Analyzing the use of CNAME cloaking in the Wild

Boris van Groeningen
supervised by Georgios Smaragdakis

Delft University of Technology

Master Thesis Presentation, July 10

Outline

- Introduction
- Background
- Research Questions
- Methodology
- Datasets
- Analysis
- Discussion
- Future Work
- Demo

Figure: Created by Midjourney AI

CNAME Cloaking and GDPR - Introduction (1/2)

In short

- CNAME (Canonical Name) cloaking is a technique used to hide the true origin of a domain by disguising it behind a CNAME record.
- The General Data Protection Regulation (GDPR) is a legal framework that aims to protect the privacy and personal data of European Union (EU) citizens.

Motivation

- CNAME cloaking presents significant challenges to GDPR compliance and data protection.
- Researching the use of CNAME cloaking is important for understanding the potential privacy risks it poses.

Criteo[1] - Introduciton (2/2)

Personalised advertising: CRITEO fined EUR 40 million

22 June 2023

On 15 June 2023, the CNIL sanctioned CRITEO, which specialises in online advertising, with a fine of EUR 40 million, in particular for failing to verify that the persons from whom it processed data had given their consent.

Figure: Snippet taken from: `https://www.cnil.fr/`

[1] Spoiler alert: we will encounter Criteo as well

DNS - Background (1/4)

Definition

The Domain Name System (DNS) is a hierarchical decentralized naming system that translates domain names into IP addresses and provides various services related to domain names.

`example.com` → 192.0.2.1

Key Components

DNS consists of several key components:

- **DNS Resolver**: Client software that initiates DNS queries and receives responses.
- **DNS Record**: A database entry that mapping domain names → IP addresses
- **DNS Server**: Stores DNS records and provides responses to DNS queries.

CDNs - Background (2/4)

Definition

A Content Delivery Network (CDN) is a distributed network of servers strategically located across the globe to deliver web content efficiently to end users.

More info

- CDNs help improve the performance, availability, and scalability
- Reducing latency via caching
- Popular CDNs: Cloudflare, Faslty, Azure, etc.

Cookies - Background (3/4)

What are cookies?

- Cookies are small text files stored on a user's computer by websites they visit
- They are used to store information and track user activity
- Types: **Session** & **Persistent**

Potential Dangers

- **Security risks**: Malicious cookies can be used for phishing, session hijacking, or cross-site scripting attacks
- **Tracking and profiling**: Enables advertisers to gather data and track users

Embedded Objects - Background (4/4)

Definition

Embedded objects are pieces of content inserted into the webpage.

Examples

- Image
- Video
- Audio
- PDF Documents
- Links

Research Questions

RQ 1

RQ 2

RQ 3

Research Questions

RQ 1

How prevalent is CNAME cloaking on the web?

RQ 2

RQ 3

Research Questions

RQ 1

How prevalent is CNAME cloaking on the web?

RQ 2

What are the characteristics of websites that use CNAME cloaking?

RQ 3

Research Questions

RQ 1

How prevalent is CNAME cloaking on the web?

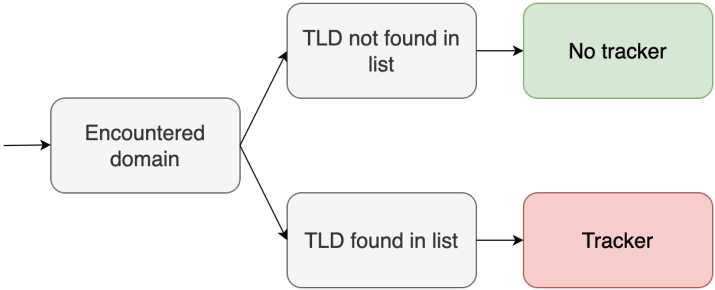
RQ 2

What are the characteristics of websites that use CNAME cloaking?

