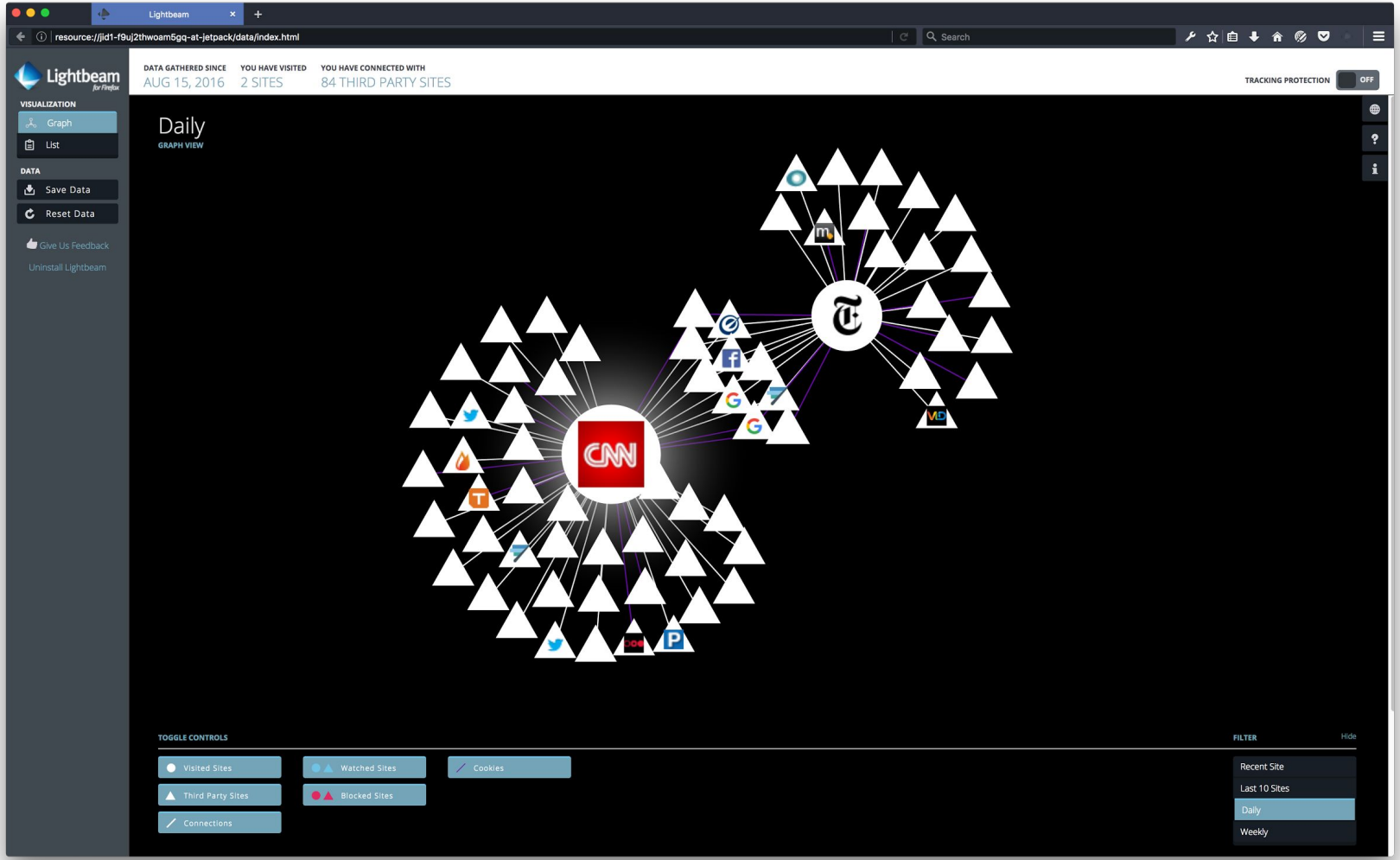
Web Privacy Through Transparency

A 1-million-site measurement and analysis

Steven Englehardt
@s_englehardt

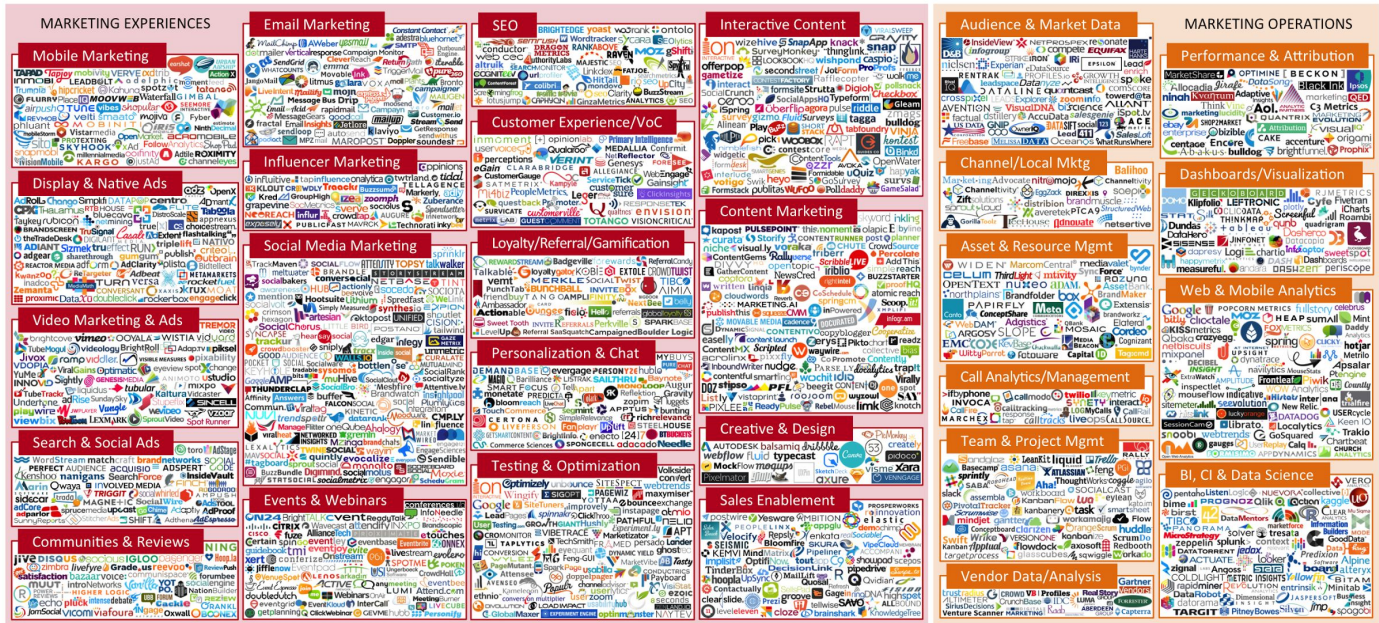
PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
PRINCETON UNIVERSITY

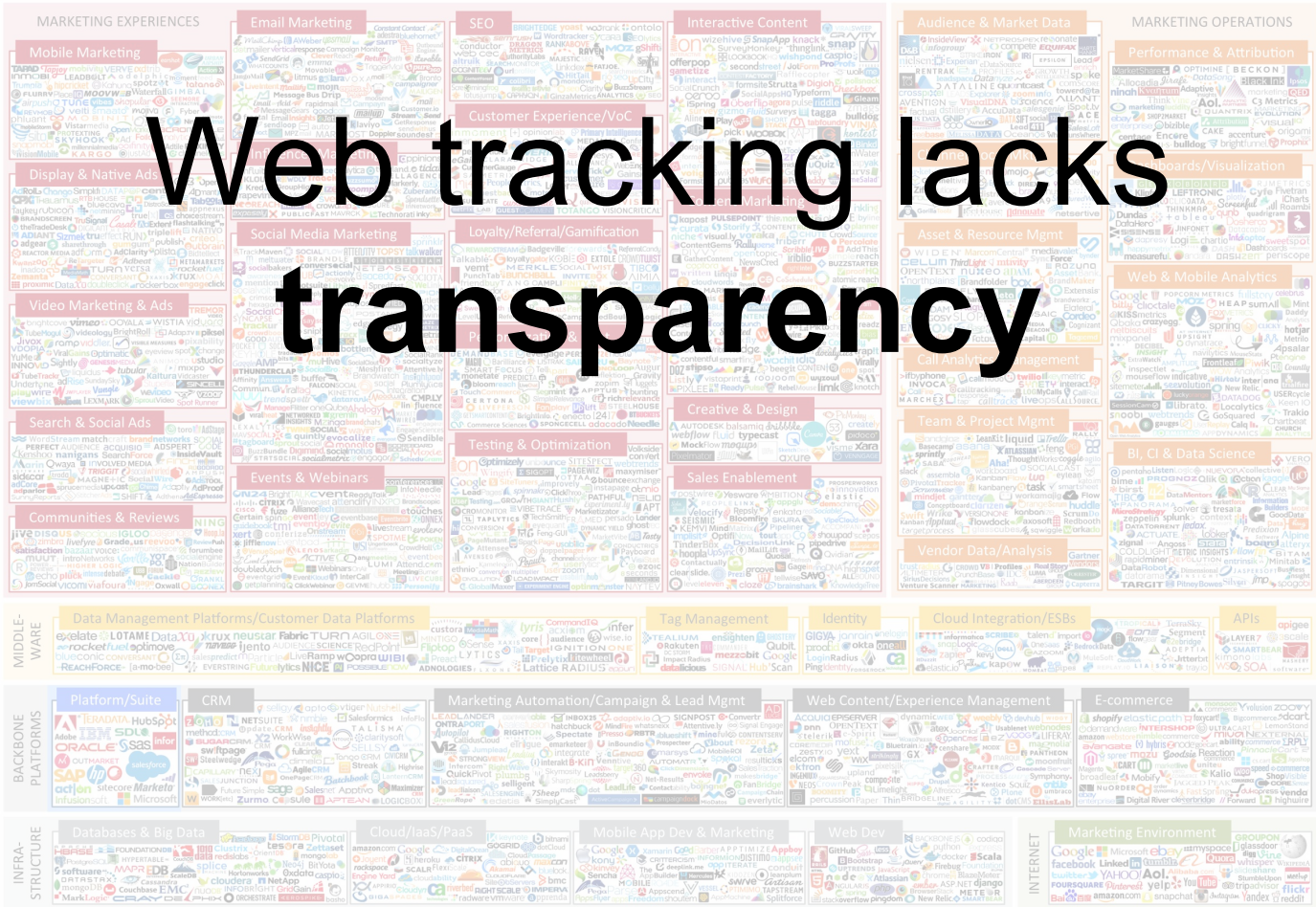

Lightbeam
resource://jid1-f9uj2thwoam5gq-at-jetpack/data/index.html
Search
Lightbeam for Firefox
VISUALIZATION
Graph
List
DATA
Save Data
Reset Data
Give Us Feedback
Uninstall Lightbeam
DATA GATHERED SINCE
AUG 15, 2016
YOU HAVE VISITED
2 SITES
YOU HAVE CONNECTED WITH
84 THIRD PARTY SITES
TRACKING PROTECTION
OFF
Daily
GRAPH VIEW
CNN
TOGGLE CONTROLS
Visited Sites
Third Party Sites
Connections
Watched Sites
Blocked Sites
Cookies
FILTER
Hide
Recent Site
Last 10 Sites
Daily
Weekly

