No boundaries: Data exfiltration by third-party tracking scripts

Steven Englehardt

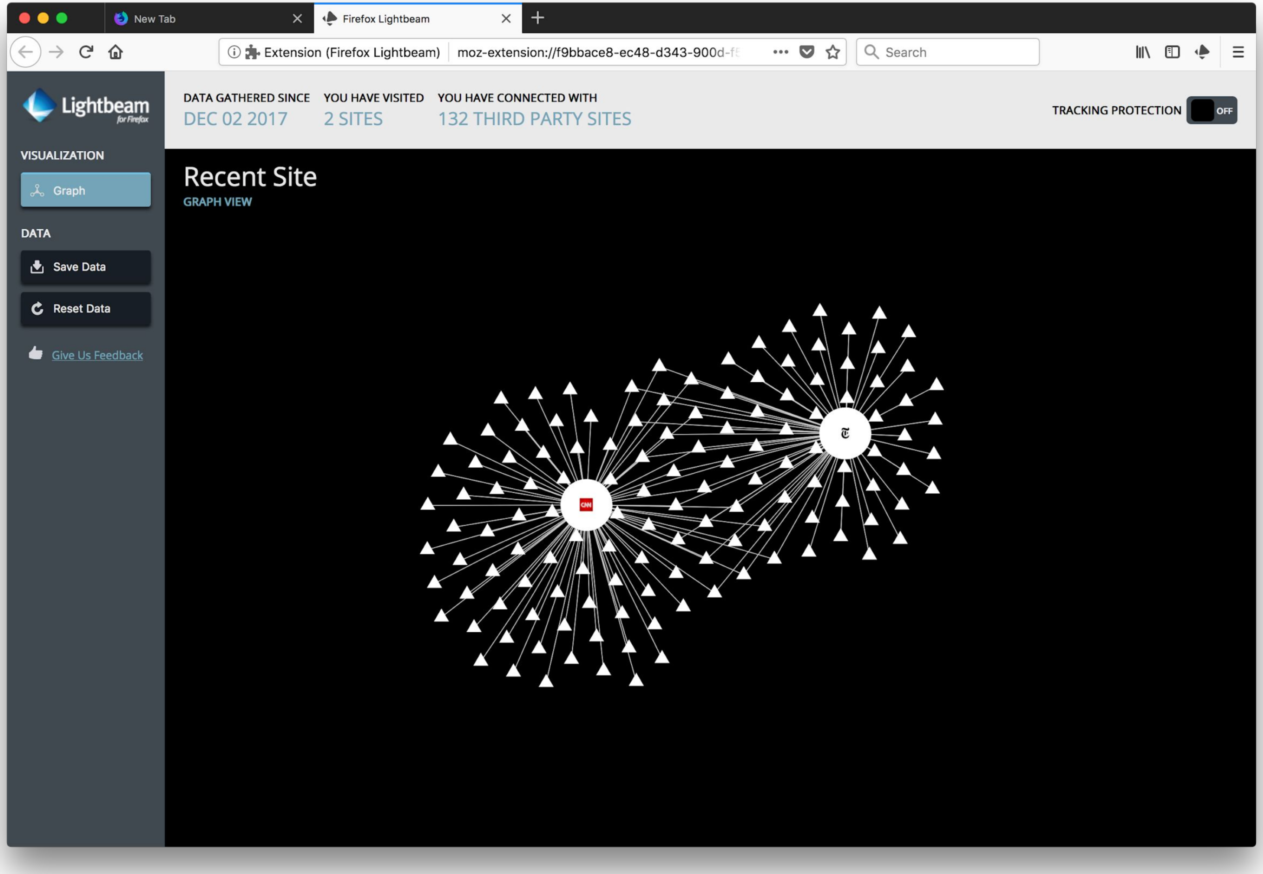
Joint work with:
Güneş Acar, Jeffrey Han, and
Arvind Narayanan

I’m now at...

