Imperial College
London

Crawling the web to identify the perseverance of cookie banners and respect for choice

MSc Computing Individual Project Presentation

Philippe Paquin-Hirtle

Web Crawl…

- measuring cookie banner blocking capabilities;
- collecting privacy-related metrics (storage, requests, third-party);
- with the Google Chrome, Brave, Firefox, and Ghostery browsers;
- using Puppeteer (Node.js library);

Background

User Perception

Annoyance

Habituation

- 96% of banners do not offer a clear “Reject All” option
- When no “Reject All” option, 22% increase in consent
- Design Dark patterns (found in 57.4% of Consent Management Providers)

Cofone, I (2017); Kulyk et al. (2018); Sanchez-Rola et al. (2019); Utz et al. (2019), Nouwens et al. (2020)

GDPR Violations

Matte et al.’s Potential Violations

- No way to opt out (6.8%)
- Preselected choices (46.5%)
- Consent stored by default (9.9%)
- Non-respect of choice (5.3%)

Other violations

- Tracked before consent (and before appearance of cookie banner)
  - On 49% of websites (Trevisan et al.)
  - On 92% of websites (Sanchez-Rola et al.)
- 97.5% of websites do not remove cookies after user refusal (Sanchez-Rola et al.)

Utz et al. (2019); Sanchez-Rola et al. (2019); Trevisan et al. (2019); Matte et al. (2020);

All in all:

Users are still at a disadvantage

Tracking is present on a vast majority of websites

- Tracking before ability to opt-out
- Non-respect of opt-out

Cookie Banners make it harder to reject cookies than to accept them

- Use of dark patterns
- Annoyance and habituation

Implementation

Crawl Parameters

- **Vantage Point**: UK (residential IP) and US (VPN IP)
- **Website Selection**: Selected 10,000 (for UK). Top-1k + randomly selected sites from across the distribution (1k to 100k) using Tranco list
- **Timeouts**: 15 seconds to load page, 30 seconds per page loaded, 5 seconds per measurement
- **Temporal variations:** Parallelism used (3 instances in parallel). Up to one day of delay.
- **Bot Detection**: headful mode with stealth plugin

Related Work: Cookie Banner Detection Algorithm

- Use of CSS element names from ”I don’t care about cookies” (Kampanos and Shahandashti, and van Eijk et al.)
- Detecting Transparency and Consent Framework compliant banners (Matte et al.)
- Corpus-based detection (Rasaii et al.)

Van Eijk et al. (2019); Matte et al. (2020); Kampanos and Shahandashti (2021); Rasaii et al. (2023)

Cookie Banner Detection Algorithm: Main Steps

- Loops through all frames
- Loops through all elements of the frame, creating sub-trees of a max size or less
- Performs word search on elements in the sub-tree, using a corpus
  - Corpus created by analyzing top-50 banner terms
- Keep best candidate, and return the class names and ID of the elements
- Assess visibility using Puppeteer’s isVisible() method

Algorithm Accuracy (on top-250 websites)

	Detection Accuracy	False Positives	False Negatives
Google Chrome (n = 132)	87%	2	15