chiefmartec.com Marketing Technology Landscape

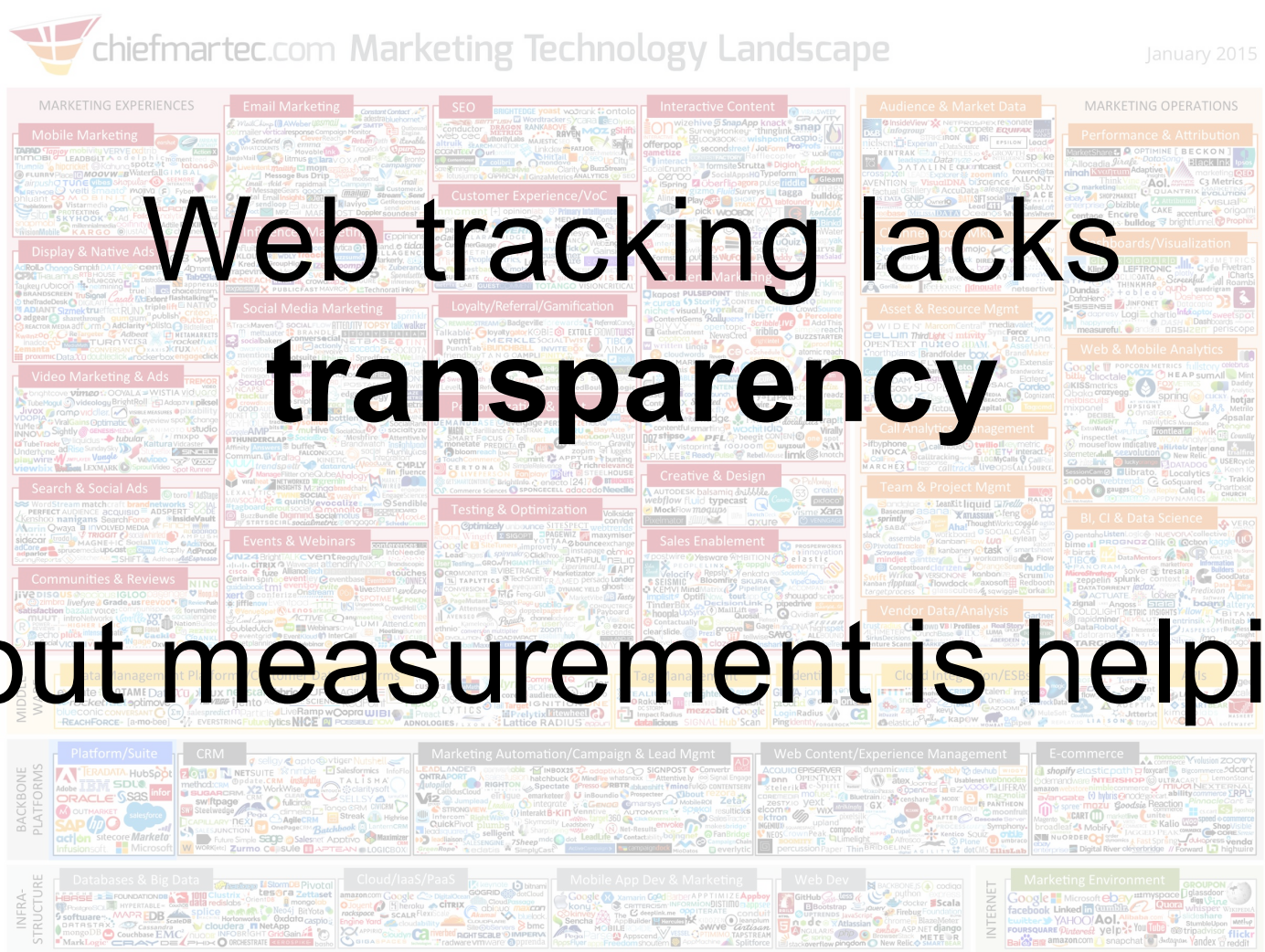
January 2015



by Scott Brinker @chiefmartec http://chiefmartec.com

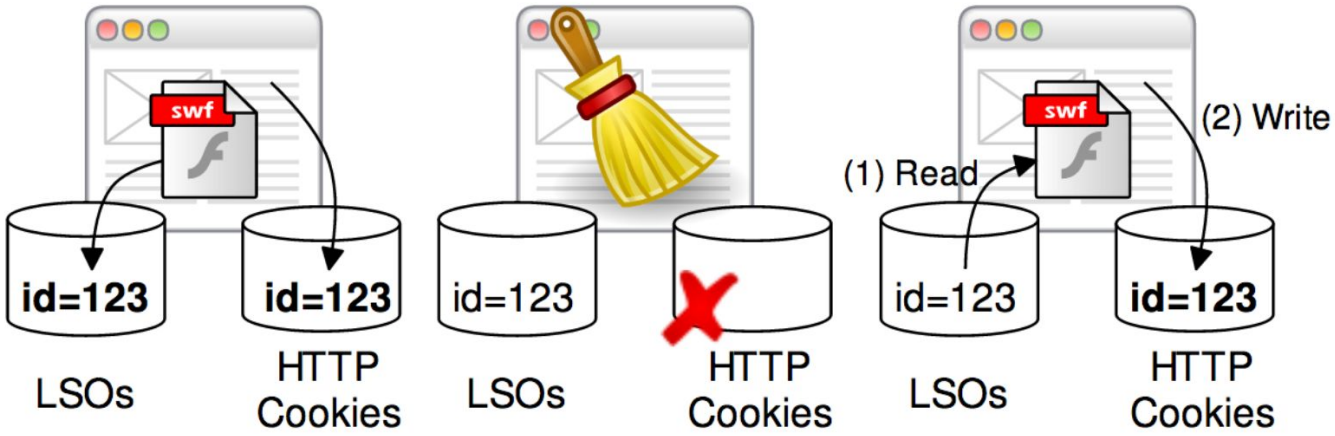


Web tracking lacks transparency



Web tracking lacks **transparency**

...but measurement is helping



Flash Cookies and Privacy (2009) Soltani, et al.
Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning (2011) Ayenson, et al.

EXTREMETECH Search Extremetech SEARCH

Computing Mobile Internet Gaming Electronics Extreme Deep Dives Deals

HOME › INTERNET › AOL, SPOTIFY, GIGAOM, ETSY, KISSMETRICS SUED OVER UNDELETABLE TRACKING COOKIES

AOL, Spotify, GigaOm, Etsy, KISSmetrics sued over undeletable tracking cookies

By Sebastian Anthony on August 4, 2011 at 7:07 am | 12 Comments

SAN FRANCISCO DIVISION

JOHN B. KIM, and DAN C. SCHUTZMAN, Individually, on Behalf of Themselves and All Others Similarly Situated,

Plaintiffs,

v.

SPACE PENCIL, INC. D/B/A KISSMETRICS, BABYPIPS.COM, INVOLVER.COM, MOO, INC., SITENING, LLC., SHOEDAZZLE.COM INC., 8TRACKS INC., ABOUT.ME, FRIENDLY, GIGA OMNI MEDIA INC., HASOFFERS.COM, KONGREGATE INC., LIVEMOCHA INC., ROCKETTHEME, LLC, FITNESS KEEPER, INC., SEOMOZ, INC., SHARECASH, LLC., SLIDESHARE.NET, SPOKEO, INC., SPOTIFY USA, INC., VISUALLY, CONDUIT USA, FLITE, INC.,

CV 11 3796

CASE NO.

JURY DEMAND

CLASS ACTION COMPLAINT FOR VIOLATIONS OF:

1. Electronic Communications Privacy Act, 18 U.S.C. § 2510;
2. Computer Crime Law, Cal. Penal Code § 502;
3. Trespass to Personal Property/Chattel; and
4. Unfair Competition Law, Cal. Bus. and Prof. Code § 17200.

Over the last few days a story has been developing about an undeletable tracking cookie used by KISSmetrics, a website analytics company. This company and more than 20 of its clients have now had a class action lawsuit filed against them. The plaintiffs claim that the Privacy Act and Electronics Communications Privacy Act have been broken, that their personal property (chattel) has been trespassed on, and that the defendants have violated unfair competition law. Anyone who has visited one of the defendants' sites is able to join the class action, and actual damages of up to $10,000 per member of the class are sought. If punitive damages are also awarded this lawsuit could be worth hundreds of millions of dollars.

Flash Cookies and Privacy (2009) Soltani, et al.
Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning (2011) Ayenson, et al.

EXTREMETECH

Search Extremetech SEARCH

Computing Mob

HOME › INTERNET › AC

AOL, Spot
undeletab

By Sebastian Anthony

JOHN B. KIM, and DAN C. SC
Individually, on Behalf of Them
Others Similarly Situated,

Plaintiffs,

v.

SPACE PENCIL, INC. D/B/A K
BABYPIPS.COM, INVOLVER
INC., SITENING, LLC., SHOE
INC., 8TRACKS INC., ABOUT
FRIENDLY, GIGA OMNI ME
HASOFFERS.COM, KONGRE
LIVEMOCHA INC., ROCKET
FITNESS KEEPER, INC., SEO
SHARECASH, LLC., SLIDESI
SPOKEO, INC., SPOTIFY USA
VISUALLY, CONDUIT USA,

Anyone who has vi
damages of up to $
this lawsuit could b

RYAN SINGEL BUSINESS 12.05.10 2:02 AM

ONLINE TRACKING FIRM SETTLES SUIT OVER UNDELETABLE COOKIES

Online tracking firm Quantcast has agreed to pay $2.4 million to settle a class action lawsuit alleging it secretly used Adobe's ubiquitous Flash plug-in to re-create tracking cookies after users deleted them, the company said Saturday.

More than $1 million of the settlement will go to fund privacy groups chosen by the plaintiffs, and 25% will go to the lawyers who filed the suit. It's unlikely that any money will go to the class, since it essentially includes every internet user in the U.S.

Flash Cookies and Privacy (2009) Soltani, et al.
Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning (2011) Ayenson, et al.

EXTREMETECH

Search Extremetech SEARCH

Computing Mob

HOME › INTERNET › AC

AOL, Spot
undeletab

By Sebastian Anthony

JOHN B. KIM, and DAN C. SC
Individually, on Behalf of Them
Others Similarly Situated,

Plaintiffs,

v.

