Insights from a 1-million-site Measurement of Online Tracking

and how our data can help your research!

Steven Englehardt @s_englehardt Oillon Reisman @dillonthehuman Arvind Narayanan @random_walker





Visiting 2 websites results in 84 third parties contacted



Open Web Privacy Measurement (OpenWPM)

Code Issues	s 45 👔 Pull requests 0 👔	🗏 Projects o 💿 🗐 Wiki	● Unw	ratch - 49 Graphs	★ Unstar	435	8 Fork	67
A web privacy measur	rement framework https://webta	p.princeton.edu/ — Edit						
480 commits	₽ 4 branches	♡ 12 releases	13 ci	ontributors		s‡a GPL	-3.0	
Branch: master - New	w pull request		Create new file	Upload files	Find file	Clone	or downloa	d 🕶
🔛 englehardt Merge br	ranch 'master' of github.com:citp/Open	WPM		L	atest comm	it 3a1441	6 7 hours a	ago
automation	Added comments about ne	ew commands					15 days a	igo
test disabling audiocontext test for travis CI 15 days						15 days a	igo	
.gitignore Merge branch 'master' of github.com:citp/OpenWPM 10 month					months a	igo		
travis.yml Add travis.yml file to run continuous integration tests. 6 months					months a	igo		
CHANGELOG	CHANGELOG Version bump to 0.6.2. Bugfix in previous version 6 months					months a	igo	
	E LICENSE Removing extra whitespace from all infrastructure files 10 months					months a	igo	
README.md	README.md Modified readme to only use travis status from master branch 15 days ago					igo		

https://github.com/citp/OpenWPM

The Princeton Web Census

Monthly 1 Million Site Crawl

Javascript Calls All javascript file

- All javascript files
- HTTP Requests and Responses
- Storage (cookies, Flash, etc)

Collecting:

Insights from the Princeton Web Census







Trackers impede HTTPS adoption



Online Tracking: A 1-million-site Measurement and Analysis (CCS 2016)

Impact of OpenWPM and the Princeton Web Census

Open-sourcing early can help spur adoption

Study using OpenWPM	Conference	Year	
The Web Never Forgets: Persistent Tracking Mechanisms in the Wild	CCS	2014	
Cognitive disconnect: Understanding Facebook Connect login permissions	OSN	2014	
Cookies that give you away: The surveillance implications of web tracking	www	2015	
Upgrading HTTPS in midair: HSTS and key pinning in practice	NDSS	2015	
Web Privacy Census	Tech Science	2015	
Variations in Tracking in Relation to Geographic Location	W2SP	2015	
No Honor Among Thieves: A Large-Scale Analysis of Malicious Web Shells	WWW	2016	
Online Tracking: A 1-million-site Measurement and Analysis	CCS	2016	
Dial One for Scam: Analyzing and Detecting Technical Support Scams	NDSS	2017	

Open-sourcing early can help spur adoption

Study using OpenWPM	Conference	Year	
The Web Never Forgets: Persistent Tracking Mechanisms in the Wild	ccs	2014	
Cognitive disconnect: Understanding Facebook Connect login permissions	OSN	2014	
Cookies that give you away: The surveillance implications of web tracking	www	2015	
Upgrading HTTPS in midair: HSTS and key pinning in practice	NDSS	2015	
Web Privacy Census	Tech Science	2015	
Variations in Tracking in Relation to Geographic Location	W2SP	2015	
No Honor Among Thieves: A Large-Scale Analysis of Malicious Web Shells	WWW	2016	
Online Tracking: A 1-million-site Measurement and Analysis	ccs	2016	
Dial One for Scam: Analyzing and Detecting Technical Support Scams	NDSS	2017	

Measurement work can influence standards



https://w3c.github.io/fingerprinting-guidance

How can measurement influence adoption of new tracking techniques?



https://webtransparency.cs.princeton.edu/webcensus/

Canvas fingerprinting returns in the absence of measurement

May 2014: 5% of sites

Aug 2014: ~0.1% of sites

Jan 2016: 2.6% of sites

Percentage of the Alexa top 100k sites

AudioContext fingerprinting the Tor Browser

271 samples from the Tor Browsers

- 7 distinct fingerprints (2 fingerprints account for 80% of samples)
- Overlap with fingerprints from Firefox shows these largely reveal OS of device

	Manua Tintanta Danama	Course Deadman	Timeline	MCL	Casuah
	View Trakets Brows	e Source Roadmap	limeline	WIKI	Search
				← Previc	ous Ticket I
t13017 ass	igned task			Opened 2 ye	ears ago
				Last modifie	d 3 weeks a
Determine if A	AudioBuffers/OfflineAudio	Context are a fir	ngerprinting	y vector	
Reported by:	mikeperry	Owned by:	arthuredel	stein	
Priority:	Very High	Milestone:			
Component:	Applications/Tor Browser	Version:			
Severity:	Critical	Keywords:	tbb-finger TorBrowse	printing-os, rTeam20161	LO
Cc:	arthuredelstein, isis, mcs, brade	Actual Points:			
Parent ID:		Points:			
Reviewer:		Sponsor:			