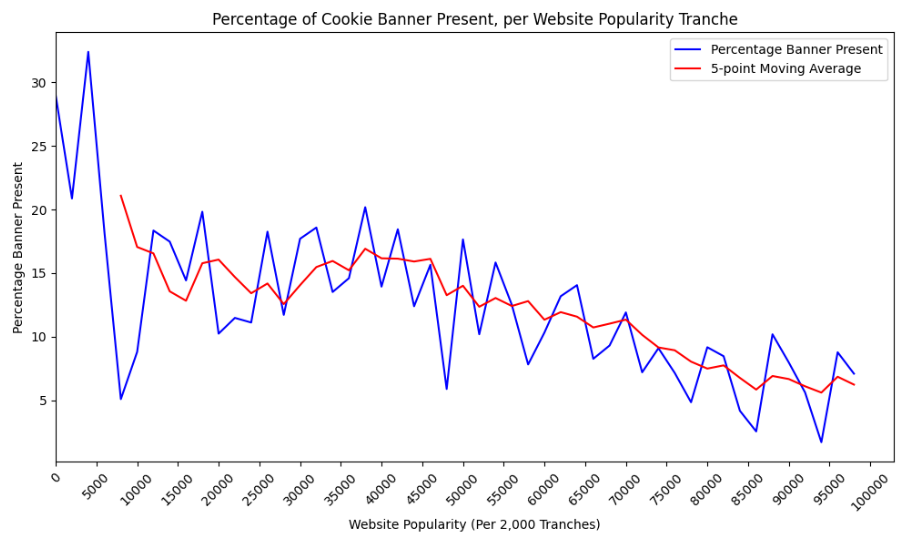
	Visibility Accuracy	False Positives	False Negatives
Brave (n = 150)	95%	4	3
Firefox (n = 102)	87%	1	12
Ghostery (n = 146)	95%	3	5
Google Chrome (n = 129)	84%	2	19

Information from: https://brave.com/shields/, https://blog.mozilla.org/en/products/firefox/firefox-rolls-out-total-cookie-protection-by-default-to-all-users-worldwide/, https://www.ghostery.com/faq#general

Results

Cookie Banner Detection by Algorithm

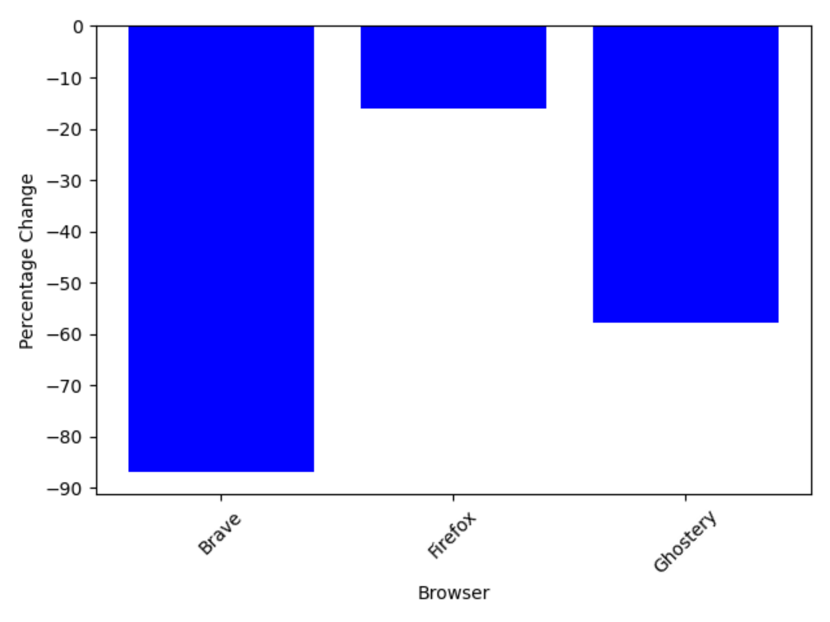
- **Algorithm detects cookie banners on 14.9% of websites**
- Underrepresentation of true value: only banners in English, and larger share of false negatives
- Declining presence, based on site popularity