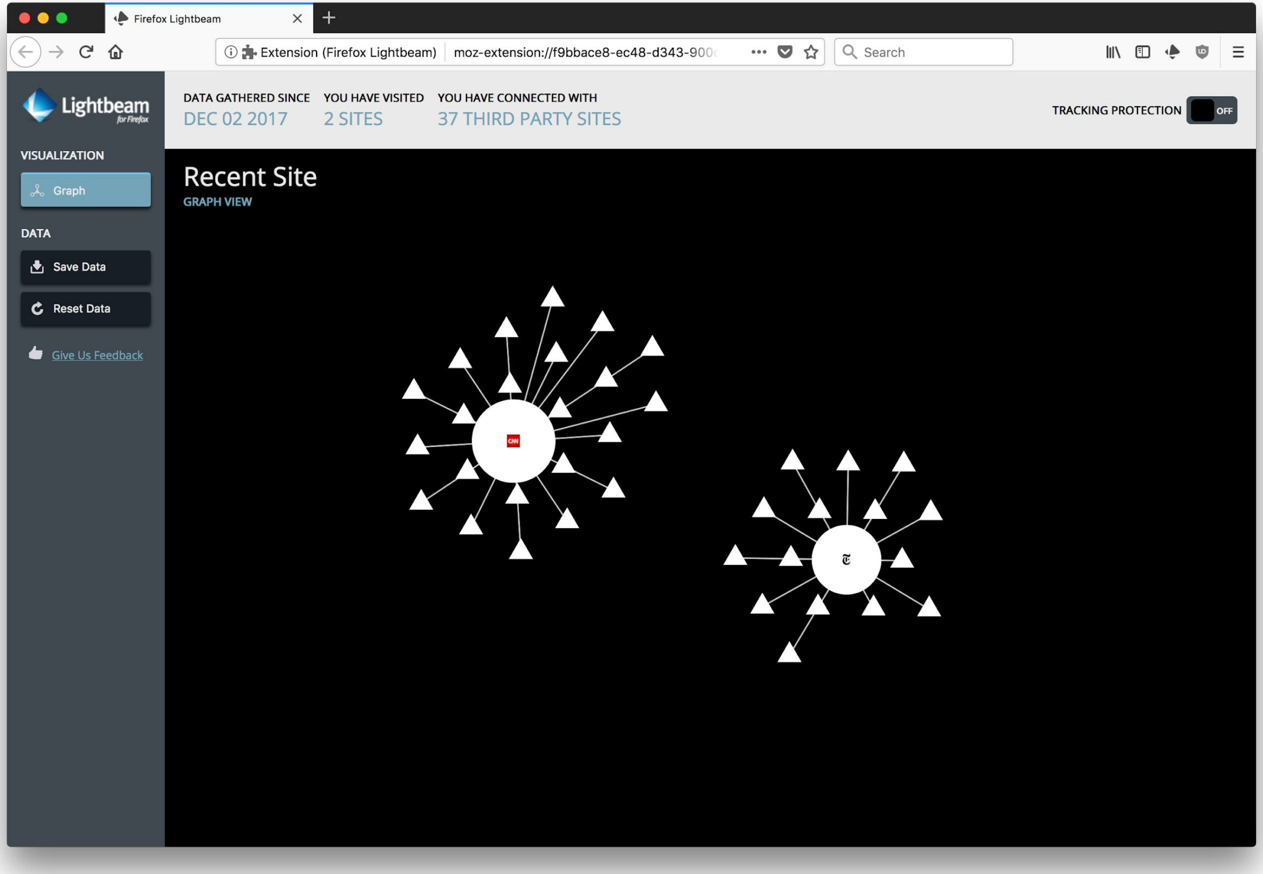
moz://a

PRINCETON
UNIVERSITY

CENTER FOR
INFORMATION
TECHNOLOGY
POLICY
PRINCETON UNIVERSITY



Just **two page visits** cause requests to **132 distinct hostnames**.