SPACE PENCIL, INC. D/B/A K
BABYPIPS.COM, INVOLVER
INC., SITENING, LLC., SHOE
INC., 8TRACKS INC., ABOUT
FRIENDLY, GIGA OMNI MEI
HASOFFERS.COM, KONGRE
LIVEMOCHA INC., ROCKET
FITNESS KEEPER, INC., SEO
SHARECASH, LLC., SLIDESI
SPOKEO, INC., SPOTIFY USA
VISUALLY, CONDUIT USA,

Anyone who has vi
damages of up to $
this lawsuit could b

RYAN SINGEL BUSINESS 12.05.10 2:02 AM

ONLINE TRACKING FIRM SETTLES SUIT OVER UNDELET

MediaPost

ONLINE MEDIA DAILY

KISSmetrics Finalizes Supercookies Settlement

by Wendy Davis @wendyndavis, January 18, 2013, 5:24 PM

On
pay
lav
ubi
tra
the

Mc
go
pla
file
wil

Analytics company KISSmetrics has finalized the settlement of a class-action lawsuit stemming from its alleged use of "supercookies" to track people online.

The company implemented an agreement calling for it to refrain from using eTags, Flash cookies or other types of hard-to-delete supercookies without first notifying users and allowing them to choose whether to accept the technology, according to recent court papers.

The company also agreed to pay around $500,000 to the attorneys who brought the case and $2,500 each to the two consumers who sued: John Kim and Dan Schutzman.

Flash Cookies and Privacy (2009) Soltani, et al.
Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning (2011) Ayenson, et al.

Litigation is an effective deterrent

Global rank	Site	CC	Respawning (Flash) domain	1st/3rd Party
16	sina.com.cn	CN	simg.sinajs.cn	3rd*
17	yandex.ru	RU	kiks.yandex.ru	1st
27	weibo.com	CN	simg.sinajs.cn	3rd*
41	hao123.com	CN	ar.hao123.com	1st
52	sohu.com	CN	tv.sohu.com	1st
64	ifeng.com	HK	y3.ifengimg.com	3rd*
69	youku.com	CN	irs01.net	3rd
178	56.com	CN	irs01.net	3rd
196	letv.com	CN	irs01.net	3rd
197	tudou.com	CN	irs01.net	3rd

The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (CCS 2014): Acar et al.

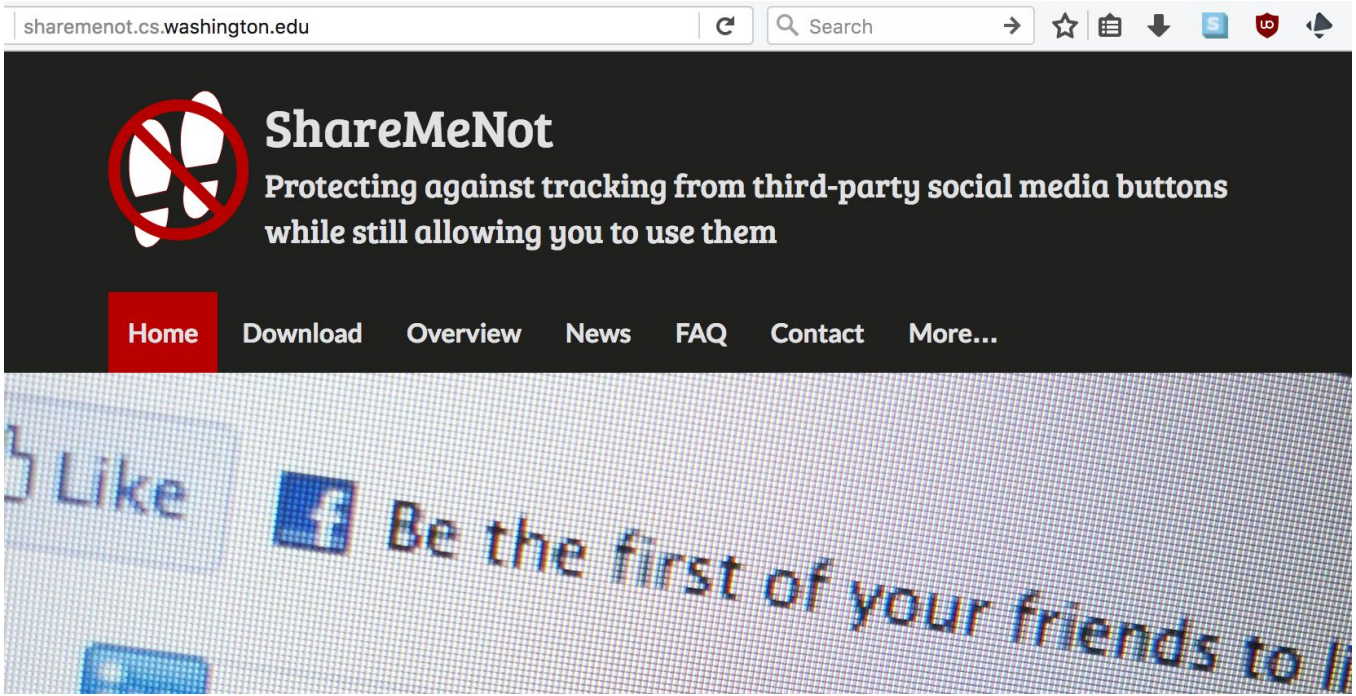
Empirical construction of tracker classification

Category	Name	Profile Scope	Summary	Example	Visit Directly?
A	Analytics	Within-Site	Serves as third-party analytics engine for sites.	Google Analytics	No
B	Vanilla	Cross-Site	Uses third-party storage to track users across sites.	Doubleclick	No
C	Forced	Cross-Site	Forces user to visit directly (e.g., via popup or redirect).	InsightExpress	Yes (forced)
D	Referred	Cross-Site	Relies on a B, C, or E tracker to leak unique identifiers.	Invite Media	No
E	Personal	Cross-Site	Visited directly by the user in other contexts.	Facebook	Yes

Table 1: *Classification of Tracking Behavior.* Trackers may exhibit multiple behaviors at once, with the exception of Behaviors B and E, which depend fundamentally on a user's browsing behavior: either the user visits the tracker's site directly or not.

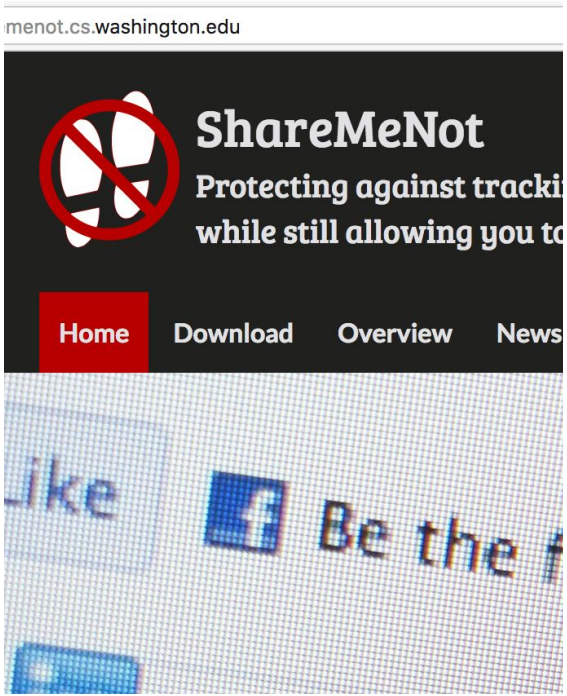
Detecting and Defending Against Third-Party Tracking on the Web (NDSI 2012): Roesner et al.

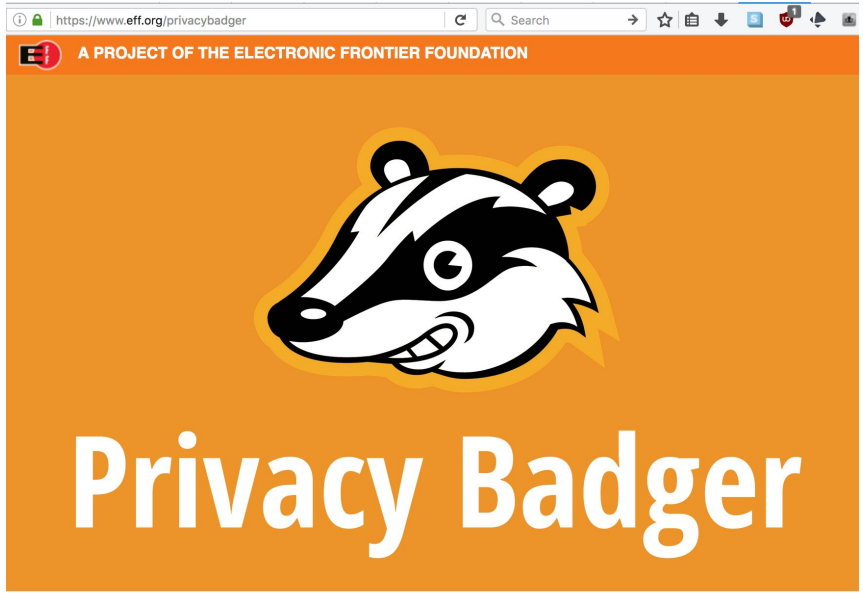
New class of trackers not effectively handled by block tools



Detecting and Defending Against Third-Party Tracking on the Web (NDSI 2012): Roesner et al.

New class of trackers not effectively handled by block tools





Privacy Badger blocks spying ads and invisible trackers.

Detecting and Defending Against Third-Party Tracking on the Web (NDSI 2012): Roesner et al.

Crying Wolf? On the Price Discrimination of Online Airline Tickets

Thomas Vissers[1], Nick Nikiforakis[1], Nataliia Bielova[2], and Wouter Joosen[1]

[1] iMinds-DistriNet, KU Leuven, 3001 Leuven, Belgium
{firstname.lastname}@cs.kuleuven.be,
[2] Inria, France
nataliia.bielova@inria.fr

Abstract. Price discrimination refers to the practice of dynamically varying the prices of goods based on a customer's purchasing power and willingness to pay. In this paper, motivated by several anecdotal accounts, we report on a three-week experiment, conducted in search of price discrimination in airline tickets. Despite presenting the companies with multiple opportunities for discriminating us, and contrary to our expectations, we do not find any evidence for systematic price discrimination. At the same time, we witness the highly volatile prices of certain airlines which make it hard to establish cause and effect. Finally, we

Transparency encourages best practices

Figure 7: Difference maps for a group on `text_arial`

Canvas
Fingerprinting
Introduced

May 2012

Transparency encourages best practices

Figure 7: Difference maps for a group on text_arial.