Sample of 6358 websites, from Google Chrome browser

Cookie Banner Visibility

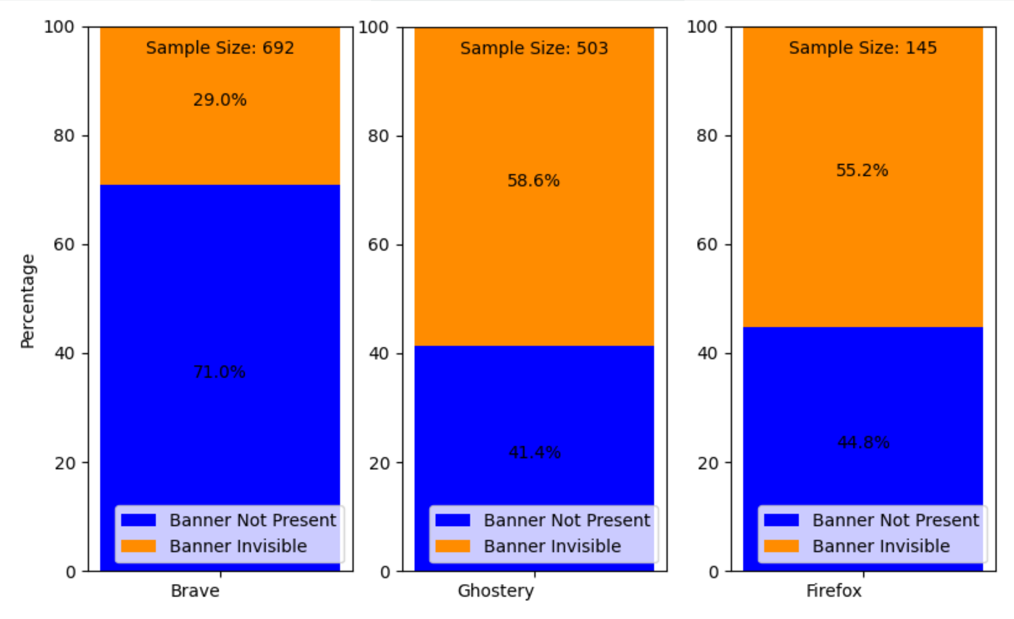
- **Brave displays 86.9% fewer cookie banners than Google Chrome.**
- Ghostery (-81% to -58%) and Firefox (-57% to -16%) results vary between test and crawl environment



Percentage Change in Cookie Banner Visibility Compared to Google Chrome
Sample of 3869 websites visited by all browsers.

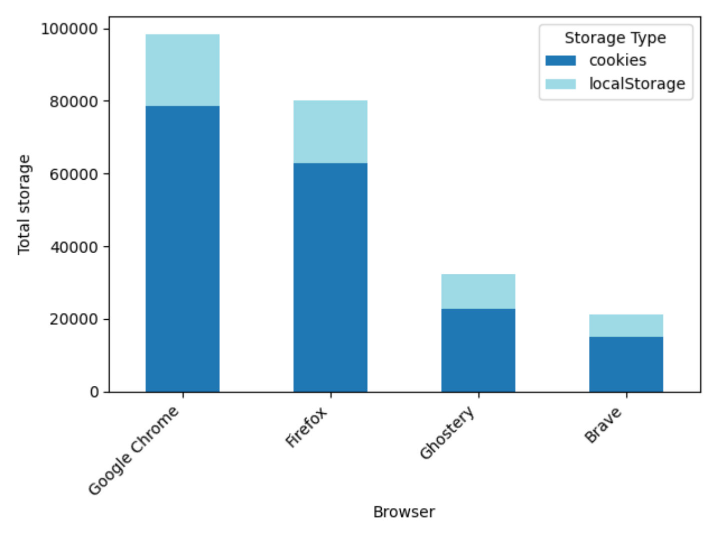
Blocking Techniques

- **Hidden state**: present, but not visible
- **Blocked state:** not-present in HTML, despite being present when visited using Google Chrome
- Brave is 29.6% more likely to be blocking a cookie banner, rather than hiding it, compared Ghostery and Firefox.

