Online Tracking and Security - report of assignement 2

Group A2 12
Igor Silaire, Nikolai Savulkin, Edvardas Ražanskas, Thomas Manolis

Contents

1 Task 1 **2**

2 Task 2 **4**

3 Task 3 **6**

4 Task 4 **8**

5 Task 5 **8**

6 Task 6 **10**

7 Task 7 **11**
7.1 Crawl-Accept 11
7.2 Crawl-Block 12
7.3 Crawl-Reject 12

8 Task 8 **13**
8.1 Crawl-Accept 13
8.2 Crawl-Block 14
8.3 Crawl-Reject 15

9 Task 9 **16**

10 Task 10 **16**

11 Task 11 **18**

12 Task 12 **19**

13 Task 13 **20**

14 Task 14 **22**

15 Task 15 **22**
15.1 What were the most challenging parts of the assignment? 22
15.2 What were the findings that surprised you, if any? 22

1 Task 1

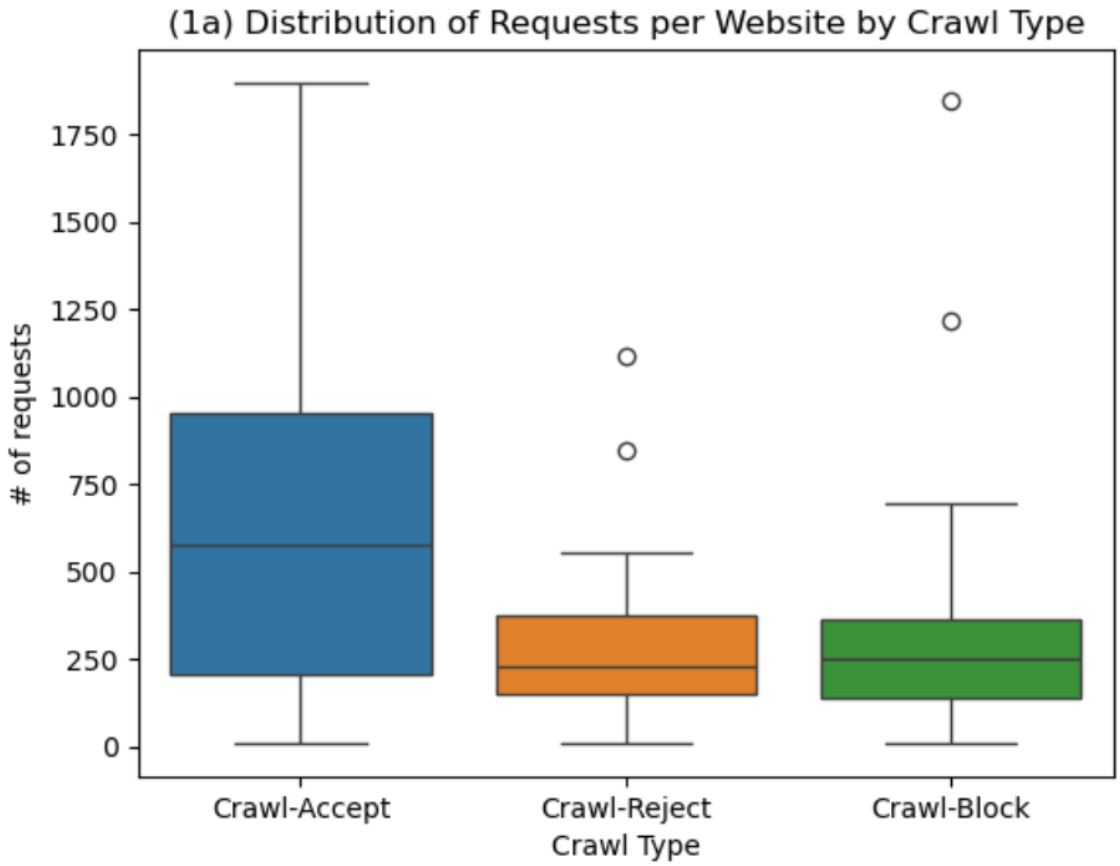

Figure 1: (1a) Distribution of Requests per Website by Crawl Type

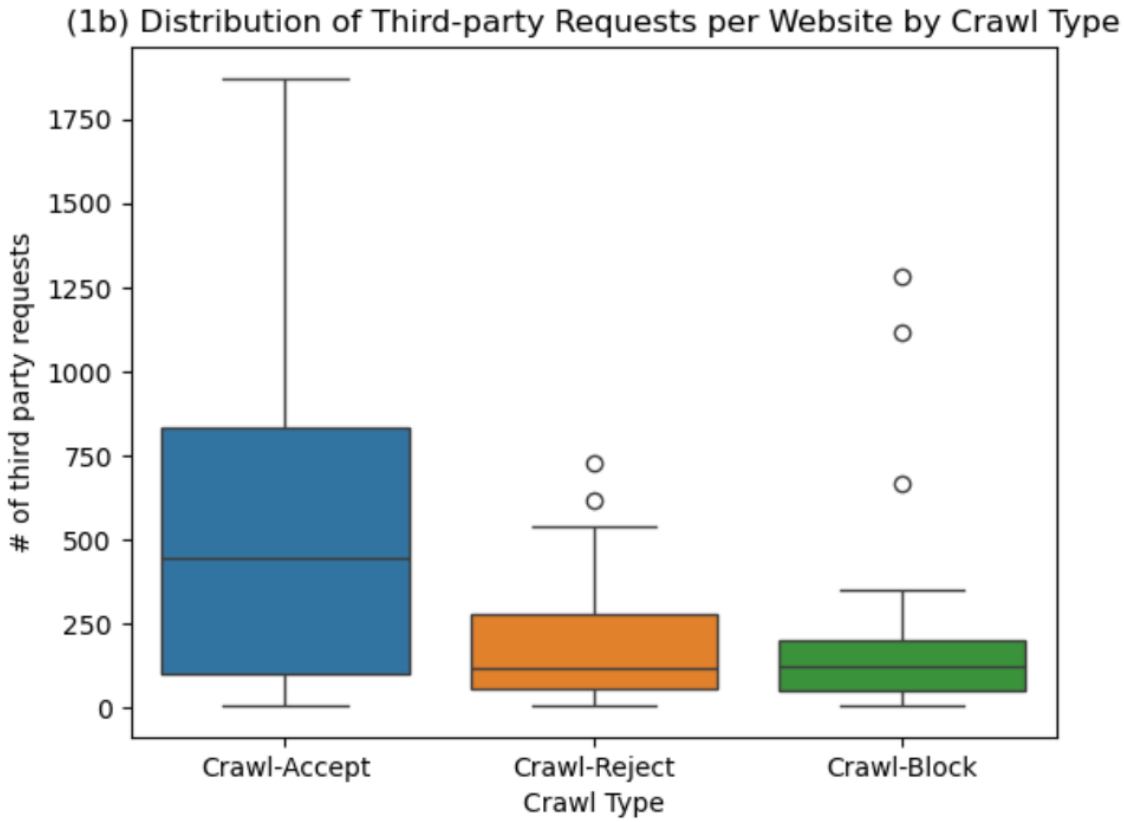

Figure 2: (1b) Distribution of Third-party Requests per Website by Crawl Type

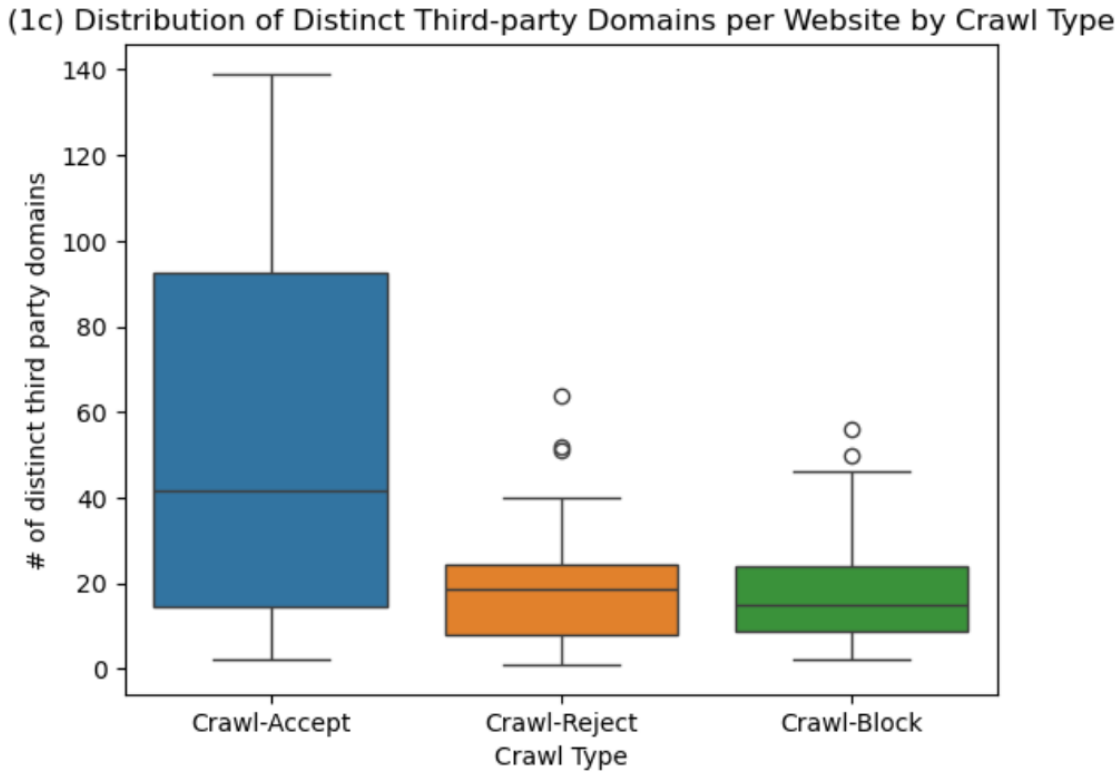

Figure 3: (1c) Distribution of Distinct Third-party Domains per Website by Crawl Type