RQ 3

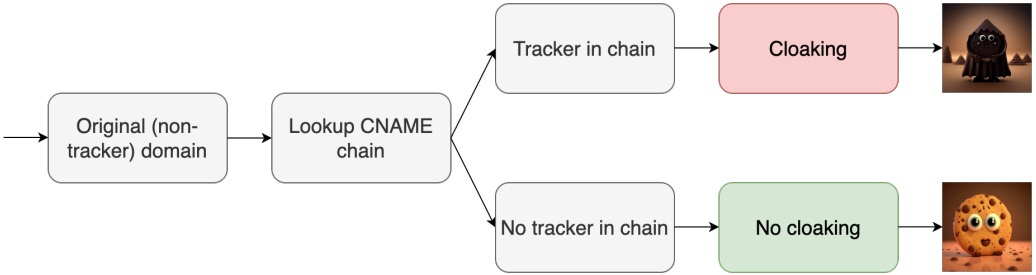
How is cloaking distributed amongst ranking intervals?

Tracking Definition - Methodology (1/6)



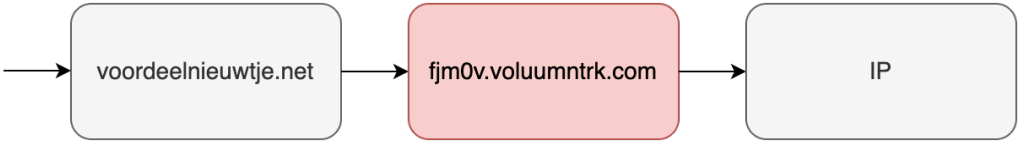
TLD: Top-Level Domain

Cloaking Definition - Methodology (2/6)

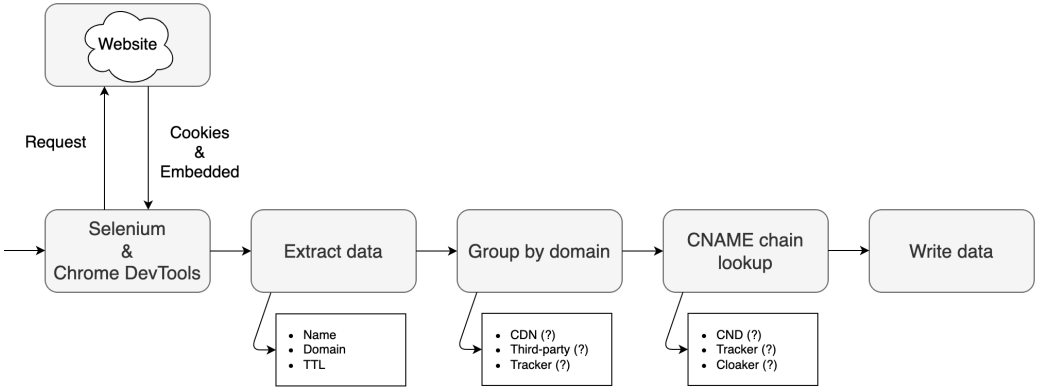


Cloaking Example - Methodology (3/6)

Found on drimble.nl. An embedded object by the domain voordeelnieuwtje.net.



Data Collection - Methodology (4/6)



Data Overview - Methodology (5/6)

- Cookies & embedded objects are grouped by domain
  - Fewer lookups
  - Better structured data

```
{
    "website_name_1": {
        "cookies": {
            "domain_1": {
                "cookie_data": [
                    {
                        "name": "str",
                        "expires": "int"
                    },
                    ...
                ],
                "is_third_party": "bool",
                "is_tracker": "bool",
                "is_CDN": "bool",
                "chain": [
                    {
                        "domain": "str",
                        "TTL": "int",
                        "is_tracker": "bool",
                        "is_CDN": "bool",
                        "is_cloaking": "bool",
                        "IPs": ["only added if cloaking is true", ...]
                    },
                    ....,
                ]
            },
            "..."
        },
        "embedded": {
            "domain_1": "chain",
            "..."
        }
    },
    "website_name_2": {
        "..."
    }
}
```

Figure: JSON structure

Analysis - Methodology (6/6)

Points of interest:

- Cloaking encounters in the dataset
  - Whether they originate from cookies and/or embedded objects
- TTLs of cloakers
- Percentage of cloakers in ranking intervals
- Type of websites for prominent cloakers

Overview - Datasets (1/5)

Multiple datasets are used to check for presence of cloaking:

- Alexa → top 1M sites
- Dutch → all sites ending in `.nl` in Alexa
- Rijksoverheid → all official links from Rijksoverheid
- G20 → official websites of G20 countries
- Covid → official websites with covid information
- Fakuda → domains which used to have cloaking in Jan 2020

Overview - Datasets (2/5)

Dataset	**# Domains**
Alexa	707k
Dutch	11k
Rijksoverheid	1.8k
G20	5.8k
Covid	198
Fakuda	1762

Table: The datasets used for the experiments.

Reachable pages - Datasets (3/5)

	Total domains	Reachable domains	Percentage (%)
Alexa	10000	8858	88.58
Dutch	10709	9770	91.23
Rijksoverheid	1812	1332	73.51
G20	5813	3435	59.09
Covid	198	156	78.79
Fakuda	1762	1698	96.37

Table: Datasets and their reachable pages.

Difference in performance (average time per domain). G20 was very slow → large number of international pages (i.e. India).

Trackers - Datasets (4/5)

Constructed a list by taking the union of the following tracker lists:

- Adguard DNS
- Easylist
- Nocoin
- Easyprivacy

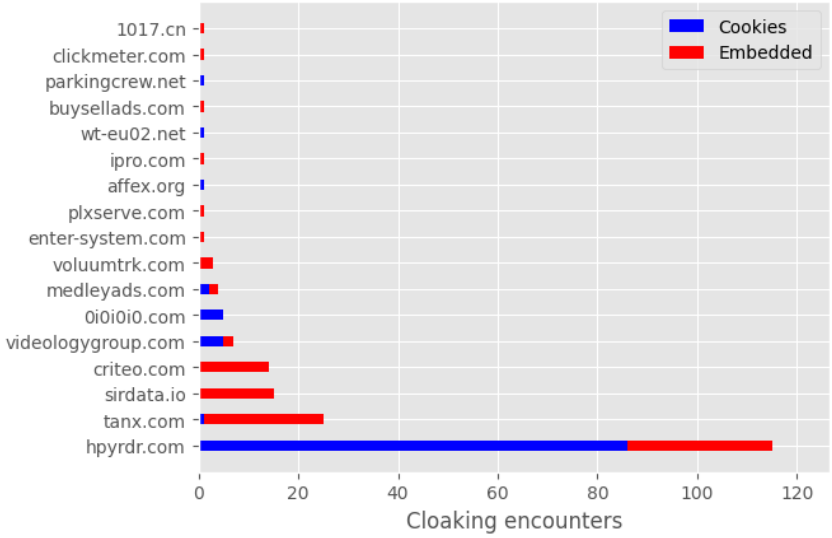
CDNs - Datasets (5/5)

The list of CDNs we will be checking for is based on the list used in the paper: Seven Years in the Life of Hypergiants' Off-Nets[2]

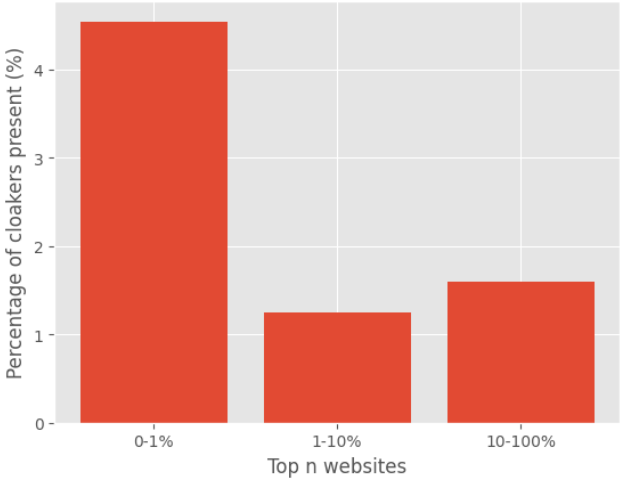
The list includes: Google, Facebook, Instagram, Netflix, Akamai, Alibaba, Cloudflare, Amazon, CDN Networks, Limelight, Apple, Twitter, Msegde, and Fastly