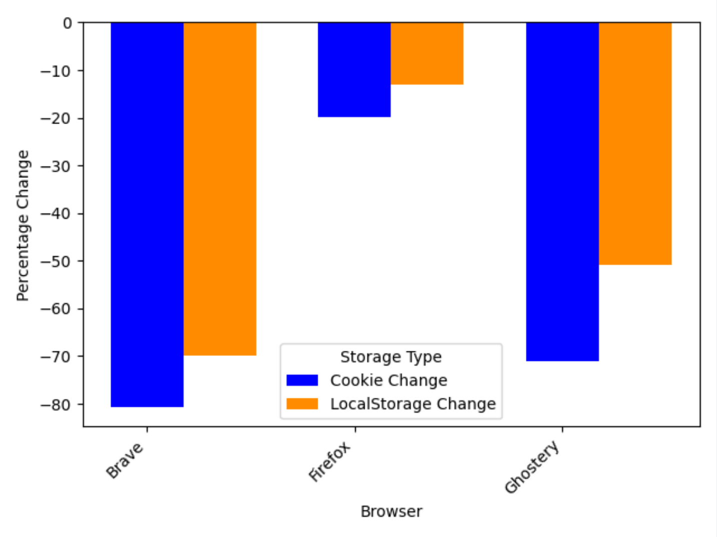


Proportion of Banner Hidden versus Blocked, per Browser
Varying sample size (see graph)

Total Storage

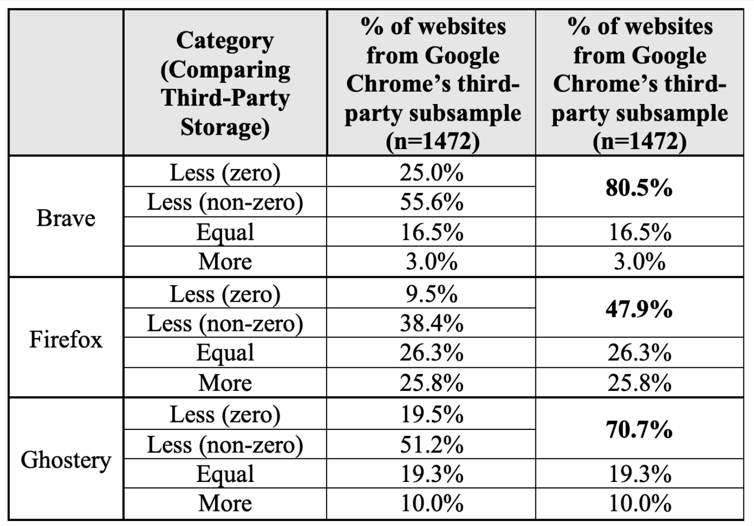


Total browser storage, per Browser
Sample size of 2,521 websites, visited by all browsers



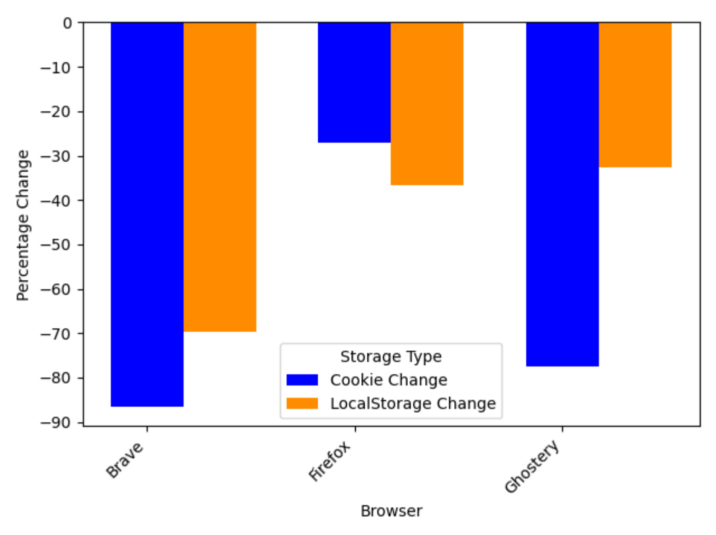
Total browser storage, per Browser
Sample size of 2,521 websites, visited by all browsers