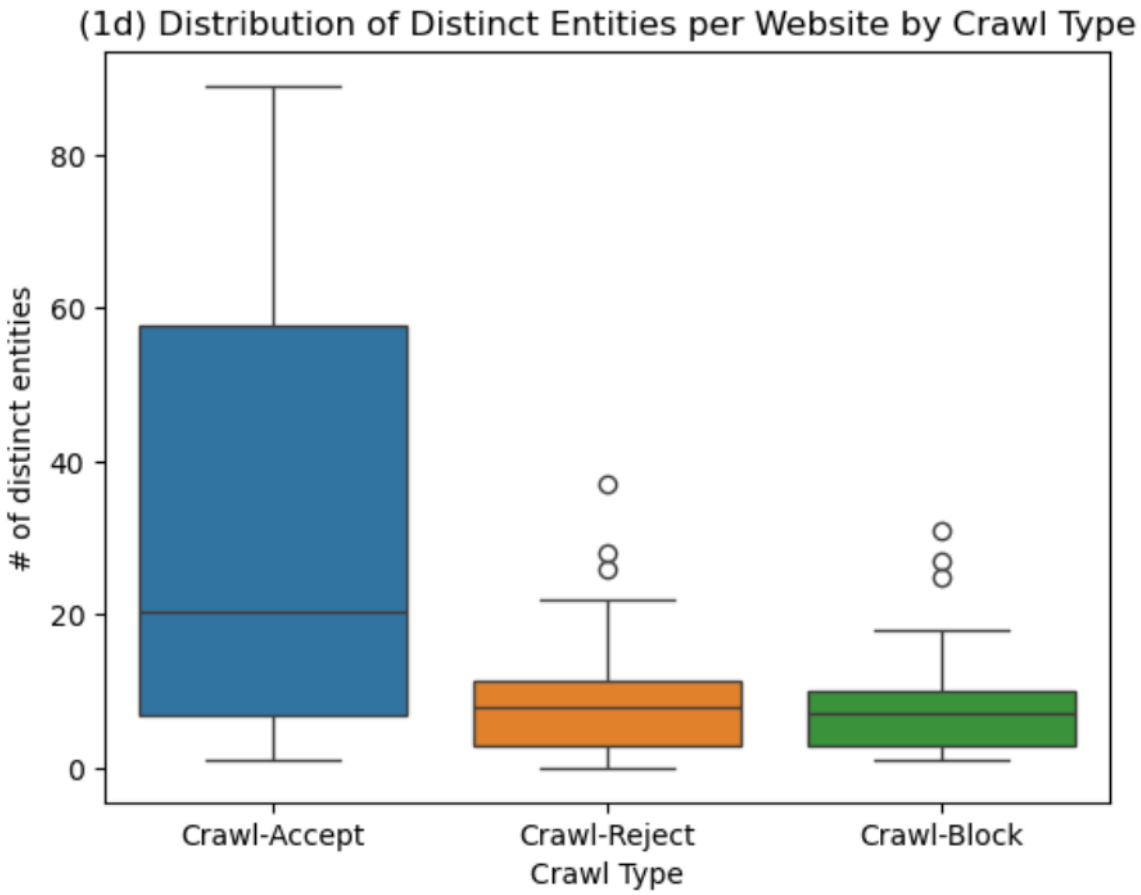

Figure 4: (1d) Distribution of Distinct Entities per Website by Crawl Type

2 Task 2

	Crawl-Accept			Crawl-Block			Crawl-Reject		
Metric	**min**	**median**	**max**	**min**	**median**	**max**	**min**	**median**	**max**
Number of requests per website	10	576.5	1897	10	251.0	1845	10	229.0	1114
Number of third-party requests per website	8	444.0	1870	8	122.0	1282	8	117.5	728
Number of distinct third party domains	2	41.5	139	2	15.0	56	1	18.5	64
Number of distinct entities	1	20.5	89	1	7.0	31	0	8.0	37

Figure 5: Metrics and statistics comparison across different crawl types

Metric	Crawl-Accept			Crawl-Block			Crawl-Reject		
	min	median	max	min	median	max	min	median	max
Number of requests per website	10	576.5	1897	10	251.0	1845	10	229.0	1114
Number of third-party requests per website	8	444.0	1870	8	122.0	1282	8	117.5	728
Number of distinct third party domains	2	41.5	139	2	15.0	56	1	18.5	64
Number of distinct entities	1	20.5	89	1	7.0	31	0	8.0	37