Canvas Fingerprinting Introduced

AddThis

LIGATUS

May 2014

May 2012

Transparency encourages best practices

Canvas Fingerprinting Introduced

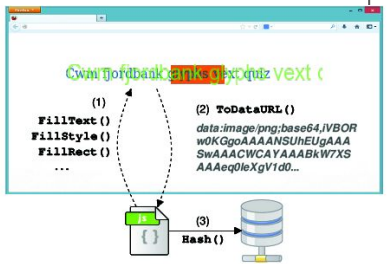
AddThis

LIGATUS

May 2012

May 2014

Canvas Fingerprinting Measured



The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (CCS 2014)

Transparency encourages best practices

Canvas Fingerprinting Introduced

AddThis

LIGATUS

News Coverage

PRO PUBLICA

BBC

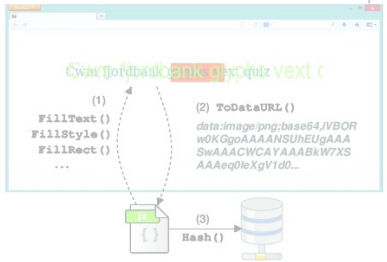
SPIEGEL ONLINE

May 2012

May 2014

July 21st 2014

Canvas Fingerprinting Measured



The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (CCS 2014)

Transparency encourages best practices

Canvas Fingerprinting Introduced

AddThis

LIGATUS

News Coverage

PRO PUBLICA

BBC

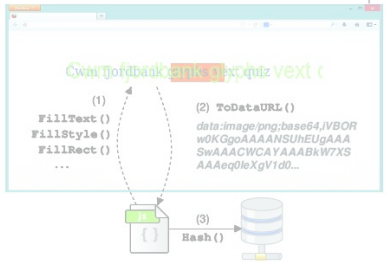
SPIEGEL ONLINE

May 2012

May 2014

July 23rd 2014

July 21st 2014



Canvas Fingerprinting Measured

AddThis

LIGATUS

Largest Fingerprinters Stopped

The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (CCS 2014)

Transparency is a necessary first step to return control to users and publishers

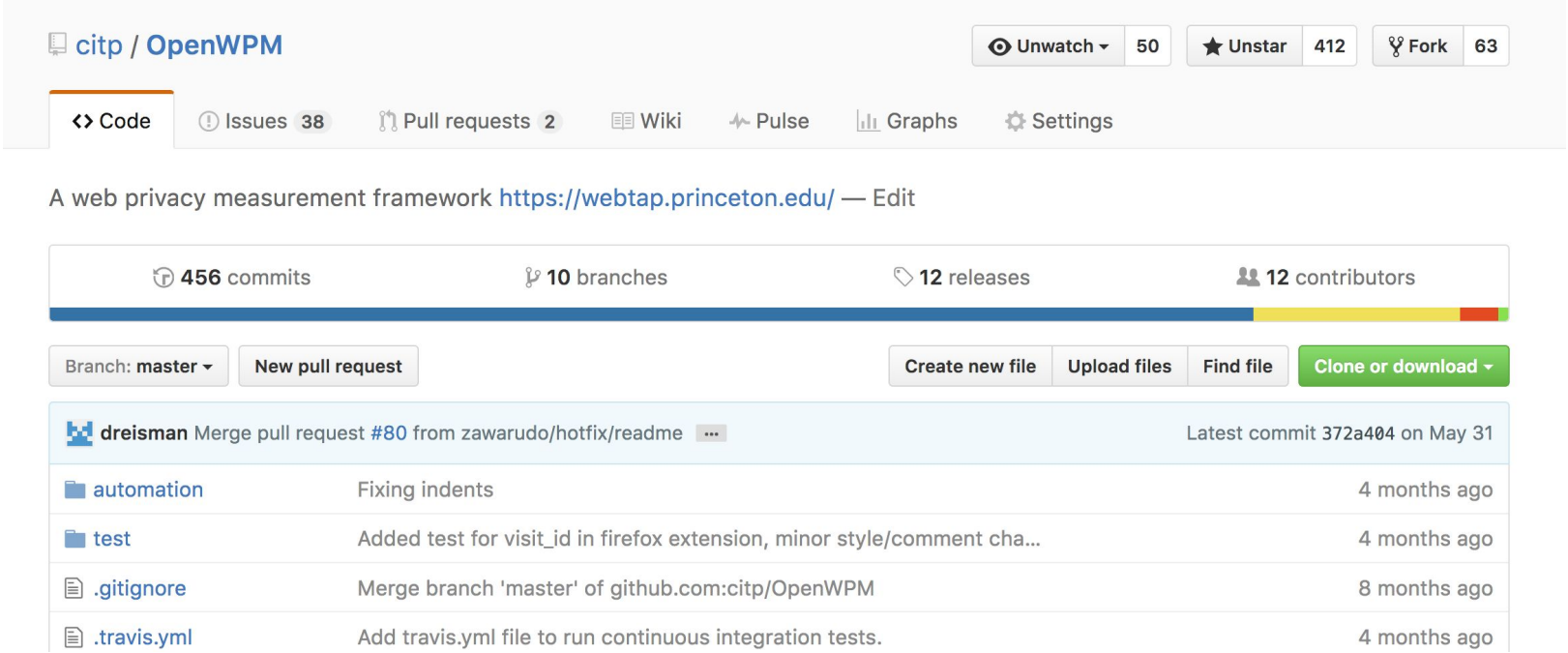
Automated, large-scale measurement is an essential part of the solution

	Targets	Infrastructure		Variable					
Paper		Automation[a]	Instrumentation	Crowd-sourced Distributed	Location	User-agent	Demographics	Interests	Privacy Tools	Scale
Leakage of PII via OSN ('09) [31]	PII leaks	M*	LHH							
Privacy diffusion on the web ('09) [30]	Tracking: cookies	F,PS	Proxy							1.2K sites
Challenges in measuring ('10) [25]	Personalization: ads		Proxy		•		•			730 queries
Flash cookies and privacy ('10) [53]	Tracking: cookies, LSOs	M*								100 sites
Privacy leakage in mOSN ('10) [32]	PII leaks	M*	Proxy							
Flash cookies and privacy II ('11) [10]	Tracking: cookies, LSOs	M*								100 sites
Privacy leakage vs. protection measures ('11) [29]	PII leaks	M*	Proxy							10 sites
Respawn HTTP Cookies ('11) [41]	Tracking: cookies, LSOs	UA*			•					600 sites
Self-help tools ('11) [38]	Tracking: cookies	UA*	FourthParty						•	500 sites
Where everybody knows your username ('11) [39]	PII leaks	M*	FourthParty				•			185 sites
Detecting and defending ('12) [52]	Tracking: cookies	FF, TT	TrackingTracker							2K sites
Detecting price and search discrimination ('12) [42]	Price discrimination	SA, CH, IE, JS	Proxy	•	•	•	•	•		200 sites
Mac users steered to pricier hotels ('12) [37]	Personalization: steering					•				
Measuring the effectiveness of privacy tools ('12) [11]	Personalization: ads	F, SL							•	
Websites vary prices ('12) [57]	Personalization: prices, deals				•					
What they do with what they know ('12) [60]	Personalization: ads		Proxy							10 days
AdReveal ('13) [34]	Personalization: ads		Proxy, Ghostery				•			103K sites
Cookieless monster ('13) [47]	Tracking: fingerprinting									10K sites
Crowd-assisted search ('13) [43]	Price discrimination	F, CH	Custom plugin	•	•	•		•		600 sites
Discrimination in online ad delivery ('13) [54]	Ads	M, UA			•	•				2184 names
FPDetective ('13) [7]	Tracking: fingerprinting, JS	CR, SL, CJ, PJ	Proxy, Browser Code							1M sites
Know your personalization ('13) [35]	Personalization: search		Custom plugin	•			•			5K queries
Measuring personalization of web search ('13) [26]	Personalization: search	PJ		•		•				120 queries
Who knows what about me? ('13) [36]	PII leaks	F, PS, SL		•			•		•	1.5K sites
Selling off privacy at auction ('13) [49]	Cookie sync, bid prices	F, SL		•	•			•		5K sites
Shining the floodlights ('13) [19]	Tracking: cookies, JS	F, JS	FourthParty			•				500 sites
Statistical approach ('13) [22]	General tracking	F, PY	FourthParty						•	2K sites
Adscape ('14) [13]	Personalization: ads	F, SL	Custom plugin				•			10K sites
Bobble ('14) [61]	Personalization: search	CH, SL	Custom plugin	•	•	•				1K queries
Information flow experiments ('14) [56]	Personalization: ads	F, SL	Proxy				•			
Third-party OSN applications ('14) [14]	PII leaks	F, SL	FourthParty				•			997 apps
Price discrimination and steering ('14) [27]	Price disc, steering	PJ		•	•	•	•	•		16 sites
Price discrimination of airline tickets ('14) [59]	Price discrimination	CJ			•	•	•	•		21 days

[a]FF = Firefox, CH = Chrome, CR = Chromium, IE = Internet Explorer, SA = Safari, SL = Selenium, JS = JavaScript, PJ = PhantomJS, PS = PageStats, PY = Python, TT = TrackingTracker, CJ = CasperJS, UA = Unknown automation, M = manual, LHH = Live HTTP Headers, Asterisk = inferred

A need for a common platform

- Constant re-engineering of similar measurement tools
- Methodological differences
  - PhantomJS vs Firefox vs Chrome
- High cost to reproduce or re-measure
  - Studies are only run once



https://github.com/citp/OpenWPM

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	[Working Paper]	2016

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	[Working Paper]	2016

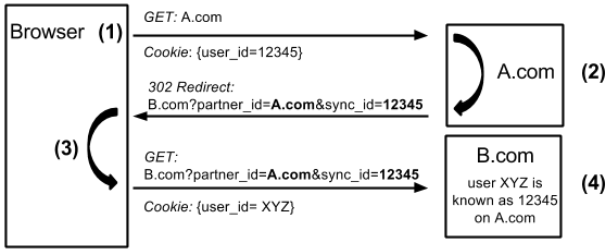
Measuring Stateful Tracking

id	crawl_id	header_id	name	value	accessed
3515	1	1819	DSID	NO_DATA	2016-08-27 14:10:47.925556
3516	1	1819	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.925556
3511	1	1818	IDE	AHWqTUmNrKKTjY3MUYIiAN6dYINl37RtBRZ1er6nJfA4WU1htrkk8luRPA	2016-08-27 14:10:47.925319
3512	1	1818	DSID	NO_DATA	2016-08-27 14:10:47.925319
3513	1	1818	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.925319
3508	1	1817	IDE	AHWqTUmNrKKTjY3MUYIiAN6dYINl37RtBRZ1er6nJfA4WU1htrkk8luRPA	2016-08-27 14:10:47.916564
3509	1	1817	DSID	NO_DATA	2016-08-27 14:10:47.916564
3510	1	1817	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.916564
3505	1	1816	IDE	AHWqTUmNrKKTjY3MUYIiAN6dYINl37RtBRZ1er6nJfA4WU1htrkk8luRPA	2016-08-27 14:10:47.890402
3506	1	1816	DSID	NO_DATA	2016-08-27 14:10:47.890402
3507	1	1816	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.890402
3503	3	1814	_ga	GA1.2.1851119688.1472321405	2016-08-27 14:10:47.854893
3504	3	1814	_gat_memega	1	2016-08-27 14:10:47.854893
3501	3	1813	_ga	GA1.2.1851119688.1472321405	2016-08-27 14:10:47.845277
3502	3	1813	_gat_memega	1	2016-08-27 14:10:47.845277
3499	3	1812	_ga	GA1.2.1851119688.1472321405	2016-08-27 14:10:47.833771
3500	3	1812	_gat_memega	1	2016-08-27 14:10:47.833771
3497	3	1811	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:47.693007
3498	3	1811	tvrg_60296	"2,1472321409"	2016-08-27 14:10:47.693007
3496	3	1810	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:47.613696
3495	3	1809	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:47.595839
3493	3	1808	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:47.532022
3494	3	1808	tvrg_60296	"2,1472321409"	2016-08-27 14:10:47.532022
3492	3	1806	ymvw	yidf42s99y6ezkIPvWjEIY92AJsx	2016-08-27 14:10:47.397447
3491	1	1802	uuid	"b06173e8-7332-43fd-b1f0-6eb415a2e0dc-20160827 14:10:46"	2016-08-27 14:10:47.237651
3490	3	1801	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:47.220842
3488	3	1800	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:47.219887
3489	3	1800	tvrg_60296	"2,1472321409"	2016-08-27 14:10:47.219887
3487	1	1798	uuid	"b06173e8-7332-43fd-b1f0-6eb415a2e0dc-20160827 14:10:46"	2016-08-27 14:10:47.171226
3485	3	1796	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:46.839196
3486	3	1796	tvrg_60296	"2,1472321409"	2016-08-27 14:10:46.839196
3484	3	1792	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:46.643236
3482	3	1790	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:46.627712

