CookieGraph: Measuring and Countering First-Party Tracking Cookies

Shaoor Munir[1], Sandra Siby[2], Umar Iqbal[3], Steven Englehardt[4], Zubair Shafiq[1], Carmela Troncoso[2]

[1]Department of Computer Science, University of California, Davis
[2]Department of Computer Science, EPFL, Lausanne
[3]Department of Computer Science, University of Washington, Seattle
[4]Independent Researcher

Problem Statement

- Major browsers have already blocked or are in the process of blocking third-party cookies
  - Safari, Firefox, and Brave have already blocked third-party cookies
  - Google Chrome aims to block third-party cookies by 2024
- Advertisers and Trackers are looking for new ways to identify and track users
- We hypothesize that first-party cookies are being used in lieu of third-party cookies by trackers and advertisers
- If first-party cookies are being used for tracking, following questions arise:
  - What is the extent of use of first-party cookies by trackers and advertisers?
  - Is it possible to isolate and block first-party tracking cookies without causing significant website breakage?

Project Goals

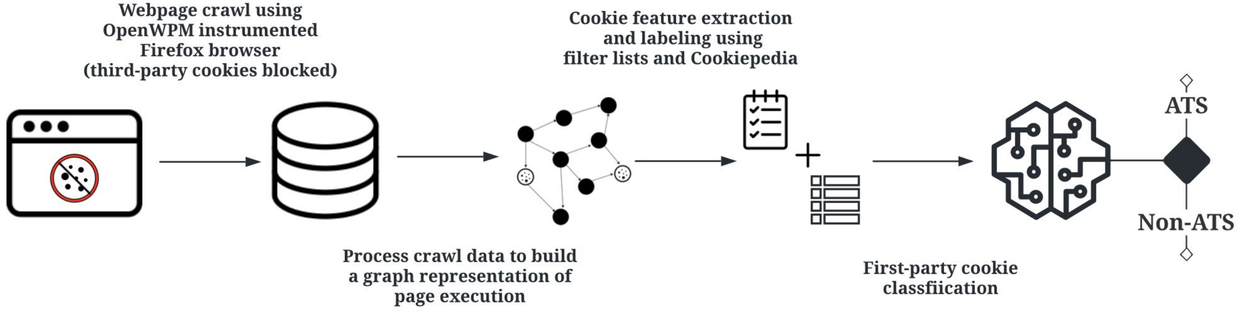
- Measure the usage of first-party cookies by trackers when third-party cookies are blocked
- Identify trackers using first-party cookies and understand the behaviors associated with that usage
- Create a threat model for cross-site tracking using the behavior of first-party tracking cookies
- Train a machine learning model to automatically detect first-party tracking cookies
- Evaluate the effectiveness of the model and its effect of website breakage

Methodology

- We crawl 10,000 websites twice – once with third-party cookies enabled, and once with third-party cookies blocked.
- This helped us measure difference between the two crawls and identify the trackers switching to first-party cookies when third-party cookies are blocked
- By studying behavior of these cookies, we identify key features and properties of first-party tracking cookies
- We identified presence of first-party tracking cookies on a majority of sites and discovered sharing of these cookies to other tracking domains

Cookie Name	Script Domain	Org.	Percentage of Sites	Destination Domains	Top-3 Destination Domains # 1	# 2	# 3
_gid	google-analytics.com	Google	77.11%	56	google-analytics.com	doubleclick.net	mountain.com
_ga	google-analytics.com	Google	68.88%	179	google-analytics.com	doubleclick.net	google.com
_fbp	facebook.net	Facebook	33.22%	73	facebook.com	appier.net	google-analytics.com
_gcl_au	googletagmanager.com	Google	14.22%	21	google.com	doubleclick.net	tealiumiq.com
__gpi	googlesyndication.com	Google	14.02%	4	doubleclick.net	googleadservices.com	ezoic.net
_ga	googletagmanager.com	Google	12.79%	48	google-analytics.com	doubleclick.net	google.com
__gads	googlesyndication.com	Google	12.35%	2	doubleclick.net	googleadservices.com	
__gads	doubleclick.net	Google	11.68%	11	doubleclick.net	googleadservices.com	ezoic.net
_uetsid	bing.com	Microsoft	10.22%	15	bing.com	hotjar.com	tealiumiq.com
_uetvid	bing.com	Microsoft	10.22%	21	bing.com	hotjar.com	tealiumiq.com
__gpi	doubleclick.net	Google	10.11%	10	doubleclick.net	googleadservices.com	ezoic.net
_clck	clarity.ms	Microsoft	8.81%	9	tealiumiq.com	drift.com	lmiutil.com
_hjTLDTest	hotjar.com	Hotjar	8.05%	1071	azercell.com	musinsa.com	google-analytics.com
_clsk	clarity.ms	Microsoft	7.88%	7	tealiumiq.com	drift.com	clicktripz.com
cto_bundle	criteo.net	Criteo	5.98%	7	criteo.com	fullstory.com	ezoic.net
_ym_d	yandex.ru	Yandex	4.85%	178	google-analytics.com	yandex.ru	doubleclick.net
_ym_uid	yandex.ru	Yandex	4.85%	48	yandex.ru	adfox.ru	google-analytics.co
_pin_unauth	pinimg.com	Pinterest	4.57%	7	tealiumiq.com	fullstory.com	azure.com
__utma	google-analytics.com	Google	4.32%	3	google-analytics.com	fullstory.com	ringostat.net
__utmb	google-analytics.com	Google	4.32%	5	google-analytics.com	piwik.pro	intellimize.co
__utmz	google-analytics.com	Google	4.32%	2	google-analytics.com	ringostat.net	
__qca	quantserve.com	Quantcast	4.19%	29	rubiconproject.com	yahoo.com	openx.net

- Using a ground truth generated by combining data from Cookiepedia (self-declaration of cookie purpose to comply with GDPR), and filter list labels of cookie setting scripts, we train a machine learning model to identify first-party tracking cookies



Countermeasure

- We train a machine learning algorithm on features computed by modeling interactions between different elements of a webpage
- CookieGraph relies on key features related to first-party tracking cookies, including exfiltration to different domains, infiltration of identifier values from trackers, use of other types of storage by setting script, etc.
- CookieGraph outperforms other methods of restricting cookies by at least 13 percent in accuracy and more than 20 percent in precision

Classifier	Accuracy	Precision	Recall
CookieGraph	91.06%	91.87%	90.59%
WebGraph	78.74%	71.59%	85.49%
CookieBlock	80.78%	69.95%	72.45%

Breakage Analysis

- Breakage analysis shows it is possible to isolate and block first-party tracking cookies
- CookieGraph outperforms existing countermeasures by only causing minor breakage on 2 percent of sites
- Other countermeasures result in major breakage on at least 10 percent of the sites

Classifier	Navigation		SSO		Appearance		Miscellaneous	
	Minor	Major	Minor	Major	Minor	Major	Minor	Major
CookieGraph	2%	0%	0%	0%	0%	0%	0%	0%
WebGraph	4%	2%	0%	6%	0%	2%	4%	10%
CookieBlock	0%	0%	0%	8%	0%	2%	0%	2%