Metrics and statistics comparison across different crawl types

3 Task 3

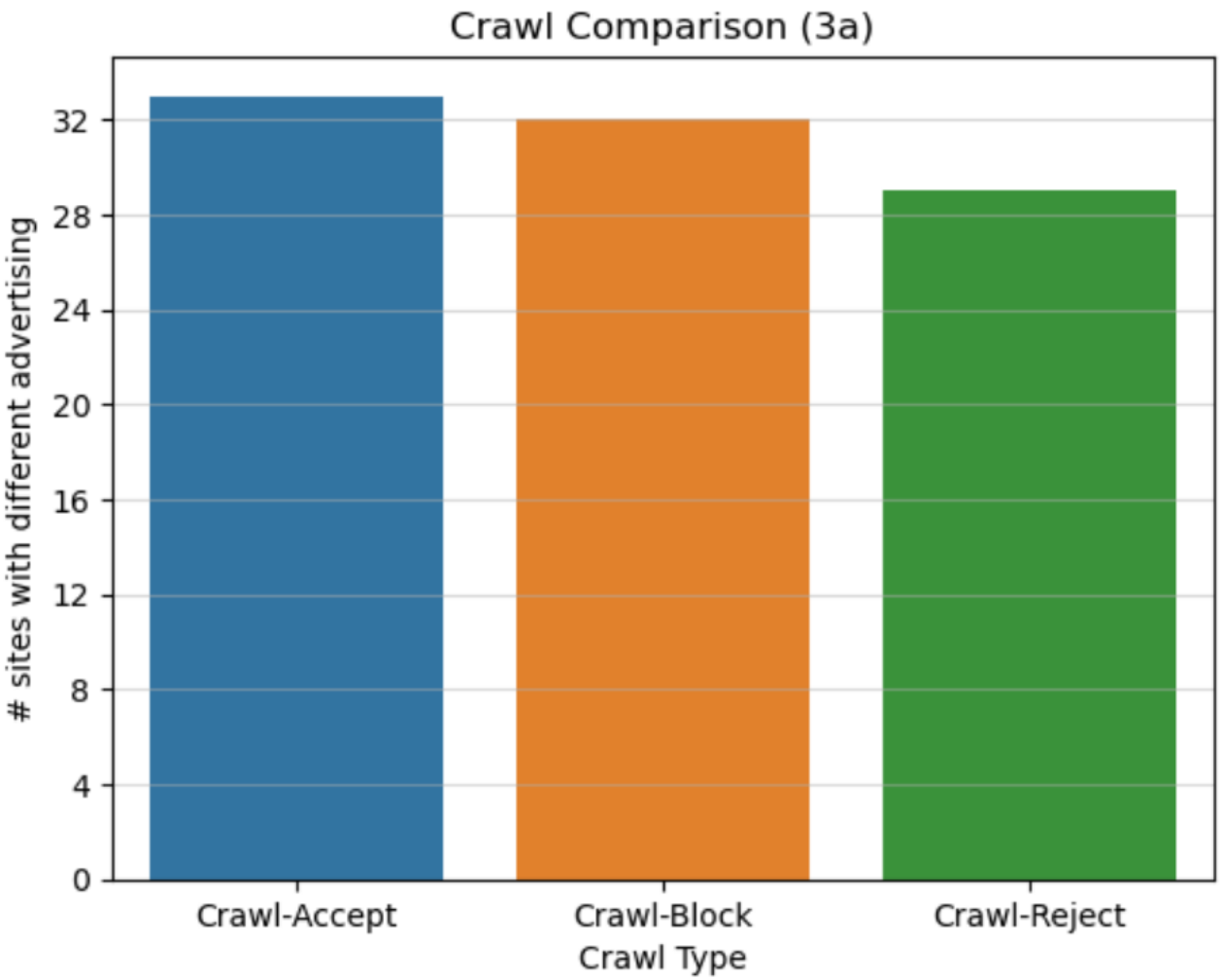

Figure 6: Number of websites with different advertising by different crawl types

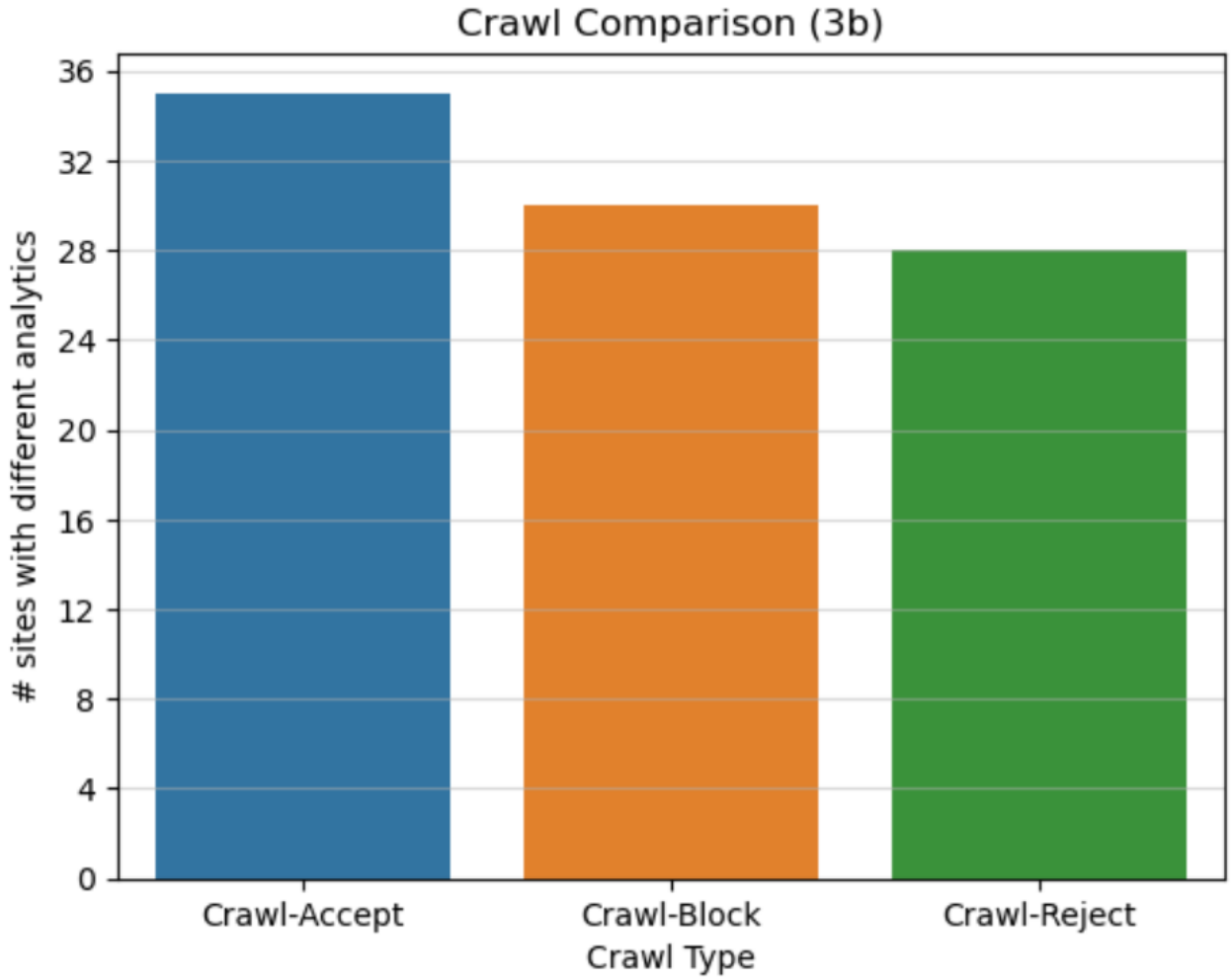

Figure 7: Number of websites with different analytics by different crawl types

4 Task 4

	European News Sites			US News Sites		
Metric	**min**	**median**	**max**	**min**	**median**	**max**
Number of requests per website	93	798.0	1897	10	469.0	1530
Number of third-party requests per website	12	636.5	1870	8	311.5	1429
Number of distinct third party domains	3	66.5	137	2	32.0	139
Number of distinct entities	1	24.5	89	1	16.5	89