Measuring Stateful Tracking

id	crawl_id	header_id	name	value	accessed
3515	1	1819	DSID	NO_DATA	2016-08-27 14:10:47.925556
3516	1	1819	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.925556
3511	1	1818	IDE	AHWqTUmNrKKTjY3MUYIiAN6dYINl37RtBRZ1er6nJfA4WU1htrkk8luRPA	2016-08-27 14:10:47.925319
3512	1	1818	DSID	NO_DATA	2016-08-27 14:10:47.925319
3513	1	1818	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.925319
3508	1	1817	IDE	AHWqTUmNrKKTjY3MUYIiAN6dYINl37RtBRZ1er6nJfA4WU1htrkk8luRPA	2016-08-27 14:10:47.916564
3509	1	1817	DSID	NO_DATA	2016-08-27 14:10:47.916564
3510	1	1817	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.916564
3505	1	1816	IDE	AHWqTUmNrKKTjY3MUYIiAN6dYINl37RtBRZ1er6nJfA4WU1htrkk8luRPA	2016-08-27 14:10:47.890402
3506	1	1816	DSID	NO_DATA	2016-08-27 14:10:47.890402
3507	1	1816	id	22af8c8cf20a00b8||t=1472321413|et=730|cs=002213fd4883c26574091b4ac4	2016-08-27 14:10:47.890402
3503	3	1814	_ga	GA1.2.1851119688.1472321405	2016-08-27 14:10:47.854893
3504	3	1814	_gat_memega	1	2016-08-27 14:10:47.854893
3501	3	1813	_ga	GA1.2.1851119688.1472321405	2016-08-27 14:10:47.845277
3502	3	1813	_gat_memega	1	2016-08-27 14:10:47.845277
3499	3	1812	_ga	GA1.2.1851119688.1472321405	2016-08-27 14:10:47.833771
3500	3	1812	_gat_memega	1	2016-08-27 14:10:47.833771
3497	3	1811	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:47.693007
3498	3	1811	tvrg_60296	"2,1472321409"	2016-08-27 14:10:47.693007
3496	3	1810	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:47.613696
3495	3	1809	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:47.595839
3493	3	1808	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:47.532022
3494	3	1808	tvrg_60296	"2,1472321409"	2016-08-27 14:10:47.532022
3492	3	1806	ymvw	yidf42s99y6ezkIPvWjEIY92AJsx	2016-08-27 14:10:47.397447
3491	1	1802	uuid	"b06173e8-7332-43fd-b1f0-6eb415a2e0dc-20160827 14:10:46"	2016-08-27 14:10:47.237651
3490	3	1801	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:47.220842
3488	3	1800	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:47.219887
3489	3	1800	tvrg_60296	"2,1472321409"	2016-08-27 14:10:47.219887
3487	1	1798	uuid	"b06173e8-7332-43fd-b1f0-6eb415a2e0dc-20160827 14:10:46"	2016-08-27 14:10:47.171226
3485	3	1796	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:46.839196
3486	3	1796	tvrg_60296	"2,1472321409"	2016-08-27 14:10:46.839196
3484	3	1792	uuid	"5540526e-9469-4f73-90f9-74280df3ac76-20160827 14:10:23"	2016-08-27 14:10:46.643236
3482	3	1790	tvid	fb71f68e4ba64548a5488ee248957066	2016-08-27 14:10:46.627712

Measuring Stateful Tracking



Cookie Syncing

Cookie Respawning

Measuring (Active) Stateless Tracking

- Custom Firefox Extension
- Log method calls and property access
  - Overwrite getters and setters
  - Resistant to tampering
- Easily ported to Chrome extension or used with Tor Browser

Transparency through Measurement

- **The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (CCS 2014)**

  Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, Claudia Diaz

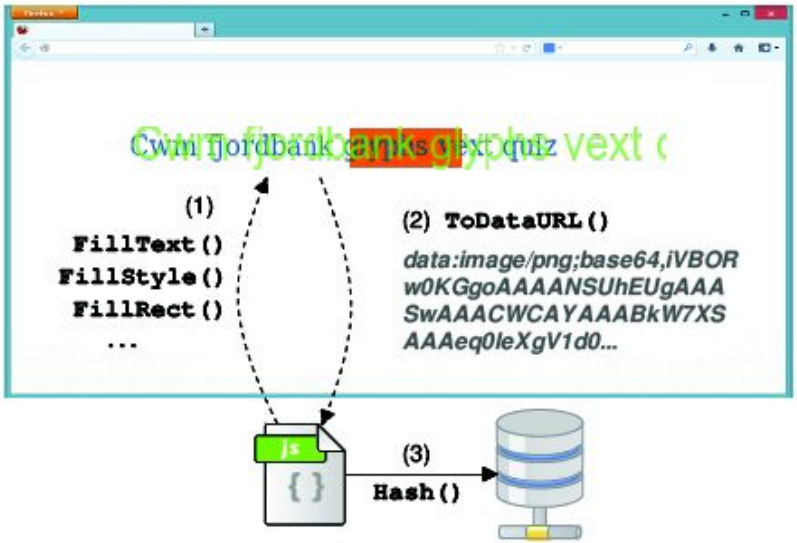
- **Cookies That Give You Away: The Surveillance Implications of Web Tracking (WWW 2015)**

  Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, Edward Felten

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

  Steven Englehardt and Arvind Narayanan

Without legal precedence, effects of press coverage of canvas fingerprinting were temporary



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

Canvas Fingerprinting

Figure 6: 13 ways to render 20px Arial

Source: Mowery and Shacham; Pixel Perfect: Fingerprinting Canvas in HTML5

Canvas Fingerprinting

Figure 6: 13 ways to render 20px Arial

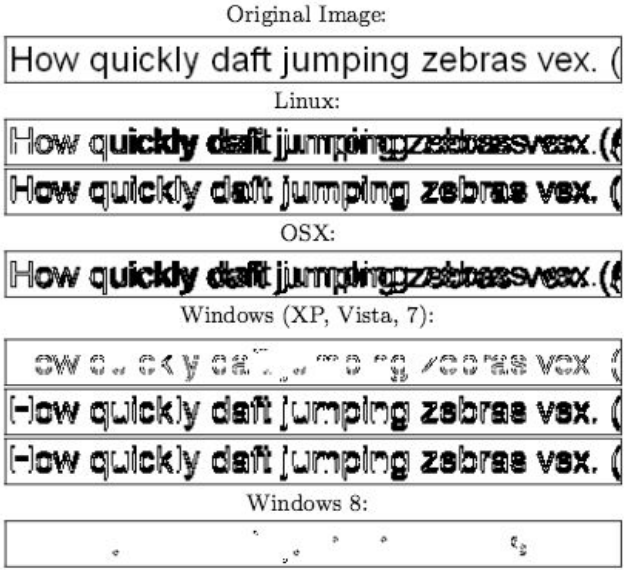


Figure 7: Difference maps for a group on `text_arial`

Source: Mowery and Shacham; Pixel Perfect: Fingerprinting Canvas in HTML5

Detection Methodology:

1. Canvas height and width >= 16px
2. Text >= 2 colors OR >= 10 characters
3. Should not call `save`, `restore`, or `addEventListener`. (Used with interactive or animated content)
4. Calls `toDataURL` or `getImageData`.

Canvas fingerprinting returns

May 2014: 5% of sites

Aug 2014: ~0.1% of sites

Jan 2016: 2.6% of sites

Canvas fingerprinting returns

May 2014: 5% of sites

Aug 2014: ~0.1% of sites

Jan 2016: 2.6% of sites

→ **Shift towards fraud detection**

Canvas: *Providing multiple ways to fingerprint since HTML5*

Font Fingerprinting Method:

1. Create a canvas and set the `font` property
2. Print some text to canvas
3. Use `context.measureText()` to determine width and height
4. If those don't match a fallback font, the user has the font installed

Image source: http://www.lalit.org/lab/javascript-css-font-detect/

Canvas: *Providing multiple ways to fingerprint since HTML5*

Font Name	Detected?
cursive	true
monospace	true
serif	true
sans-serif	true
fantasy	true
default	true
Arial	true
Arial Black	true
Arial Narrow	true
Arial Rounded MT Bold	true
Bookman Old Style	false
Bradley Hand ITC	false
Century	false
Century Gothic	false