Browsers remove BatteryStatus API citing privacy

•••	1313580 – Remove web content ac 🗙 🕂						
🗲 🕕 🌢 Mozilla Found	dation (US) https://bugzilla.mozilla.org/show_b	ug.cgi?id 🛛 🤇 🧐 Search	☆ 自 🔸 🕻	■ 🗢 🗢 🔳			
Bugzilla@Mozill	Bugzilla@Mozilla New Account Log In Forgot Password mozilla						
Home New Bro	Home New Browse Search Search [help] Reports Product Dashboard						
	Persona is no longer an option for authentication on BMO. For more details see Persona Deprecated.						
Bug 1313580 -	Remove web content access	to Battery API		Last Comment			
Status: Whiteboard:	VERIFIED FIXED	Reported:	2016-10-27 23:28 PDT by Chris [:cpeterson]	s Peterson			
Keywords:	addon-compat, dev-doc-needed, privacy, site-compat	Modified: CC List:	2016-11-02 09:53 PDT (History 7 users (show)))			
Product: Component:	Core (show info) DOM: Device Interfaces (show other bugs) (show info)	Flags: See Also:	ryanvm: in-testsuite-				
Version: Platform:	unspecified Unspecified Unspecified	Crash Signature:	(edit)				
Importance: Target Milestone:	normal (vote) mozilla52	QA Whiteboard: Iteration:					
Assigned To:	Chris Peterson [:cpeterson]	Points: Has Regression Range					

Browsers remove BatteryStatus API citing privacy

•••	1313580 – Remove web content ac 🗴	+											
🗧 🛈 🔒 Mozilla Foundation (US) https://bugzilla.mozilla.org/show_bug.cgi?id 🛛 C 🔍 Search 🖓 🖆 🔸 🧕 🤠 🔶 🖝													
Bugzilla@Mozilla			🥝 Bug 164213 – Remo	ove Battery Stat	< +								
Home New Browse Search		(← → (i) △)	ttps://bugs.webkit.org/s	show_bug.cgi?id=	164213		C,	Ct Search	Ľ	2 🗎 🦊	5	ļ 🔶	
			WebKit Bugzilla										
	Persona is no longer an option for a	1		Bug 16	4213: Remo	ve Battery	Status	API from the tree					
Bug 1313580 - Remove web content acces		Home New Bro	wse Search		Search	[?] Report	ts Rec	quests Help New /	Account	Log In F	orgot Pass	word	
		« First Last » « Prev Next » This bug is not in your last search results.											
Status: VERIFIED FIXED		Bug 164213 - Remove Battery Status API from the tree											
Whiteboard: Keywords: addon-compat, dev-doc-needed, privac	Sta	us: RESOLVED FIXE	D		Re	eporte	d: 2016-10-30 20:	26 PDT	by Brady	Eidson			
site-compat		Produ	t: WebKit			M	1odifie CC Lis	d: 2016-11-02 14: st: 8 users (<u>show</u>)	32 PDT	(<u>History</u>)			
Product:	Core (show info)	<u>V</u> ers	on: WebKit Nightly E	Build									
Component:	DOM: Device Interfaces (show other bug (show info)	Platfo	rm: Unspecified Unsp	pecified		Se	ee Also	<u>129040</u>					
Version:	unspecified	Importa Assigned	ce: P2 Normal	n									
Platform:	Unspecified Unspecified	Abbiglica											
Importance	normal (vote)	<u>UF</u> Keywo	<u>L:</u> de										
Target Milestone:	mozilla52	<u>Reywo</u>	<u>us</u> .										
Assigned To: Chris Peterson [:cpeterson]		Depends of Block	<u>n:</u> s:										
			Show dependent	cy <u>tree</u> / <u>graph</u>									

Blocking tools miss less popular trackers & fingerprinters



https://webtransparency.cs.princeton.edu/webcensus/

Future directions of the Web Census



Our data is available!

Data

The data is available as bzipped PostgreSQL dumps. The schema file used in all of the datasets is available here.