Figure 8: Metrics and statistics comparison across different crawl types

5 Task 5

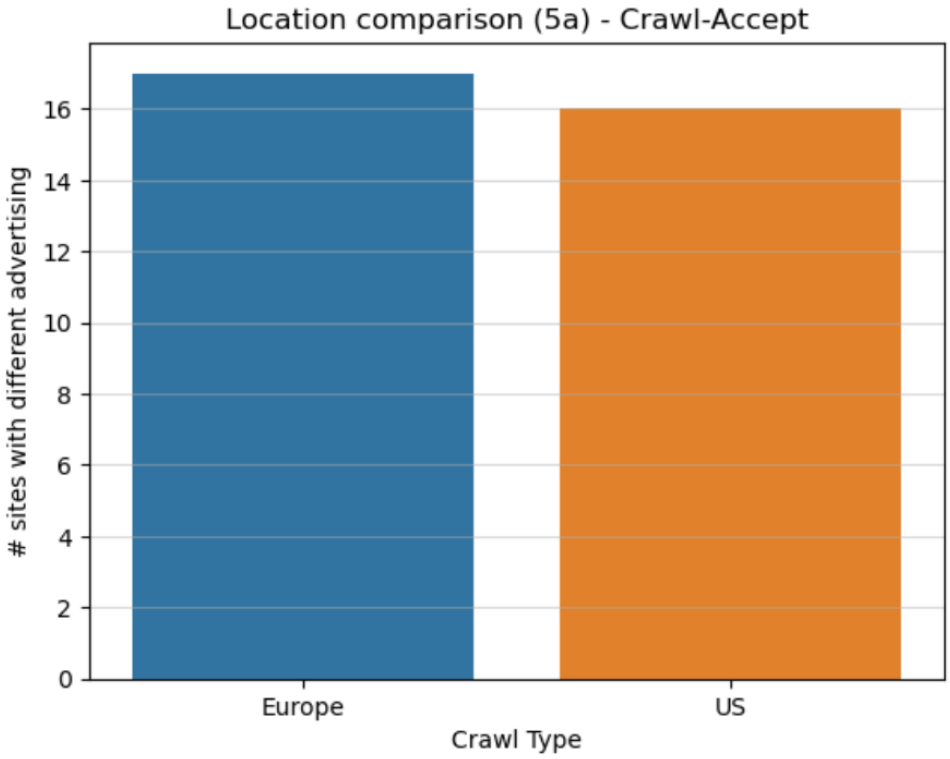

Figure 9: Number of websites with different advertising by different location

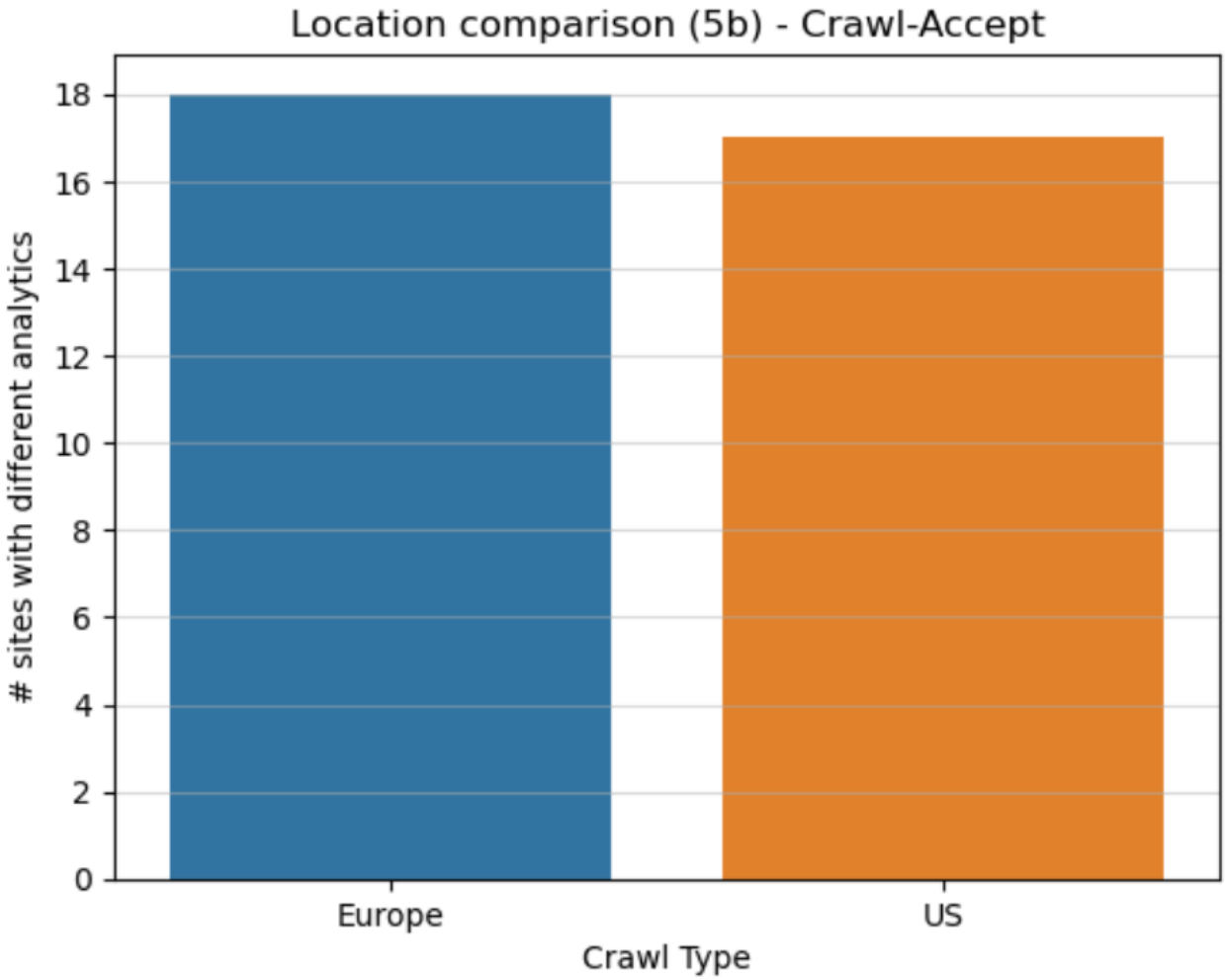

Figure 10: Number of websites with different analytics by different location

6 Task 6

METHOD: Process entries chronologically. For each domain, track which cookies were set via Set-Cookie. If a cookie appears in a request but was never set by Set-Cookie for that specific earlier, it's likely JS-set.