[2]By one of my favorite authors

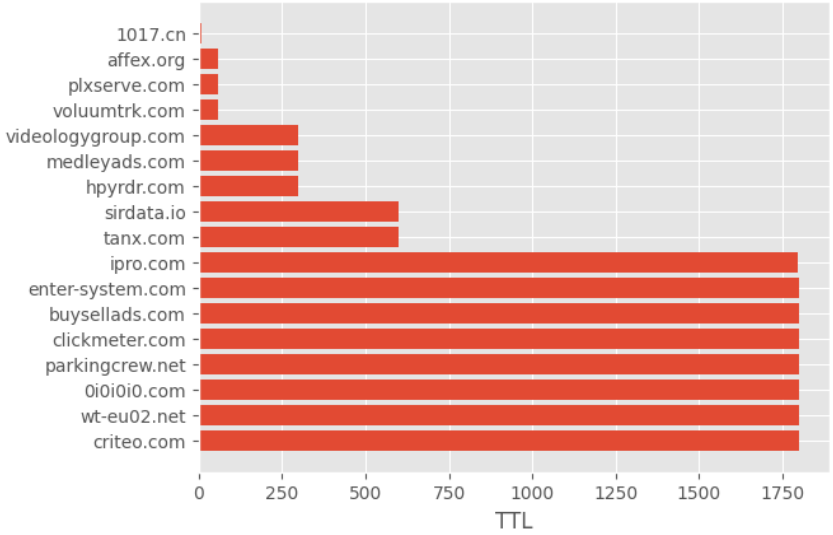
Alexa Encounters - Analysis (1/14)



Alexa Ranking - Analysis (2/14)

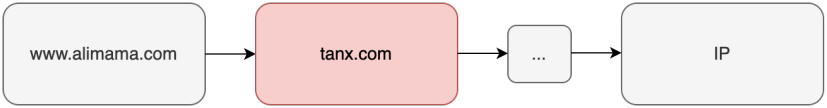


Alexa TTL - Analysis (3/14)



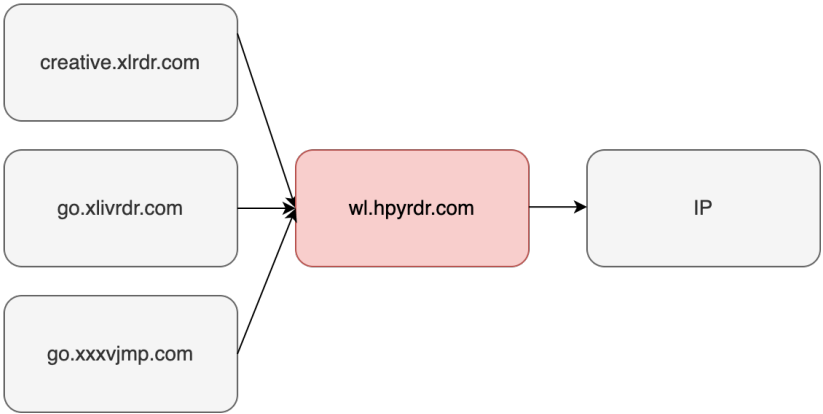
Alexa Example 1 - Analysis (4/14)

On the website `alieexpress.com` (and other ali-related websites), an embedded object by the domain of `www.alimama.com` is found.