Dataset	Comments					
1 Million Site Stateless	Parallel Stateless Crawl					
100k Site Stateful	Parallel Stateful Crawl 10,000 site seed profile					
10k Site ID Detection (1)	Sequential Stateful Crawl Flash enabled Synced with ID Detection (2)					
10k Site ID Detection (2)	Sequential Stateful Crawl Flash enabled Synced with ID Detection (1)					
55k Site Stateless with cookie blocking	Parallel Stateless Crawl Firefox set to block all third-party cookies					
55k Site Stateless with Ghostery	Parallel Stateless Crawl Ghostery extension installed and set to block all possible trackers					
55k Site Stateless with HTTPS Everywhere	Parallel Stateless Crawl HTTPS Everywhere installed					

https://webtransparency.cs.princeton.edu/webcensus/index.html#data

Making data exploration easier

Problem: Querying our data — and making sense of the output — comes with a steep learning curve

```
query = "SELECT DISTINCT res.url, v1.name, v1.value FROM " \
    "http_responses_view as res " \
    "LEFT JOIN http_response_cookies_view as v1 " \
    "ON v1.response_id = res.id " \
    "WHERE res.top_url = %s AND v1.name != ''" \
    " union " \
    "SELECT DISTINCT req.url, v2.name, v2.value FROM " \
    "http_requests_view as req " \
    "LEFT JOIN http_request_cookies_view as v2 " \
    "ON v2.request_id = req.id " \
    "WHERE req.top url = %s AND v2.name != ''"
```

Making data exploration easier

Problem: Querying our data — and making sense of the output — comes with a steep learning curve



Solution: We are making our data analysis libraries available via Jupyter Notebook!

Census.py

- We'll give you access to our Notebook server, complete with tools that will provide an abstraction layer over our web census data.
- Example API:
 - o get_third_party_responses_by_domain(domain)
 - o get_cookie_syncs_on_domain(domain)

 - o get_trackers(domain)

Get access: https://groups.google.com/forum/#!forum/web-census-explorers

Example: Getting third party requests by domain

```
In [14]: results = census.get_third_party_responses_by_domain(con, 'http://nytimes.com')
third_party_trackers = {results[x]['url_ps'] for x in results if results[x]['is_tracker']}
print "Number of third_party trackers on domain: " + str(len(third_party_trackers))
for url in results:
    print url
    print '\tIs a script? ' + str(results[url]['is_js'])
    print '\tIs a tracker? ' + str(results[url]['is_tracker'])
    print '\tPS+1: ' + results[url]['url ps']
```

Get access: https://groups.google.com/forum/#!forum/web-census-explorers

Example: Getting third party requests by domain

```
In [14]: results = census.get_third_party_responses_by_domain(con, 'http://nytimes.com')
third_party_trackers = {results[x]['url_ps'] for x in results if results[x]['is_tracker']}
print "Number of third_party trackers on domain: " + str(len(third_party_trackers))
for url in results:
    print url
    print '\tIs a script? ' + str(results[url]['is_js'])
    print '\tIs a tracker? ' + str(results[url]['is_tracker'])
print '\tIs a tracker? ' + str(results[url]['is_tracker'])
print '\tPS+1: ' + results[url]['url_ps']
Number of third_party trackers on domain: 27
https://static01.nyt.com/images/2016/08/18/business/19WHEELS-ss-slide=0LQP/19WHEELS-ss-slide=0LQP-thumbStandard.jpg
    Is a script? False
        Is a tracker? False
        PS+1: nyt.com
        bttps://static01.nyt.com/images/2016/08/18/business/19wHEELS-ss-slide=0LQP/19wHEELS-ss-slide=0LQP-thumbStandard.jpg
        Is a tracker? False
        PS+1: nyt.com
        bttps://static01.nyt.com/images/2016/08/18/businesider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/02ignider_sequence1/0
```

```
Is a script? False
Is a tracker? False
PS+1: nyt.com
https://static01.nyt.com/images/2016/08/03/insider/03insider-savage01/03insider-savage01-thumbStandard-v2.jpg
Is a script? False
Is a tracker? False
PS+1: nyt.com
https://tpc.googlesyndication.com/simgad/15453271384116304559
Is a script? False
Is a tracker? True
PS+1: googlesyndication.com
http://googleads.g.doubleclick.net/pagead/gen_204?id=wfocus&gqid=&qqid=CMqE0Km81c4CFVVcDAodOMoOTQ&fg=1
Is a script? False
Is a tracker? True
```

Get access: https://groups.google.com/forum/#!forum/web-census-explorers

Thanks for listening!

Full Paper:

senglehardt.com/papers/ccs16_online_tracking.pdf

Princeton Web Census Data and Analysis:

webtransparency.cs.princeton.edu/webcensus/

Collaborate:

webtap.princeton.edu/research/

Email: ste@cs.princeton.eduTwitter: @s_englehardtWeb: senglehardt.comdreisman@princeton.edu

Image Assets from the Noun Project:

Robotic Arm by Creative Stall; Browser Network and Browser Battery by Aybige; Computer and Magnifying Glass by Edward Boatman; Server by Yazmin Alanis; Cookie by Rashida Luqman Kheriwala; JS File by Michael Finlay; Blocked Computer by arejoenah