1. Track all cookies set via Set-Cookie response headers (server-side)
2. Track all cookies sent in Cookie request headers (requests to server)
3. Cookies appearing in requests but NEVER in Set-Cookie are JS-set

RATIONALE:

1. Server-side cookies appear in Set-Cookie response headers
2. JS-set cookies (document.cookie=...) only appear in Cookie request headers
3. By tracking server-set cookies per domain, we identify client-side cookies

In total 2533 potential JavaScript-set cookies were found. Full result set is attached as 'task6_js_cookies.csv', below there is a table of 5 potentially JS-set cookies for ad.nl website with cookie domain:

	cookie_name	website	cookie_domain
0	_sp_ses.e23f	ad.nl	dpgmedia.nl
1	_sp_id.e23f	ad.nl	dpgmedia.nl
2	gtm_temptation_template_id	ad.nl	ad.nl
3	tcf20_purposes	ad.nl	ad.nl
4	_vwo_uuid_v2	ad.nl	ad.nl

Figure 11: Table of 5 potentially JS-set cookies for ad.nl website with cookie domain

7 Task 7

Below there are three tables (one table for each crawl type) of the ten most prevalent third-party domains (based on the number of distinct websites where the third-party domain is present). The Disconnect Categories column contain categories of the domain in Disconnect's blocklist (or NA, if absent).

7.1 Crawl-Accept

	Third-party domain	Number of distinct websites	Disconnect categories
0	googlesyndication.com	29	NA
1	google.com	28	Advertising, Analytics, FingerprintingGeneral, Social
2	doubleclick.net	28	NA
3	amazon-adsystem.com	27	NA
4	googleapis.com	25	NA
5	adtrafficquality.google	25	NA
6	criteo.com	25	Advertising
7	googletagmanager.com	23	NA
8	adsrvr.org	22	NA
9	bidswitch.net	22	NA

Figure 12: Ten most prevalent third-party domains for Crawl-Accept

7.2 Crawl-Block

	Third-party domain	Number of distinct websites	Disconnect categories
0	googletagmanager.com	22	NA
1	google.com	21	Advertising, Analytics, FingerprintingGeneral, Social
2	doubleclick.net	19	NA
3	googleapis.com	18	NA
4	googlesyndication.com	15	NA
5	amazon-adsystem.com	14	NA
6	gstatic.com	12	NA
7	scorecardresearch.com	11	NA
8	jsdelivr.net	9	NA
9	cookielaw.org	9	NA

Figure 13: Ten most prevalent third-party domains for Crawl-Block

7.3 Crawl-Reject

	Third-party domain	Number of distinct websites	Disconnect categories
0	googletagmanager.com	23	NA
1	doubleclick.net	22	NA
2	googlesyndication.com	20	NA
3	googleapis.com	20	NA
4	google.com	18	Advertising, Analytics, FingerprintingGeneral, Social
5	gstatic.com	15	NA
6	amazon-adsystem.com	14	NA
7	scorecardresearch.com	11	NA
8	google-analytics.com	11	NA
9	adtrafficquality.google	10	NA

Figure 14: Ten most prevalent third-party domains for Crawl-Reject

8 Task 8

Below there are three tables (one table for each crawl type) of the top ten websites that send requests to the highest number of distinct third-party domains.

8.1 Crawl-Accept

	Website	Country code	Number of distinct third-party domains
0	apnews.com	us	139
1	libero.it	it	137
2	gazzetta.it	it	133
3	the-sun.com	uk	129
4	corriere.it	it	128
5	nypost.com	us	126
6	newsweek.com	us	119
7	dailymail.co.uk	uk	107
8	telegraaf.nl	nl	106
9	huffpost.com	us	95

Figure 15: Top ten websites that send requests to the highest number of distinct third-party domains for Crawl-Accept

8.2 Crawl-Block

	Website	Country code	Number of distinct third-party domains
0	nbcnews.com	us	56
1	newsweek.com	us	50
2	apnews.com	us	46
3	ad.nl	nl	36
4	nypost.com	us	35
5	gazzetta.it	it	35
6	telegraaf.nl	nl	26
7	businessinsider.com	us	25
8	the-sun.com	uk	25
9	repubblica.it	it	24

Figure 16: Top ten websites that send requests to the highest number of distinct third-party domains for Crawl-Block

8.3 Crawl-Reject

	Website	Country code	Number of distinct third-party domains
0	nbcnews.com	us	64
1	newsweek.com	us	52
2	apnews.com	us	51
3	businessinsider.com	us	40
4	nypost.com	us	35
5	gazzetta.it	it	34
6	bbc.co.uk	uk	28
7	corriere.it	it	27
8	cnbc.com	us	26
9	telegraaf.nl	nl	26

Figure 17: Top ten websites that send requests to the highest number of distinct third-party domains for Crawl-Reject

9 Task 9

The objective of this task is to identify the 10 visits that generated the highest number of distinct server IP addresses.

	Website	Crawl type	Country code	Number of distinct server IPs
0	ad.nl	Accept	nl	0
1	apnews.com	Accept	us	0
2	axios.com	Accept	us	0
3	bbc.co.uk	Accept	uk	0
4	bloomberg.com	Accept	us	0
5	businessinsider.com	Accept	us	0
6	buzzfeed.com	Accept	us	0
7	cbsnews.com	Accept	us	0
8	cnbc.com	Accept	us	0
9	cnn.com	Accept	us	0

Figure 18: Top 10 visits by number of distinct server IPs

Analysis:
The analysis is unable to determine the number of distinct IP addresses because the corresponding information (the `serverIPAddress` field of the HAR entries) does not appear to have been recorded correctly by the crawler (Playwright) during data acquisition. This field was consistently null or missing, preventing any valid count.

