

Unsigned primary domains (92.3%)

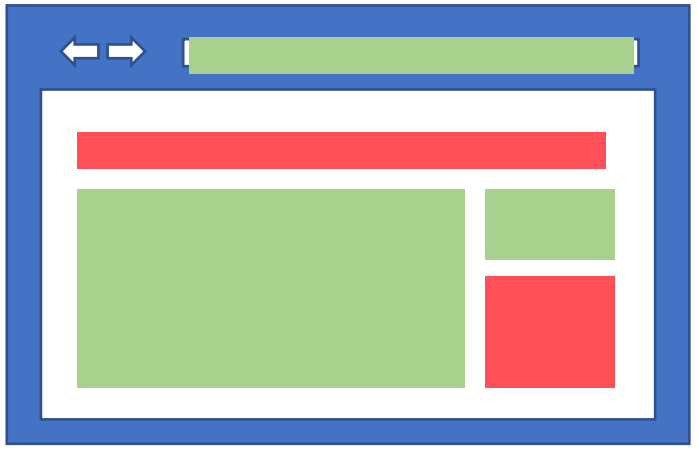


Unsigned Domain
No signed content
78.23%

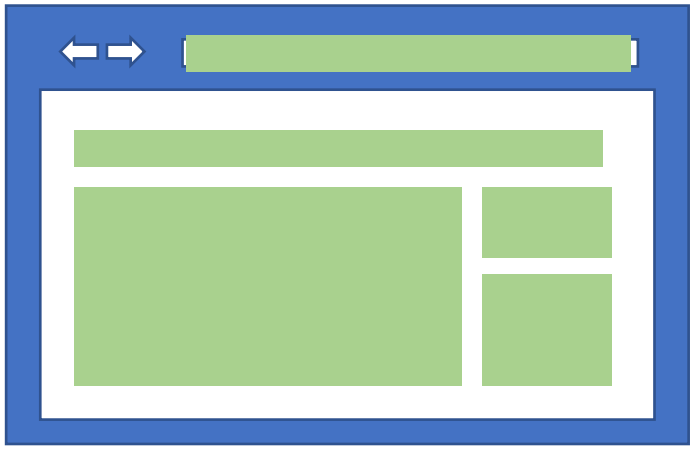


Unsigned Domain
1+ signed content
14.11%

Signed primary domains (7.7%)