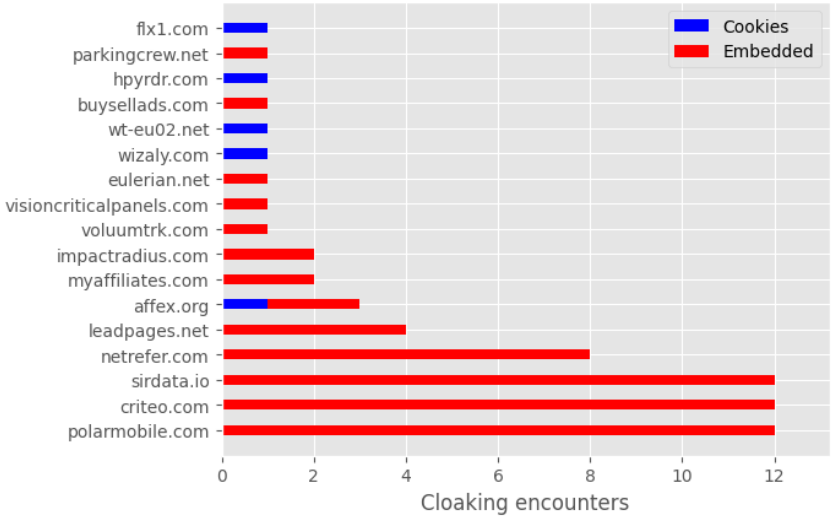


Alexa Example 2 - Analysis (5/14)

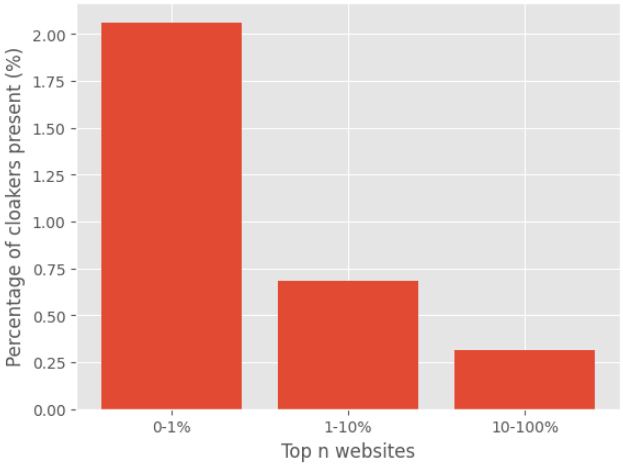
On multiple adult websites, the cloaker `wl.hpyrdr.com` has been detected through both cookies or embedded objects.



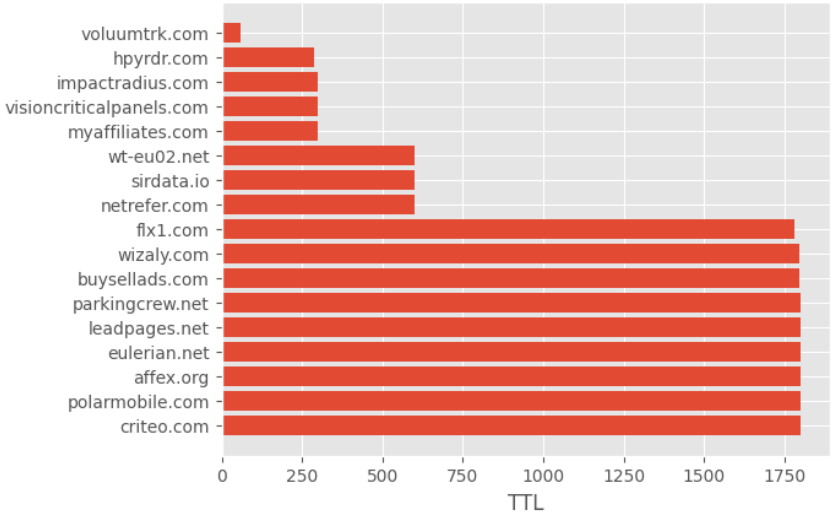
Dutch Encounters - Analysis (6/14)



Dutch Ranking - Analysis (7/14)

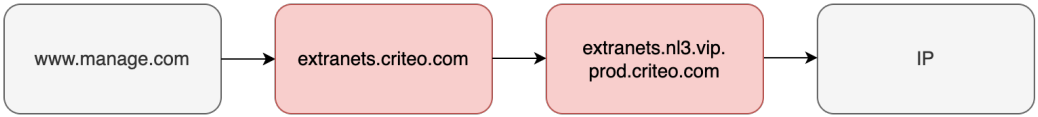


Dutch TTL - Analysis (8/14)



Dutch Example - Analysis (9/14)

Criteo is another popular cloaker. It originates from an embedded object associated with the domain `www.manage.com`



Rijksoverheid - Analysis (10/14)

Fortunately[3] no cloaking-based tracking has been detected in the dataset of Rijksoverheid.