10 Task 10

This analysis focuses on `Permissions-Policy` headers (formerly `Feature-Policy`) that control access to sensitive browser features. The goal was to identify the five permissions most frequently **completely disabled**.

Permission	Crawl-Accept	Crawl-Reject	Crawl-Block
geolocation	geolocation: 15	geolocation: 8	geolocation: 8
microphone	microphone: 15	microphone: 8	microphone: 8
camera	camera: 15	camera: 8	camera: 8
midi	midi: 14	midi: 8	midi: 8
interest-cohort	interest-cohort: 13	interest-cohort: 8	interest-cohort: 8

Figure 19: Number of distinct sites that have completely disabled sensitive permissions per crawl mode.

Analysis:

1. **Hardware Privacy:** Permissions related to user hardware (`geolocation`, `microphone`, `camera`) lead the ranking, being disabled on **15 sites** in `Crawl-Accept` mode. This signals a strong intention on the part of first-party sites to **prevent the abuse of these sensitive resources** by third-party scripts.
2. **Rejection of New Tracking:** The frequent disabling of the `interest-cohort` permission (13 sites) reflects a **widespread rejection** by sites of Google's new advertising tracking API.
3. **Impact of Crawl Mode:** Deactivation is observed less often in `Crawl-Reject` and `Crawl-Block` modes (8 sites).

11 Task 11

The objective of this task is to determine how many websites explicitly define a `referrer policy` to control the amount of URL information transmitted to third parties.

	Crawl-Accept	Crawl-Reject	Crawl-Block
0	strict-origin-when-cross-origin: 22	strict-origin-when-cross-origin: 16	strict-origin-when-cross-origin: 21
1	same-origin: 9	same-origin: 6	same-origin: 5
2	no-referrer: 8	no-referrer: 4	no-referrer: 4
3	unsafe-url: 3	unsafe-url: 3	unsafe-url: 3
4	origin: 2	strict-origin: 1	strict-origin: 1
5	strict-origin: 1	origin-when-cross-origin, strict-origin-when-cross-origin: 1	origin-when-cross-origin, strict-origin-when-cross-origin: 1
6	origin-when-cross-origin, strict-origin-when-cross-origin: 1		

Figure 20: Distribution of observed reference policies by crawl mode.

Analysis

1. **Dominance of Security Policies:** The most frequently observed policy is `strict-origin-when-cr` (22 observations in Accept mode). This is the recommended default policy: it enhances privacy while maintaining some functionality.

2. **Persistent Risky Policy:** The `unsafe-url` policy is observed on 3 sites in all crawl modes. This policy is considered the most risky because it transmits the full URL, potentially exposing sensitive information to third parties.

3. **Impact of Restrictive Modes:** The `Crawl-Reject` (30 observations) and `Crawl-Block` (34 observations) modes show a lower total number of observations than the `Crawl-Accept` mode (46 observations).The reduction observed is due to the **blocking of third-party requests** in these modes.

12 Task 12

The objective of this task is to analyze the use of the `Accept-CH` header by websites, which allows them to request detailed information (Client Hints) about the user's device and software from the browser. This information, can be used for fingerprinting (unique user identification).

	Crawl-Accept	Crawl-Reject	Crawl-Block
0	Sec-CH-UA-Model: 28	Sec-CH-UA-Model: 20	Sec-CH-UA-Model: 13
1	Sec-CH-UA-Arch: 28	Sec-CH-UA-Arch: 20	Sec-CH-UA-Arch: 13
2	Sec-CH-UA-Full-Version-List: 27	Sec-CH-UA-Full-Version-List: 19	Sec-CH-UA-Full-Version-List: 12

Figure 21: Frequency of observed Client Hints requests

1. **Dominant Fingerprinting Information:** The three most requested hints aim to collect technical information about the device and browser: the **Device Model** (`Sec-CH-UA-Model`), the **CPU architecture** (`Sec-CH-UA-Arch`), and the **full list of versions** of the browser (`Sec-CH-UA-Full-Version-List`). This combined information is useful for **fingerprinting** and user identification.

2. **Dependence on Acceptance:** These hints are most frequently requested in `Crawl-Accept` mode. Sites clearly use this technique when they have obtained consent for tracking.

3. **Impact of Restrictive Modes:** The `Crawl-Reject` and `Crawl-Block` modes show a significant reduction in the number of sites requesting these hints.

    - The reduction in `Crawl-Reject` mode indicates that the consent management script, when refused, **removes the request** for Client Hints.
    - The reduction in `Crawl-Block` mode is even more pronounced. Blocking third-party resources prevents the loading of advertising and analytics scripts that are often responsible for issuing `Accept-CH` headers.

13 Task 13

The purpose of this analysis is to detect redirects between domains belonging to different entities.