Third-Party Storage Reduction



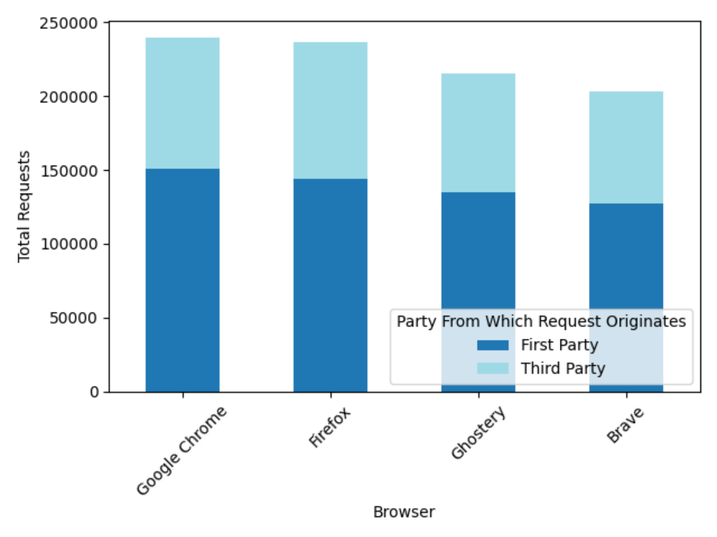
	Category (Comparing Third-Party Storage)	% of websites from Google Chrome's third-party subsample (n=1472)	% of websites from Google Chrome's third-party subsample (n=1472)
Brave	Less (zero)	25.0%	**80.5%**
	Less (non-zero)	55.6%	
	Equal	16.5%	16.5%
	More	3.0%	3.0%
Firefox	Less (zero)	9.5%	**47.9%**
	Less (non-zero)	38.4%	
	Equal	26.3%	26.3%
	More	25.8%	25.8%
Ghostery	Less (zero)	19.5%	**70.7%**
	Less (non-zero)	51.2%	
	Equal	19.3%	19.3%
	More	10.0%	10.0%

Classification of Websites, per Browser, Comparing The Number of Third-Party Storage Units to the Google Chrome Value

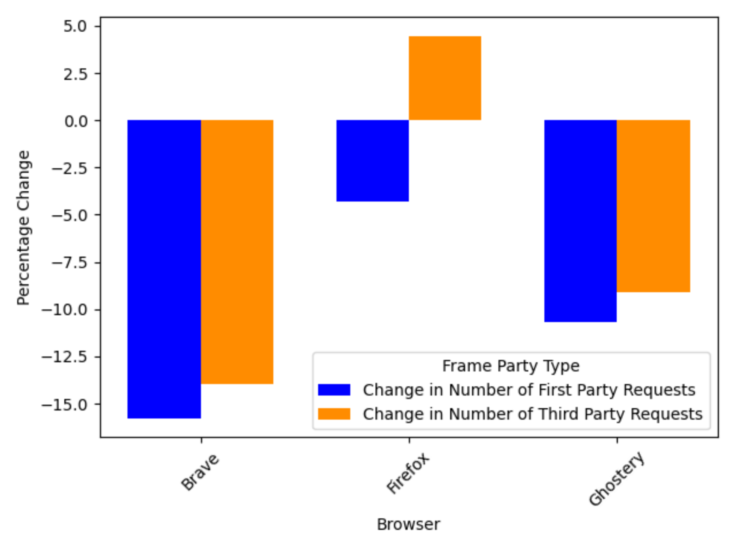


% Change in Third-Party Storage, Compared to Google Chrome, per Browser
Sample size of 1,472 websites (websites with third-party storage in Google Chrome)