[3] But sadly for me...

G20 - Analysis (11/14)

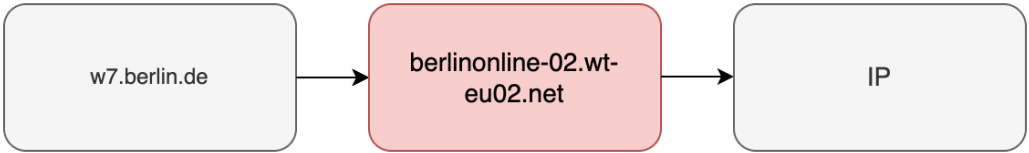
Three cases of cloaking have been detected on:

- berlin.de
- ale.ombudsrat.de
- michiganlotter.com

berlin.de is the official portal website of Germany's capital

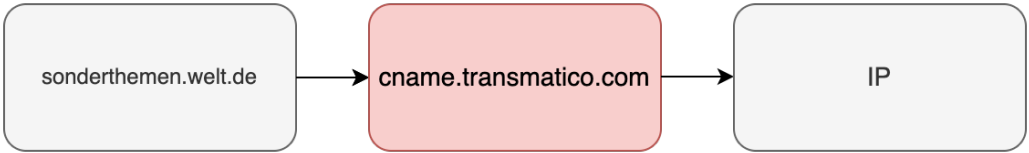
Berlin - Analysis (12/14)

`berlin.de` contains a cookie by the domain of `w7.berlin.de` which resolves to:



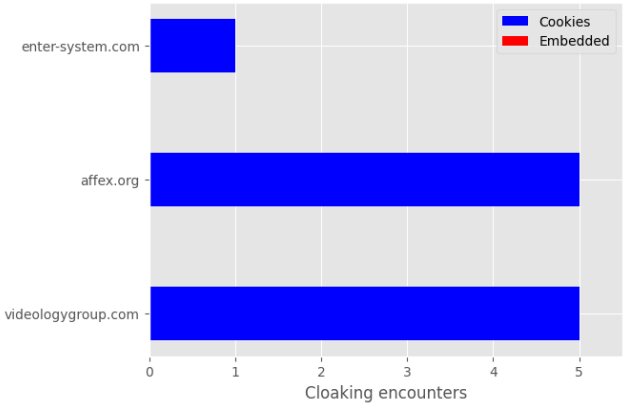
Covid - Analysis (13/14)

Only 1 case of cloaking has been detected in the Covid dataset. Namely, on the page: `welt.de/themen/coronavirus-epidemie/`, which is a German news website.



Fakuda - Analysis (14/14)

Cloaking has drastically decreased for this dataset.



Research Questions - Discussion (1/3)

- Happens quite often:
  - Alexa dataset $\rightarrow$ 1.59%
  - Dutch daatset $\rightarrow$ 0.37%
- Cloakers are mostly category specific (i.e. `wl.hpyrdr.com` with adult websites)
- Cloakers are most present at the higher ranked webpages (see Alexa and Dutch datasets)

Limitations - Discussion (2/3)

- The web is ever-changing, meaning different outcomes at different times
- We could have missed cloaking due to our 60 seconds timeout per page (for efficiency)
- Cloaking detection is as good as the provided trackers list
  - Tracker not in list $\rightarrow$ not considered cloaking
  - Requires an up-to-date list of trackers

Future Work - Discussion (3/3)

- More domains to crawl
- Having more data could lead to a better analysis/pattern recognition
- Tool can be extended upon $\rightarrow$ Developing countermeasures (browser extension perhaps)

Up next...

Public Repo

Code available at:
github.com/Boris304/cname-cloaking