Cross-entity redirects – Crawl-Accept

	Source domain	Source entity	Target domain	Target entity	Website
0	1rx.io	Nexxen	adsrvr.org	The Trade Desk	apnews.com
1	1rx.io	Nexxen	adsrvr.org	The Trade Desk	apnews.com
2	1rx.io	Nexxen	adsrvr.org	The Trade Desk	libero.it
3	1rx.io	Nexxen	adsrvr.org	The Trade Desk	nbcnews.com
4	1rx.io	Nexxen	adsrvr.org	The Trade Desk	newsweek.com
...	...	...	...	...	...
1034	zemanta.com	Outbrain	rubiconproject.com	Magnite	gazzetta.it
1035	zemanta.com	Outbrain	rubiconproject.com	Magnite	the-sun.com
1036	zemanta.com	Outbrain	seedtag.com	SeedTag	corriere.it
1037	zemanta.com	Outbrain	seedtag.com	SeedTag	telegraaf.nl
1038	zemanta.com	Outbrain	smartadserver.com	Equativ	libero.it

Figure 22: Redirection in accept mode

Cross-entity redirects – Crawl-Reject

	Source domain	Source entity	Target domain	Target entity	Website
0	admanmedia.com	AdmanMedia	criteo.com	Criteo	nbcnews.com
1	casalemedia.com	IndexExchange	doubleclick.net	Google	nbcnews.com
2	doubleclick.net	Google	criteo.com	Criteo	nbcnews.com
3	doubleclick.net	Google	demdex.net	Adobe	nbcnews.com
4	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
5	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
6	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
7	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
8	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
9	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
10	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk
11	doubleclick.net	Google	doubleverify.com	DoubleVerify	bbc.co.uk

Figure 23: Sample of redirects in reject mode (30 in total)

Cross-entity redirects – Crawl-Block

	Source domain	Source entity	Target domain	Target entity	Website
0	doubleclick.net	Google	adnxs.com	Microsoft	ad.nl
1	doubleclick.net	Google	demdex.net	Adobe	nbcnews.com
2	doubleclick.net	Google	doubleverify.com	DoubleVerify	ad.nl
3	yahoo.com	Yahoo!	360yield.com	Improve Digital	libero.it

Figure 24: Redirection in block mode

Analysis:

There is a **drastic drop** in the number of inter-entity redirects between acceptance mode and restrictive modes.

- In Crawl-Acceptance mode, a very high number of redirects (approximately **1038**) occur, confirming that this technique is used extensively when consent is given.
- In `Crawl-Reject` mode, and especially in `Crawl-Block` mode, the number of redirects drops to marginal levels (**30** and **3**), demonstrating that consent management and blocking mechanisms effectively target and disable most advertising redirect chains.

14 Task 14

This task analyzes the phenomenon of **CNAME Cloaking** (*CNAME masking*).

	First-party subdomain	CNAME record	CNAME entity	Disconnect categories
0	inessinsider.com	atcvt1h5rg5p.customer.permutive.app	Permutive	Advertising
1	plates.cnbc.com	d3qj5wuhimh3ra.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
2	metrics.cnn.com	cnn.com.102.122.2o7.net	Adobe	Advertising, Email, FingerprintingGeneral
3	orriere.corriere.it	d3mixdy9px9a4c.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
4	video.corriere.it	d1p15w45f27aqk.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
5	i-video.corriere.it	d24p96brdjmljd.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
6	ew.foxnews.com	:247preview.foxnews.com.akamaized.net	Take-Two	Analytics, FingerprintingGeneral, FingerprintingInvasive
7	ics.foxnews.com	foxnews.com.ssl.d1.sc.omtrdc.net	Adobe	Advertising, Email, FingerprintingGeneral
8	vice2.gazzetta.it	d2arvzuswg4o6w.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
9	2-mtc.gazzetta.it	d1fm7fuuekry8j.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
10	menti.gazzetta.it	d23yhnnqssijqq.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
11	video.gazzetta.it	d9kqvgumogq80.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
12	wst.lefigaro.fr	www.wysistat.com	Wysistat	Analytics
13	www.libero.it	d31d9gezsyt1z8.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
14	www.n-tv.de	d2tsqtep9kts9c.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
15	metrics.n-tv.de	971764.eu-central-1.elb.amazonaws.com	Amazon	Content
16	bilder.n-tv.de	d2tsqtep9kts9c.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
17	urcepoint.n-tv.de	cdn-1239.privacy-mgmt.com	Sourcepoint	ConsentManagers
18	nt.nbcnews.com	nbcnews.com.ssl.sc.omtrdc.net	Adobe	Advertising, Email, FingerprintingGeneral
19	i.newsweek.com	cdn-newsweek.tinypass.com	Piano	Analytics
20	i.newsweek.com	auth-newsweek.tinypass.com	Piano	Analytics
21	!.newsweek.com	c2-newsweek.tinypass.com	Piano	Analytics
22	:.newsweek.com	vx-newsweek.tinypass.com	Piano	Analytics
23	i.newsweek.com	d2examor23l10l.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
24	pbc.nypost.com	d126peu52a12md.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
25	dd.nytimes.com	sslgen-nyt.datadome.co	Datadome	Anti-fraud
26	nts.repubblica.it	d2gidjd0m1xg92.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
27	www.rtl.de	d120w0w1honsn7.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive
28	ourcepoint.rtl.de	cdn-1239.privacy-mgmt.com	Sourcepoint	ConsentManagers
29	cdn.the-sun.com	cdn-259.privacy-mgmt.com	Sourcepoint	ConsentManagers
30	www.wsj.com	dlp0y1mxy0v3u.cloudfront.net	dz7188oz6lnyb.cloudfront.net	Analytics, Content, FingerprintingInvasive

Figure 25: CNAME Cloaking

15 Task 15

15.1 What were the most challenging parts of the assignment?

The hardest part was automating cookie acceptance and blocking using Playwright. Every country uses different cookie consent designs—some use pop-up windows, others use banners or embedded frames. This made it difficult to write code that worked across all websites. Each region needed custom handling because there's no standard way websites ask for cookie consent

15.2 What were the findings that surprised you, if any?

Several things were unexpected. First, Playwright proved to be less developed than expected for handling complex cookie scenarios. The tool lacks built-in features for tracking cookie changes, requiring manual workarounds.

Second, we discovered lordofthefrogs.com listed as a data broker in the Disconnect blocklist.

Third, cookie consent varies significantly by country. Some UK and French websites force users to pay for a subscription if they want to block tracking cookies. These "pay or consent" walls charge money for ad-free browsing, which raises questions about whether consent is truly voluntary under GDPR rules.