Image source: http://www.lalit.org/lab/javascript-css-font-detect/

Canvas: *Providing multiple ways to fingerprint since HTML5*

<u>Detection Methodology:</u>

1. Canvas created and text written
2. >= 50 distinct, valid fonts set
3. >= 50 calls to `measureText()`

Image source: http://www.lalit.org/lab/javascript-css-font-detect/

Canvas: *Providing multiple ways to fingerprint since HTML5*

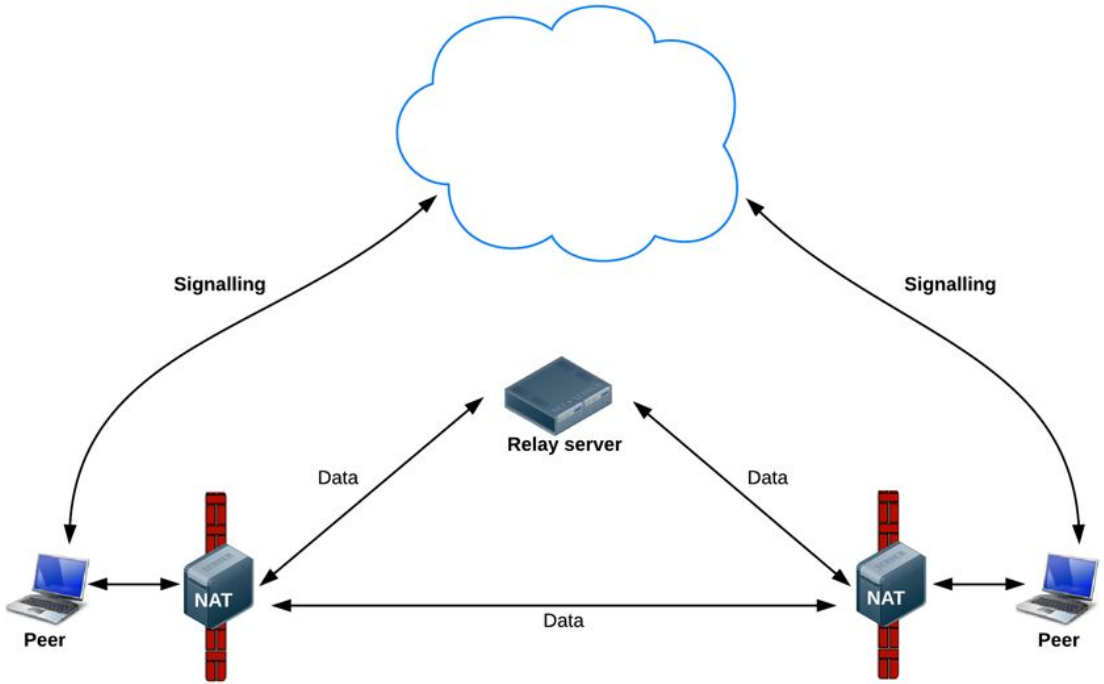
- 3,250 of the top 1 million sites
- Almost all Media Math (90%)
- Skew towards top sites (2.5% of top 1k)

Image source: http://www.lalit.org/lab/javascript-css-font-detect/

The Diversity of Fingerprinting

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

Abusing WebRTC candidate generation for tracking



Source: http://www.html5rocks.com/en/tutorials/webrtc/basics/

Abusing WebRTC candidate generation for tracking

Detection Methodology:

1. Select all scripts calling `createDataChannel` and `createOffer`, which also access the `onicecandidate` event handler

2. Manually examine the script to determine if it's a tracker

Abusing WebRTC candidate generation for tracking

Detection Methodology:

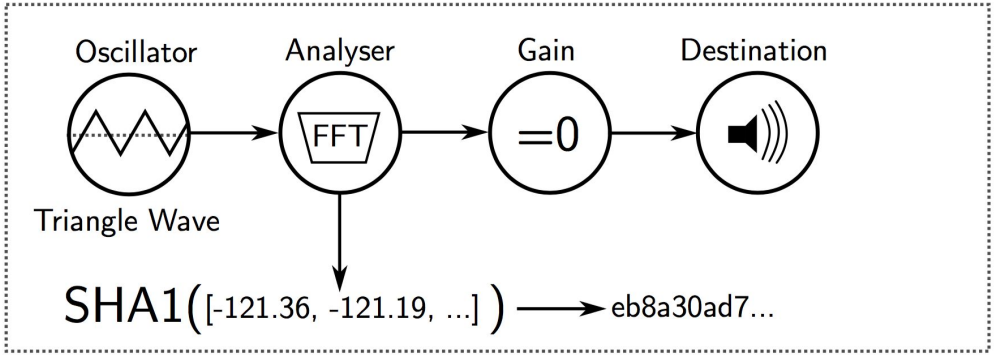
1. Select all scripts calling `createDataChannel` and `createOffer`, which also access the `onicecandidate` event handler

2. Manually examine the script to determine if it's a tracker

~90% of uses were tracking. 57 scripts on 625 sites.

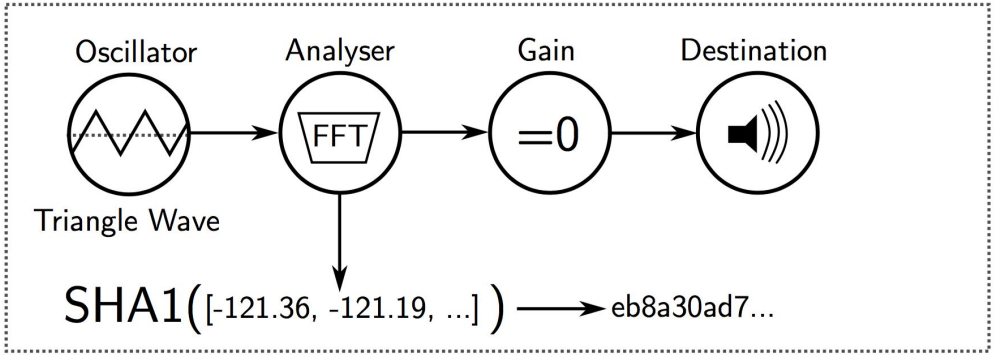
Using AudioContext for fingerprinting

Used by:
`cdn-net.com` script

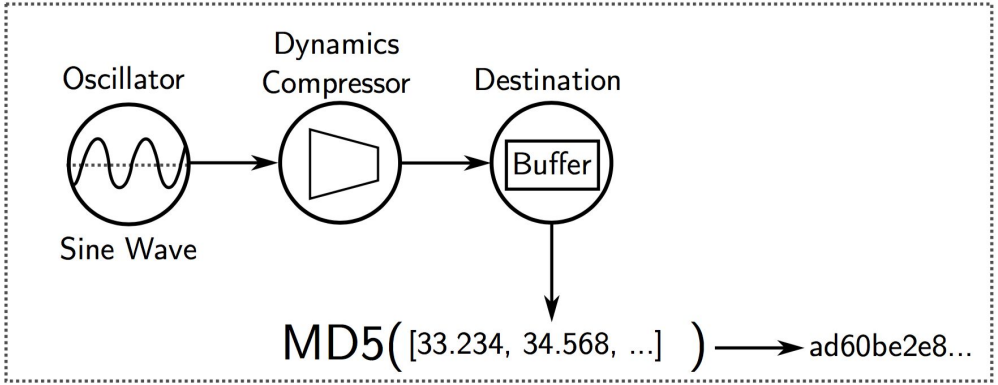


Using AudioContext for fingerprinting

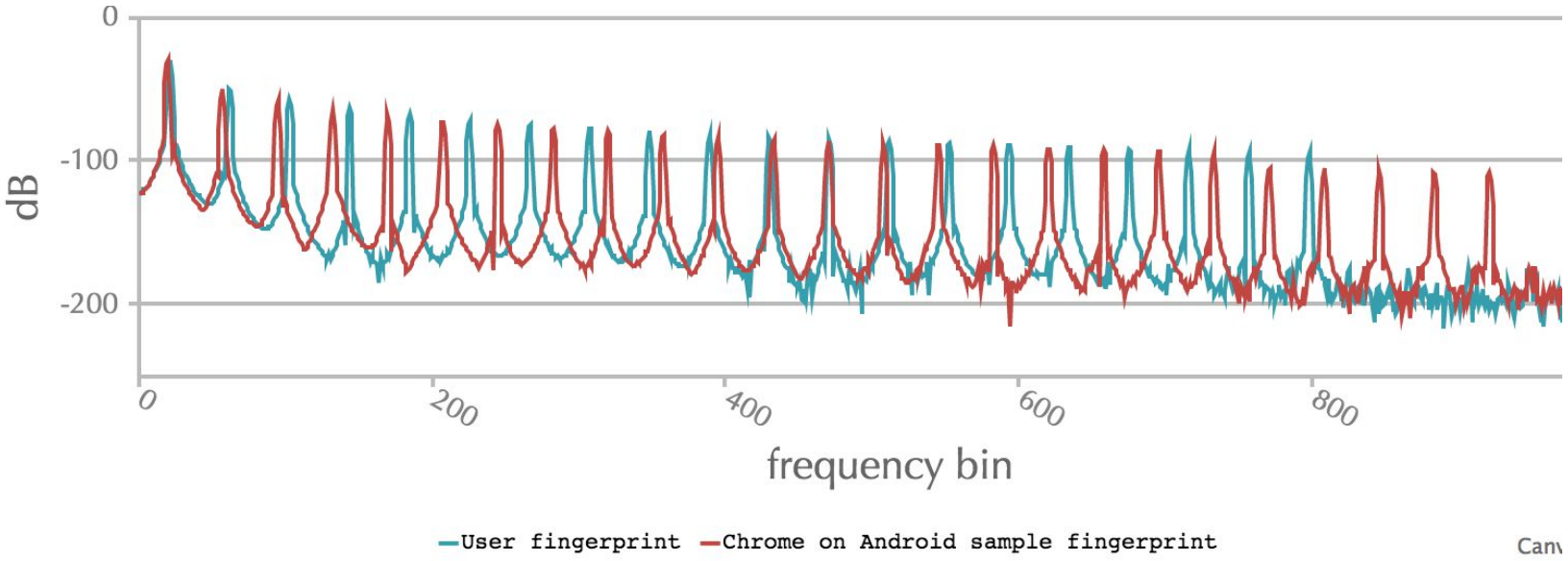
Used by:
`cdn-net.com` script



Used by:
`pxi.pub` and
`ad-score.com` scripts



Using AudioContext for fingerprinting



Live test page: https://audiofingerprint.openwpm.com/

Third parties (and trackers) may impede HTTPS adoption

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

Sites may avoid adopting HTTPS if they include HTTP 3rd parties

Half of all third parties are HTTP only

5%

~~Half~~ of all third parties are HTTP only

...when weighted by popularity

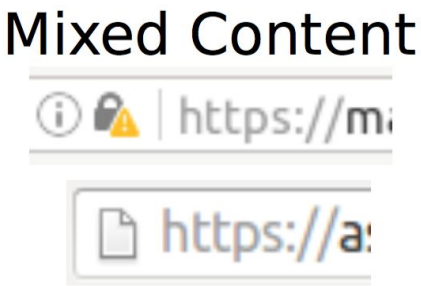
5%

~~Half~~ of all third parties are HTTP only

...when weighted by popularity

~25% of HTTP sites contain at least one HTTP-only resource

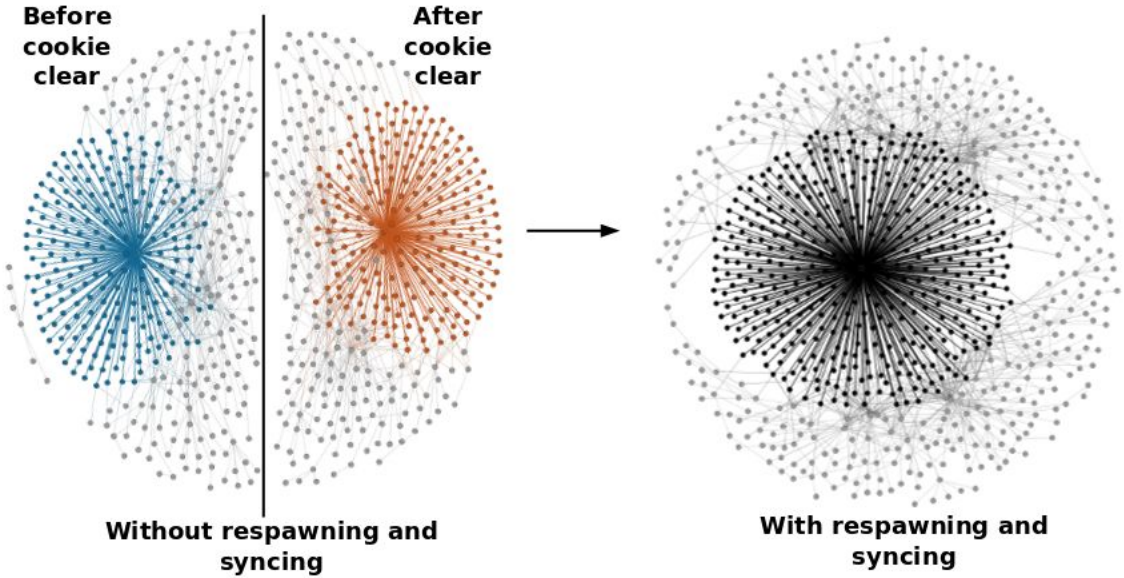
HTTP-Only third parties Impede HTTPS Adoption