Signed Domain
1+ signed content
6.08%



Signed Domain
Fully signed content
1.58%

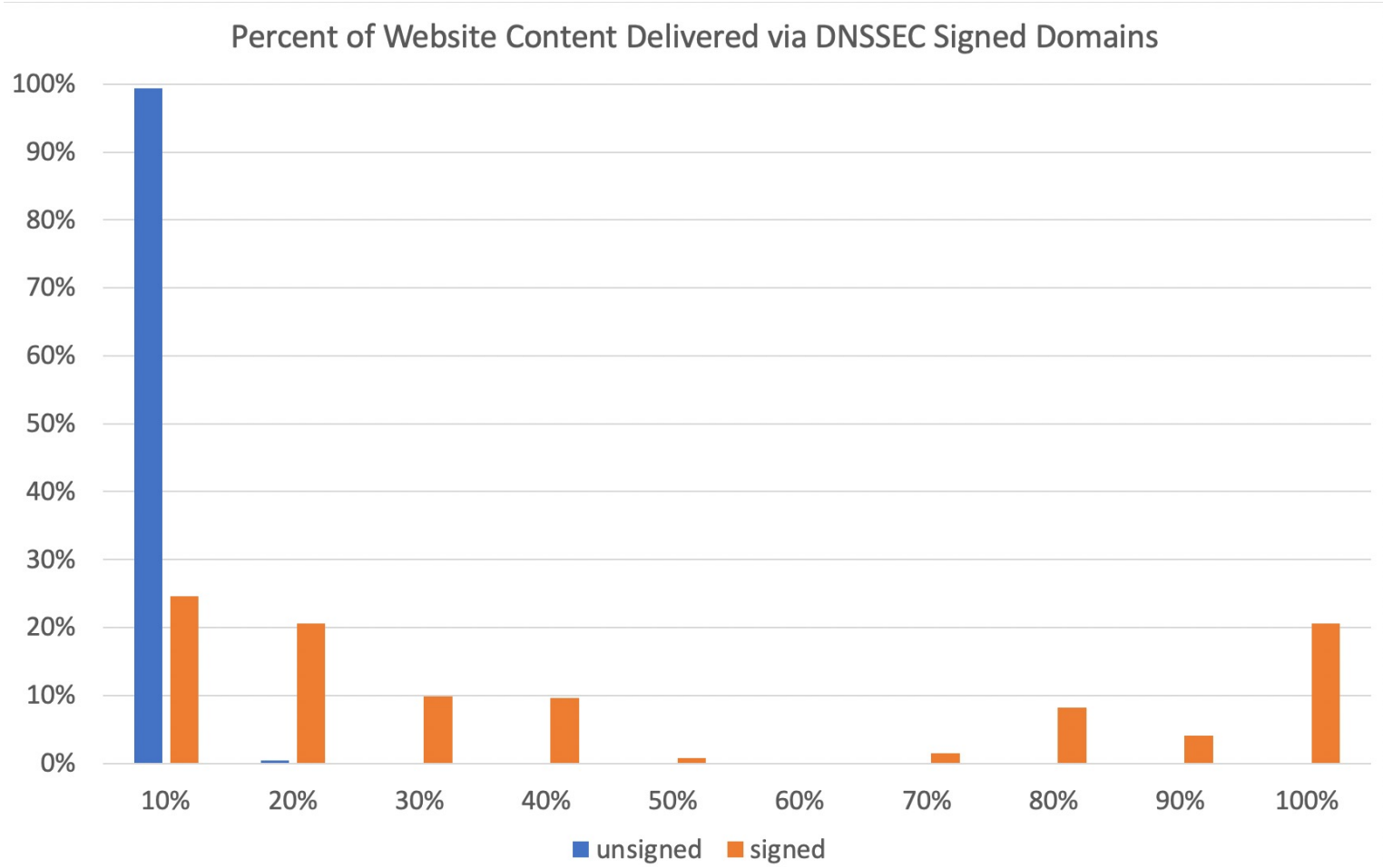
Analysis

Typically, how *much* of the website content is signed?

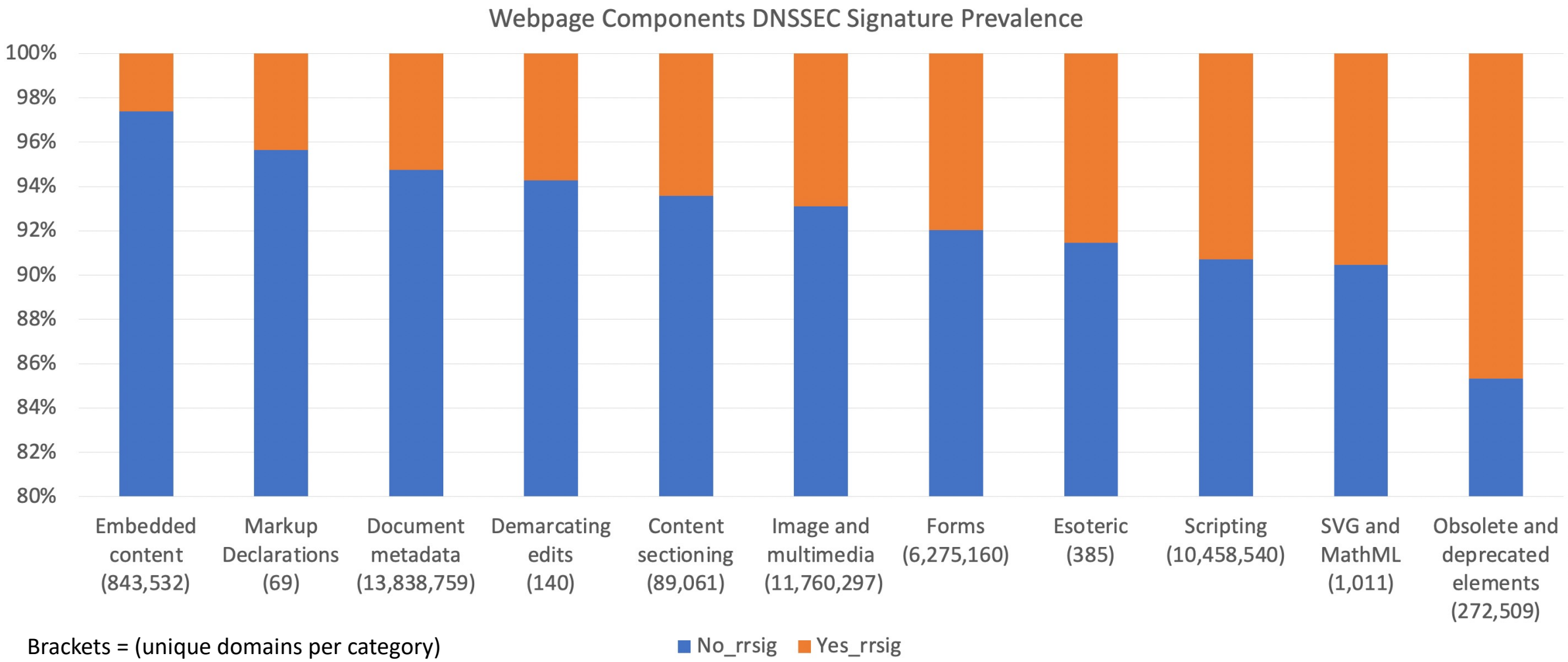
Unsigned primary domains have content that is mostly under 10% signed.