Total Requests

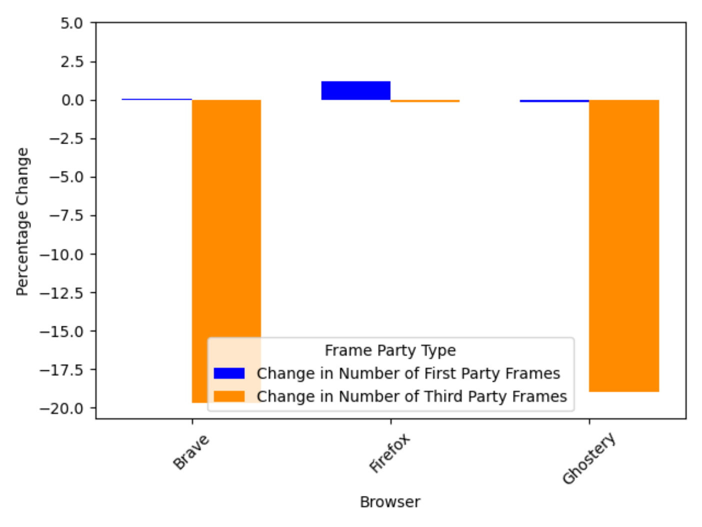


Total Number of Requests, per Party Type, per Browser
Sample size of 3,865 websites