~55% of mixed content warnings caused only by third parties

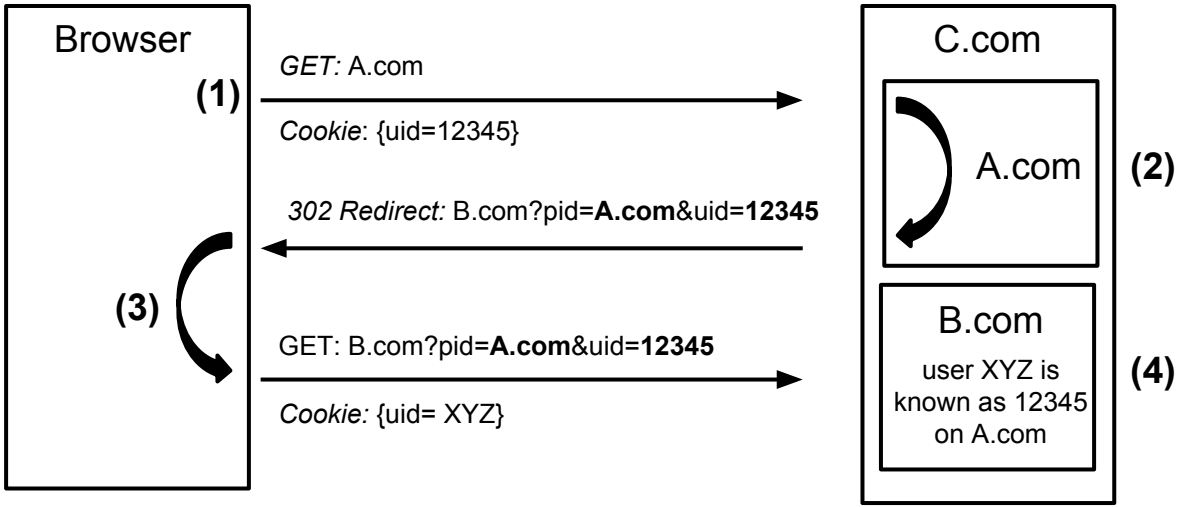
~10% caused only by trackers

What does it take to start fresh on the web?

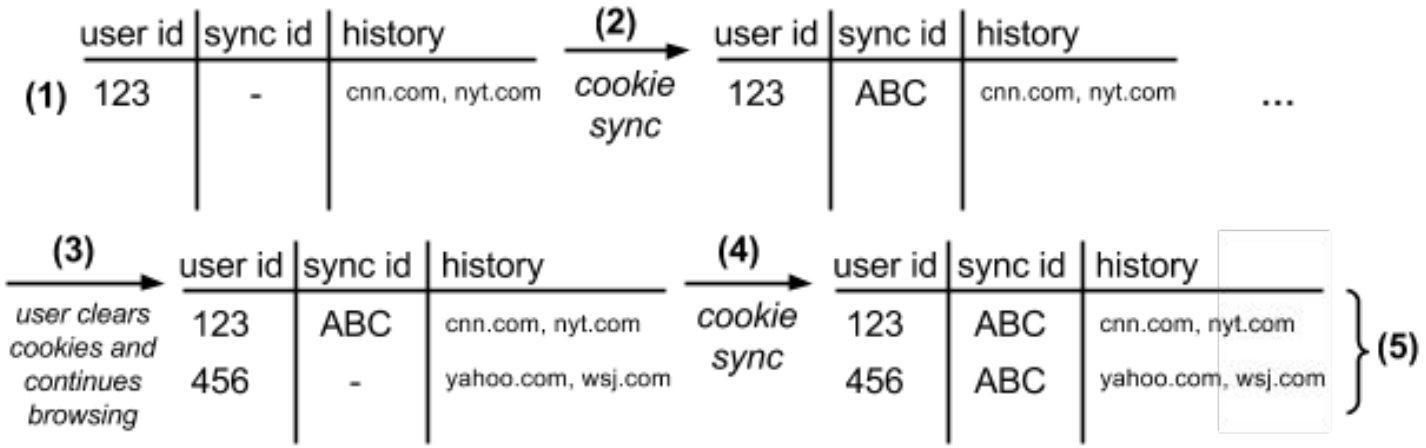


The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (CCS 2014)

Cookie Syncing



Network effects amplify bad actors



Network effects amplify bad actors

→ Only need 1 party to respawn cookies or fingerprint

→ If ID synced with large exchange, identity reintroduced

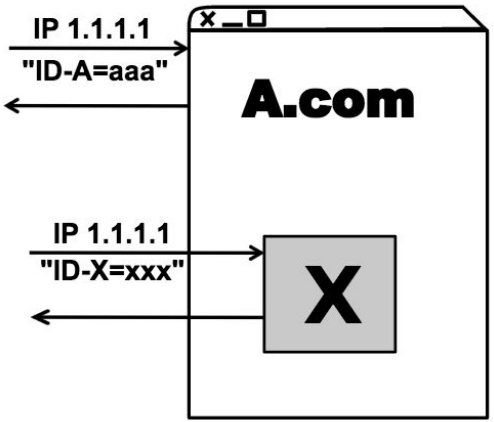
Real example:

→ Respawning by third-party found on 1 site
→ Sync with ad exchange found on 11% of sites

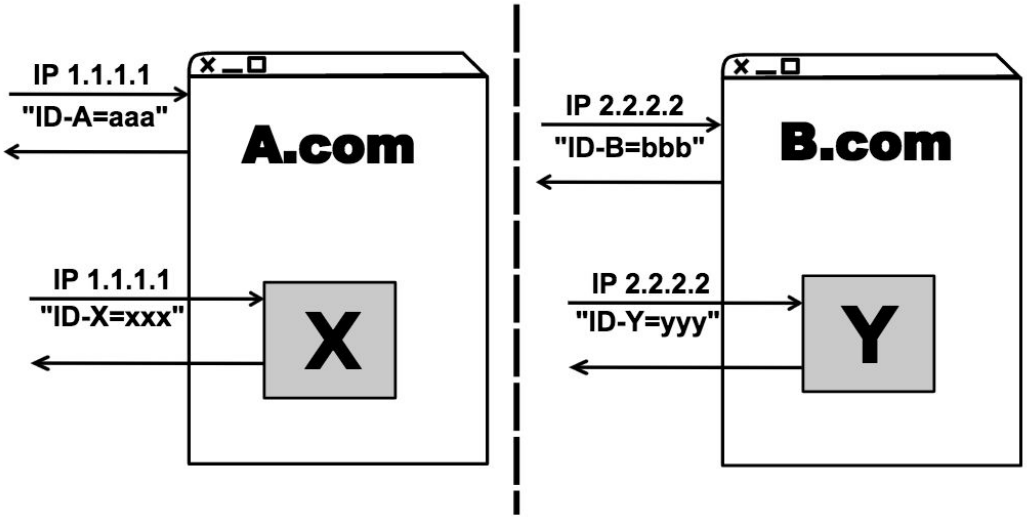
How well does tracking help network adversaries?

Cookies That Give You Away: The Surveillance Implications of Web Tracking (WWW 2015)