Signed primary domains have content that is much more likely to be signed including above 80% in many cases

Numbers are driven by the primary domain which delivers content itself

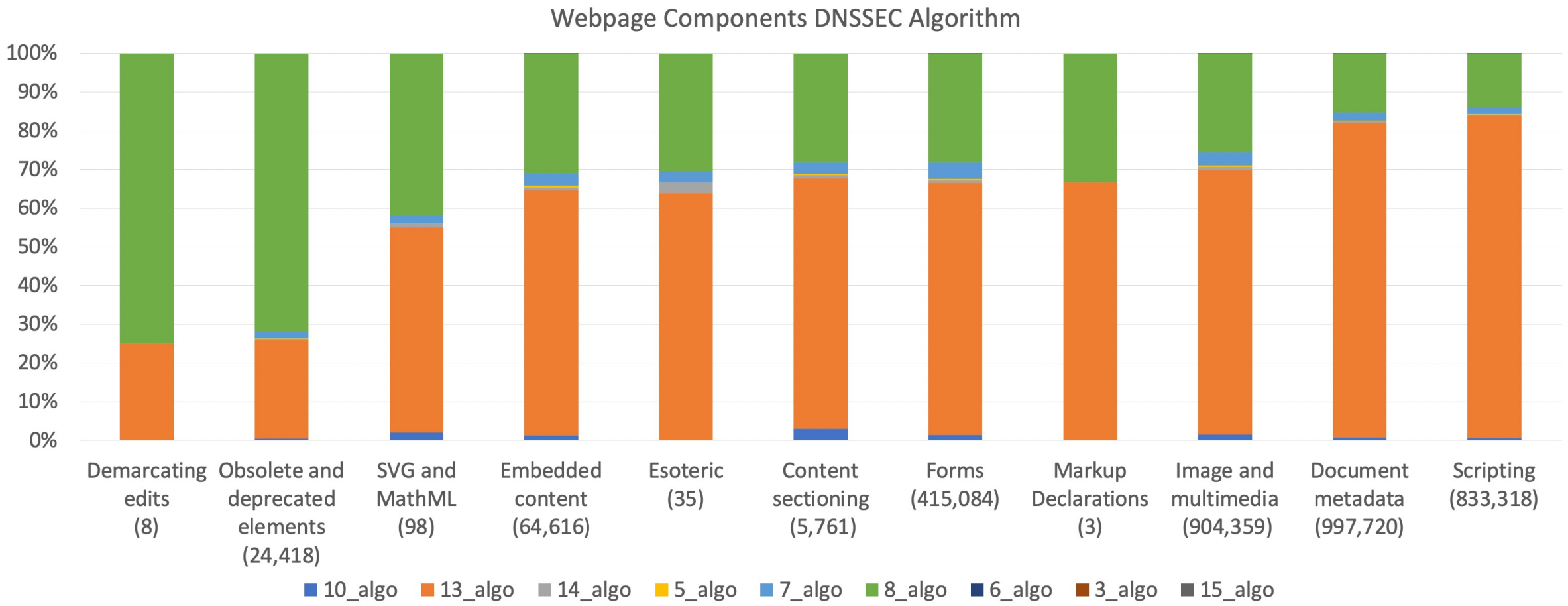


Analysis



Brackets = (unique domains per category)

Analysis



Brackets = (unique domains per category)