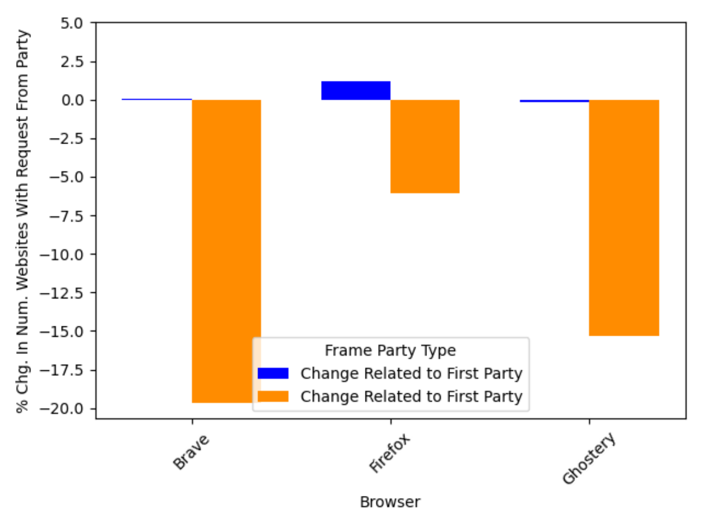


% change of the Number of Requests, per Party Type, Compared to Google Chrome

Third-Party Requests: Distinct Frames vs. Num. Websites

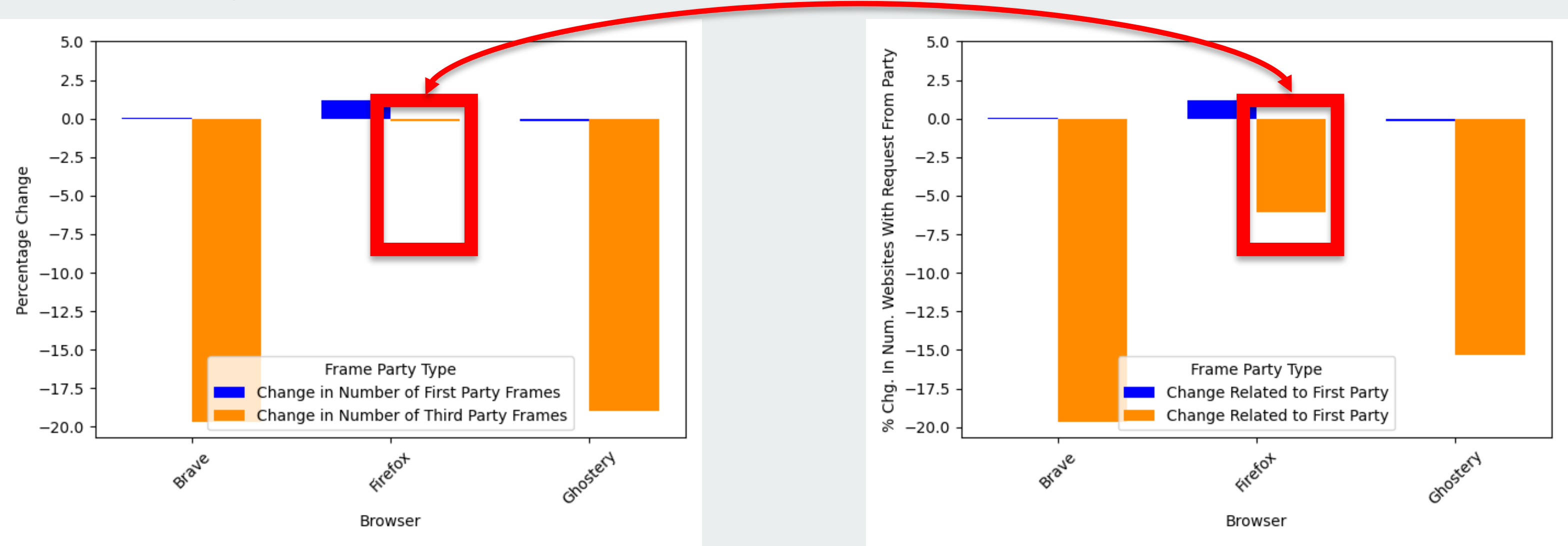


Percentage Change in **Number of Frames**, per Party Type, per Browser



Percentage Change in **Number of Websites** Registering Requests of a Certain Kind, per Browser

Third-Party Requests: Distinct Frames vs. Num. Websites

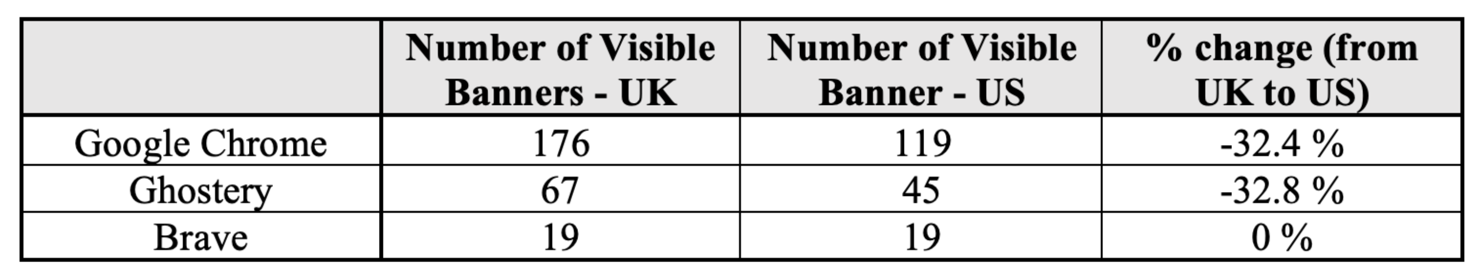


Percentage Change in **Number of Frames**, per Party Type, per Browser

Percentage Change in **Number of Websites** Registering Requests of a Certain Kind, per Browser