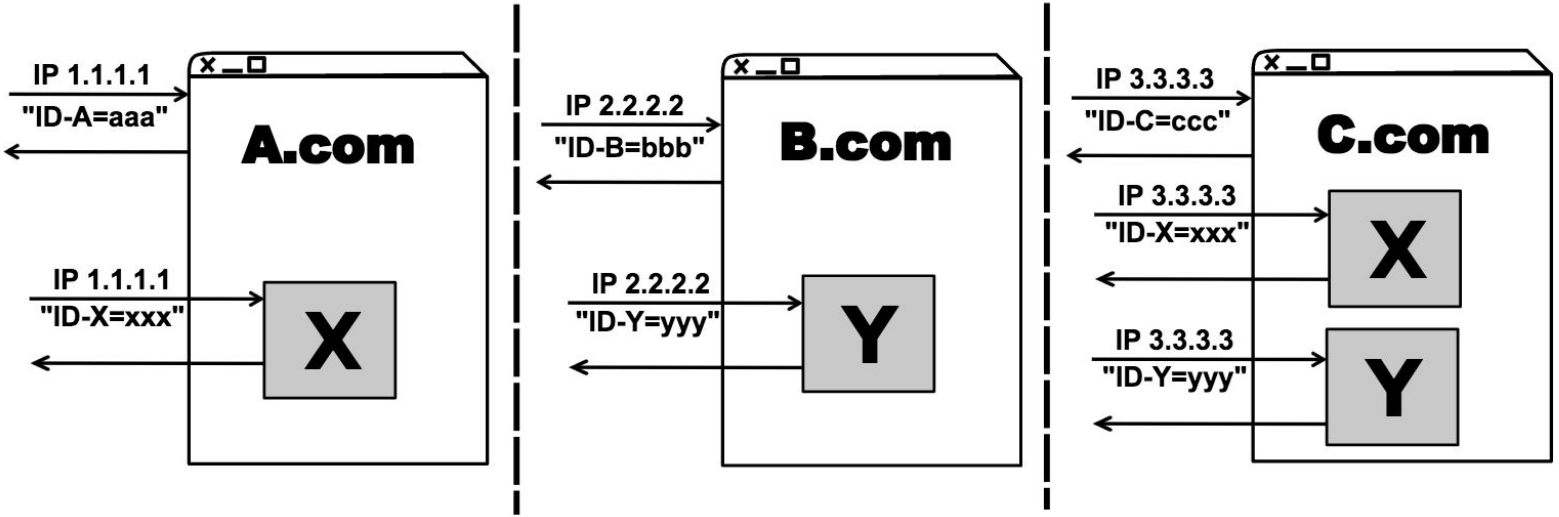
Transitive linking of cookies

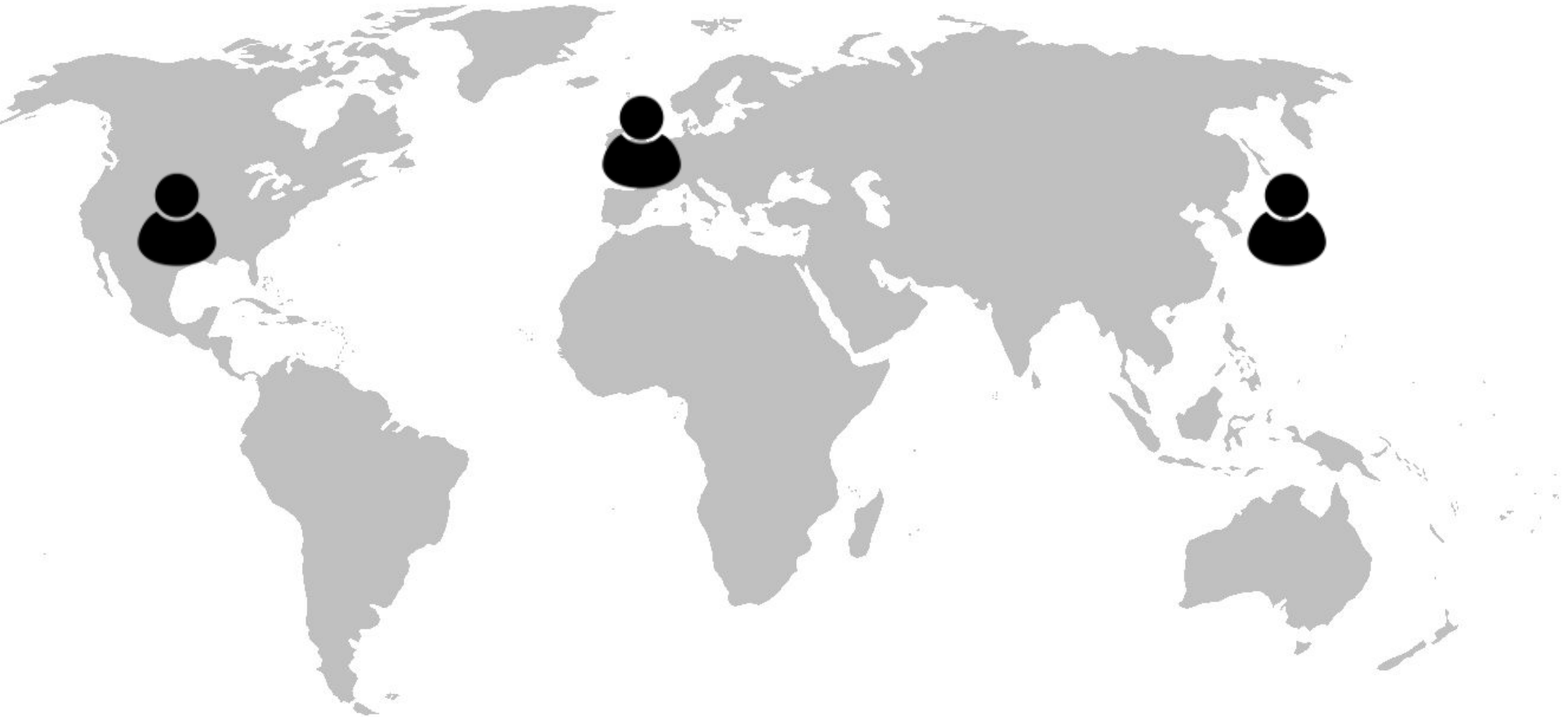


Transitive linking of cookies



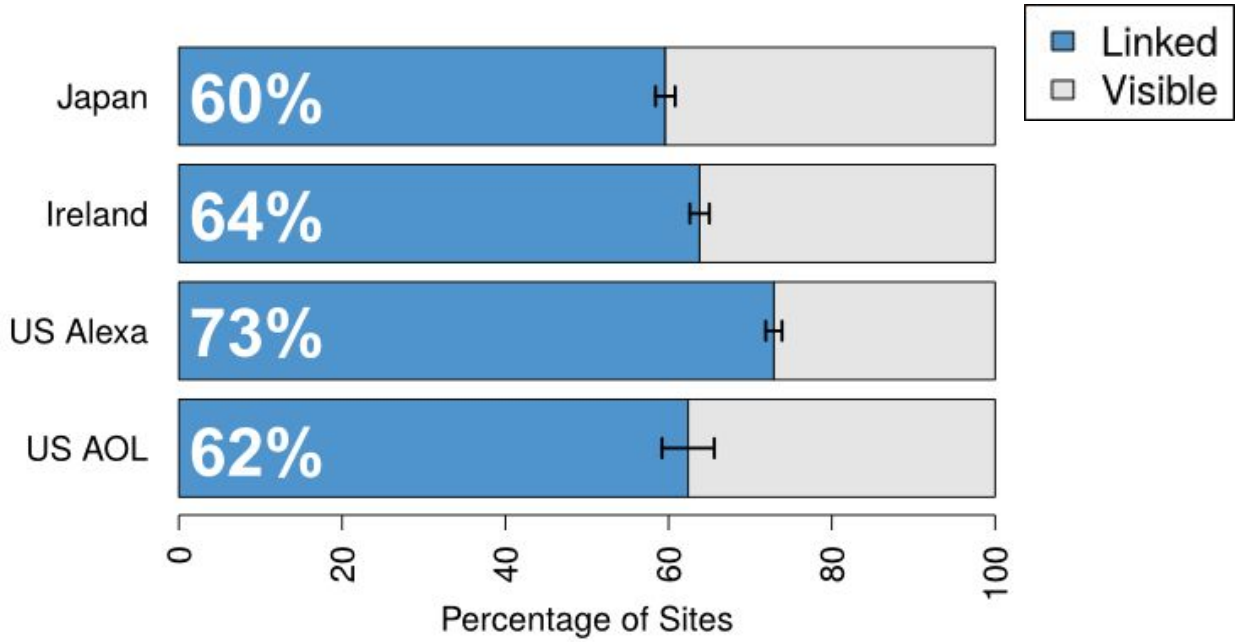
Transitive linking of cookies





Measurement under different legal models

Average percentage of first-party sites linked



Average number of identity leakers

4

6

10

Do Privacy Tools Help?

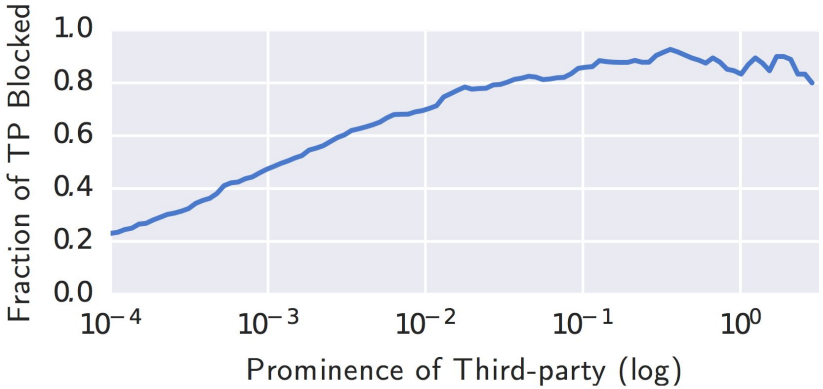
Online Tracking: A 1-million-site Measurement and Analysis (CCS 2016)

Blocking stateful tracking

- Third-party cookie blocking
  - Only a handful of sites work around this by redirecting the top-level domain
  - Average number of third-parties per site reduced from ~18 to ~13
- Ghostery
  - Average number of third-parties per site reduced from ~18 to ~3
  - Very few third-party cookies are set

Blocking stateful tracking

- Third-party cookie blocking
  - Only a handful of sites work around this by redirecting the top-level domain
  - Average number of third-parties per site reduced from ~18 to ~13
- Ghostery
  - Average number of third-parties per site reduced from ~18 to ~3
  - Very few third-party cookies are set



Blocking Fingerprinting

	EasyList + EasyPrivacy	
Technique	Percentage of Scripts	Percentage of Sites

Blocking Fingerprinting

	EasyList + EasyPrivacy	
Technique	Percentage of Scripts	Percentage of Sites
Canvas	25%	88%

Blocking Fingerprinting

	EasyList + EasyPrivacy	
Technique	Percentage of Scripts	Percentage of Sites
Canvas	25%	88%
Canvas Font	10%	91%

Blocking Fingerprinting

	EasyList + EasyPrivacy	
Technique	Percentage of Scripts	Percentage of Sites
Canvas	25%	88%
Canvas Font	10%	91%
WebRTC	5%	6%

Blocking Fingerprinting

	EasyList + EasyPrivacy	
Technique	Percentage of Scripts	Percentage of Sites
Canvas	25%	88%
Canvas Font	10%	91%
WebRTC	5%	6%
AudioContext	6%	2%

Crowdsourced lists are insufficient

→ Lists miss less popular trackers

→ Lists fail to block new techniques

→ Relatively high false positive (anecdotal breakage)

Next Steps

The Princeton Web Census

Monthly
1 Million Site Crawl

The Princeton Web Census

Monthly
1 Million Site Crawl

Collecting:

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

Detection Heuristics as Ground truth

Detection Methodology:

1. Canvas created and text written
2. >= 50 distinct, valid fonts set
3. >= 50 calls to `measureText()`

Detection Methodology:

1. Select all scripts calling `createDataChannel` and `createOffer`, which also access the `onicecandidate` event handler
2. Manually examine the script to determine if it's a tracker

Detection Methodology:

1. Canvas height and width >= 16px
2. Text >= 2 colors OR >= 10 characters
3. Should not call `save`, `restore`, or `addEventListener`. (Used with interactive or animated content)
4. Calls `toDataURL` or `getImageData`.

- Largely a manual effort
- Benefit from overall low API usage
+ Fingerprinting techniques clustered
+ Fingerprint scripts tend to be standalone

Machine Learning for Tracker Detection

Category	Description	Number of features
URL String	Keywords like ‘ad’, ‘popup’, ‘banner’, are query parameters valid, number of commas, etc.	16
Third Party Statistical	How many different first parties a third party domain exists on and similar	7
Http-Cookies	Number of cookies set, if session or secure cookies are set, entropy in cookie values, etc.	9
URL Content	If url is an image or a script	3
Javascript Content	Tf-idf based various function calls in the javascript code as features	451

Master’s Thesis: *Using Machine Learning for Online Tracking Protection and Ad Blocking* by Shivam Agarwal

Data Access

DATA
TRANSPARENCY
LAB

Contribute:

github.com/citp/OpenWPM

Collaborate:

webtap.princeton.edu

Email: ste@cs.princeton.edu **Twitter:** @s_englehardt

Image Assets from the Noun Project:
Database by Creative Stall; programmer by Hadi Davodpour