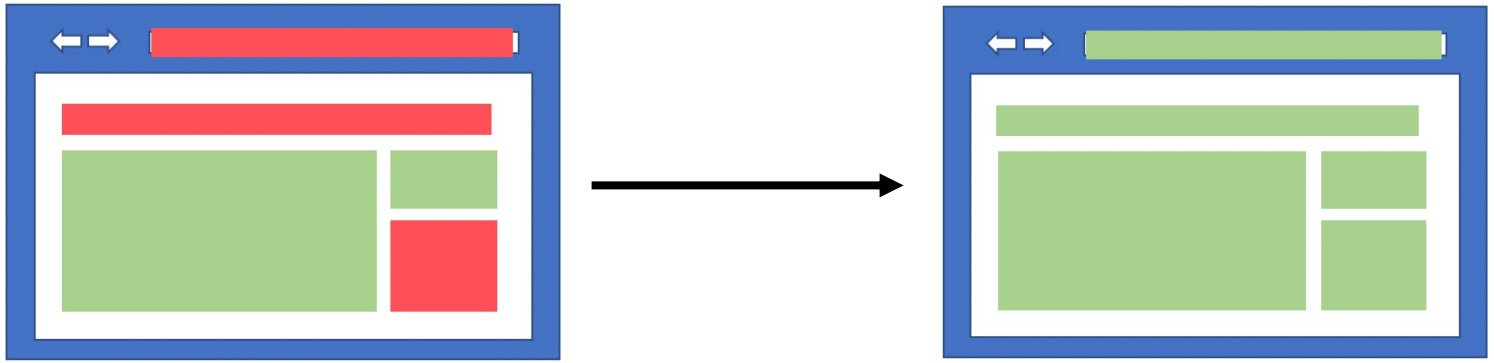
Analysis

Some content domains are seen on many different websites. Not a surprise.

The top 10 content domains are observed within 68% of websites in the dataset

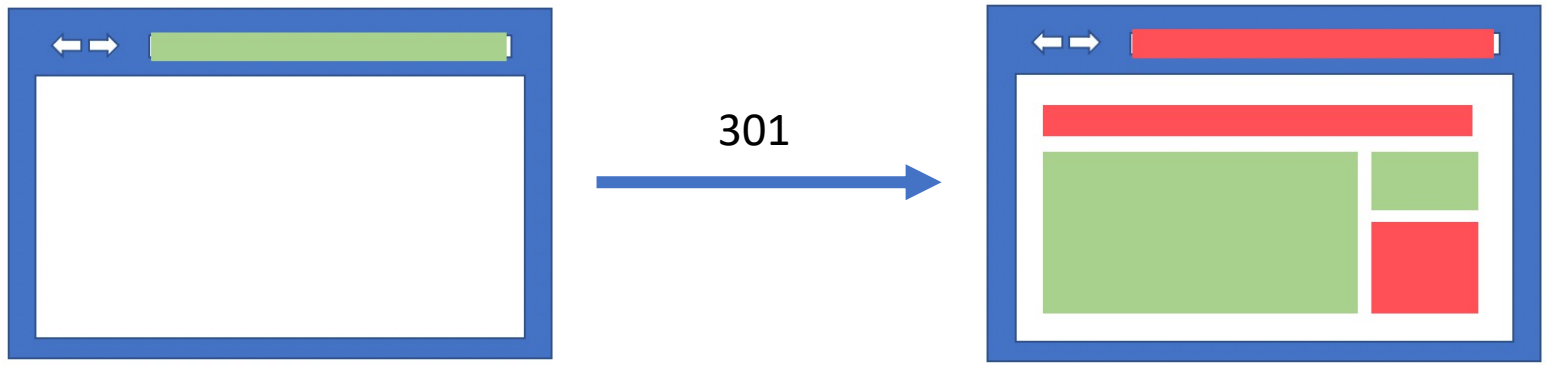
Popular content domains have significant influence on the whole dataset.

Content Domain	Count	Signed
<font cdn>	6,434,507	No
<popular website builder cdn domain>	4,915,602	No
<domain to deliver linked data schema>	3,458,998	No
<tracking>	2,717,833	No
<large tech company>	1,328,860	No
<popular script cdn domain>	1,153,884	No
<font cdn>	1,046,510	No
<tracking>	979,022	No
<popular cdn domain>	890,003	Yes
<large tech company>	766,794	No



Redirects

- What about HTTP redirects?
- Analysed 750k of the DNSSEC signed (primary) domains for redirects:



- 0.09% of the DNSSEC signed domains redirected to a non-signed domain
- Much better than I expected
- Probably more work to do here around studying the how and why… and what about redirects on the content domains too?

Conclusions and Thoughts

- DNSSEC has higher impact on website delivery than raw DNSSEC deployment statistics suggest. What do we want to measure?

  some DNSSEC 21.77% ------- > 7.7% ------- > 1.58% *all* DNSSEC

- We rarely see signed domains use HTTP redirects to unsigned ones. Good!
- 10 popular domains are observed in 68% of websites. Only one of these is signed
- Does a website owner have complete control over all content in their domain being delivered with DNSSEC? Large enterprises running their own CDN, quite possibly. Smaller website owners who rely on common supply chains, potentially not

Improvements:

headless browser, scanning from user endpoints, more robust redirect analysis, closer coupling between web scrape and DNS data collection, more robust tag dictionary