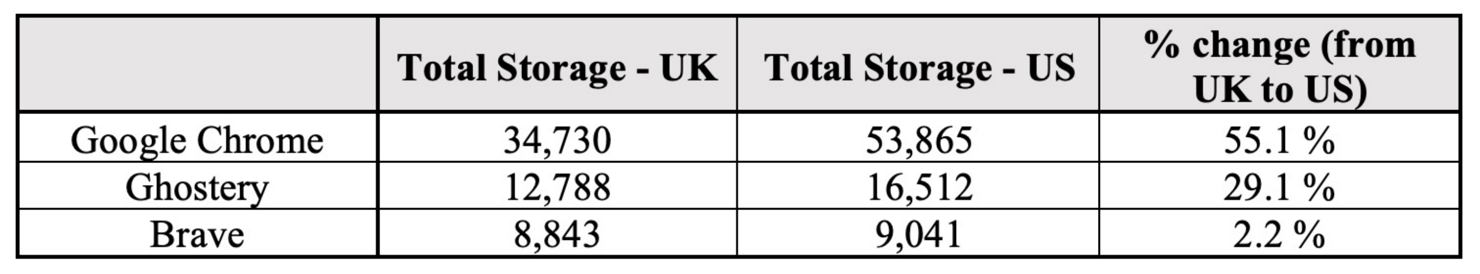
UK-US Comparison

Cookie Banner Visibility Comparison (Sample of 1643 websites)



	Number of Visible Banners - UK	Number of Visible Banner - US	% change (from UK to US)
Google Chrome	176	119	-32.4 %
Ghostery	67	45	-32.8 %
Brave	19	19	0 %

Total Storage Comparison (Sample of 1015 websites)

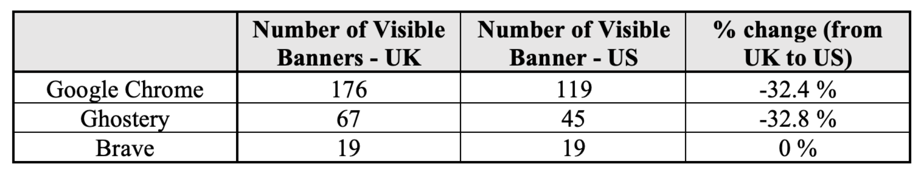


	Total Storage - UK	Total Storage - US	% change (from UK to US)
Google Chrome	34,730	53,865	55.1 %
Ghostery	12,788	16,512	29.1 %
Brave	8,843	9,041	2.2 %

UK-US Comparison

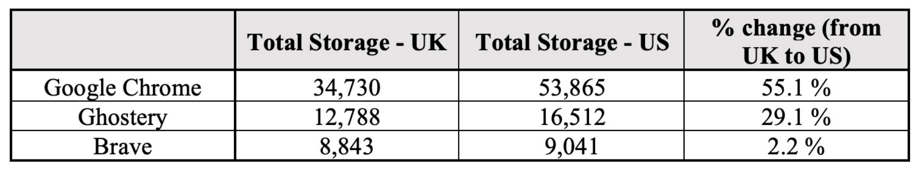
- Brave filtered out 99% of the extra storage from the US vantage point
- Third-party storage increase is larger for Google Chrome, but smaller for Ghostery and Brave

Cookie Banner Visibility Comparison (Sample of 1643 websites)



	Number of Visible Banners - UK	Number of Visible Banner - US	% change (from UK to US)
Google Chrome	176	119	-32.4 %
Ghostery	67	45	-32.8 %
Brave	19	19	0 %

Total Storage Comparison (Sample of 1015 websites)



	Total Storage - UK	Total Storage - US	% change (from UK to US)
Google Chrome	34,730	53,865	55.1 %
Ghostery	12,788	16,512	29.1 %
Brave	8,843	9,041	2.2 %

Evaluation and Conclusion

Evaluation

Limitations

- Use of a VPN IP address for US crawl
- Webdriver flag enabled for Firefox
- Lack of testing on websites in the rest of the distribution

Successes

- 87% banner detection accuracy and 84% to 95% banner visibility accuracy
- Resilient crawler (little to no human interactions required)
- Data and analysis answer project goal

Conclusion

Browser Comparison

- Brave performs better than both Ghostery and Firefox across all the metrics tested:
  - -86.9% cookie banners
  - -78.5% total storage
  - -15.0% number requests

Future Work

- Increase the number of browsers (or extensions) considered
- Crawl from a mobile device
- Isolate cookie banner blocking technique as a variable to see its impact on storage and requests