DATA GATHERED SINCE DEC 02 2017

YOU HAVE VISITED 2 SITES

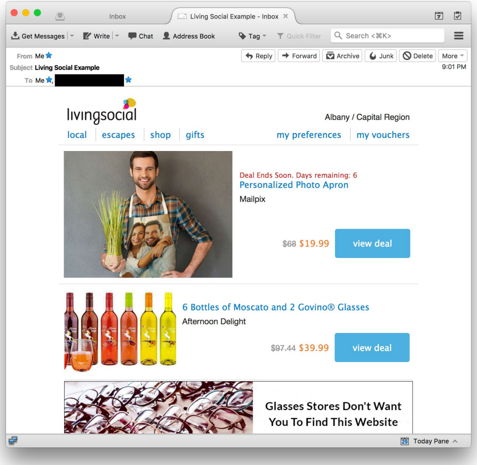
YOU HAVE CONNECTED WITH 37 THIRD PARTY SITES

TRACKING PROTECTION OFF

Recent Site

GRAPH VIEW

With uBlock Origin enabled, the number of hostnames requested is down to 37. **Nearly 100 of the hosts loaded were ads, trackers, and analytics.**



From Me

Subject Living Social Example

To Me, [redacted]

9:01 PM

livingsocial

Albany / Capital Region

local | escapes | shop | gifts | my preferences | my vouchers

Deal Ends Soon. Days remaining: 6

Personalized Photo Apron

Mailpix

~~$68~~ $19.99 view deal

6 Bottles of Moscato and 2 Govino® Glasses

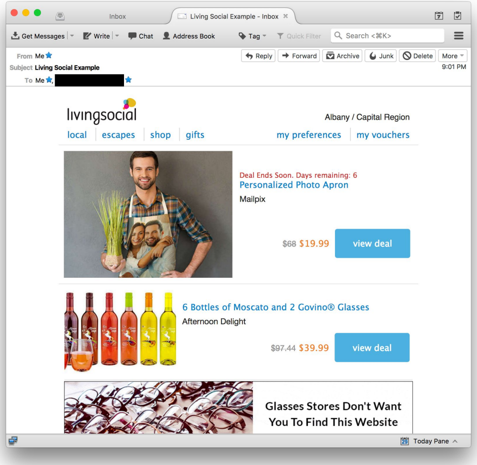
Afternoon Delight

~~$97.44~~ $39.99 view deal

Glasses Stores Don't Want You To Find This Website

What happens when you load remote content in an email?

Englehardt, Han, and Narayanan, "I never signed up for this! Privacy implications of email tracking" (PETS 2018)



From Me
Subject Living Social Example
To Me, [redacted]
9:01 PM

livingsocial

Albany / Capital Region

Deal Ends Soon. Days remaining: 6
Personalized Photo Apron
Mailpix
~~$68~~ $19.99

6 Bottles of Moscato and 2 Govino® Glasses
Afternoon Delight
~~$97.44~~ $39.99

Glasses Stores Don't Want You To Find This Website

Your device contacts 24 companies
→ 20 can track you (if supported)
→ 10 receive an MD5 hash of your email address

Receives MD5(email address) & Sets a Cookie

American List Counsel (alcmpn.com)
LiveIntent (liadm.com)
Oracle (nexac.com)
Acxiom (rlcdn.com, pippio.com, acxiom-online.com)
Criteo (criteo.com)
Conversant Media (dotomi.com)
V12 Data (v12group.com)
VideoAmp (videoamp.com)
<Unknown> (alocdn.com)

Sets a Cookie

OpenX (openx.net)
comScore (scorecardresearch.com, voicefive.com)
Oracle (bluekai.com)
Google (doubleclick.net)
Realtime Targeting Aps (mojn.com)
MediaMath (mathtag.com)
TapAd (tapad.com)
IPONWEB (bidswitch.net)
AOL (advertising.com)
Centro (sitescout.com)
The Trade Desk (adsrvr.org)
Adobe (demdex.net)

Receives MD5(email addr.)

Criteo (emailretargeting.com)
Neustar (agkn.com)

Receives Bare Request

LiveIntent (licasd.com)
Google (2mdn.net)
Akamai (akamai.net)

Englehardt, Han, and Narayanan, “I never signed up for this! Privacy implications of email tracking” (PETS 2018)

A user’s email address is the perfect identifier!

- It’s unique
- It rarely changes
- It’s the same across devices
- Consumers freely provide it to stores
- There’s a lot of associated data

PII-based tracking

```
UUID = {
  MD5(bob@example.com),
  SHA1(bob@example.com),
  SHA256(bob@example.com)
}
```

Are trackers also collecting PII on the web?

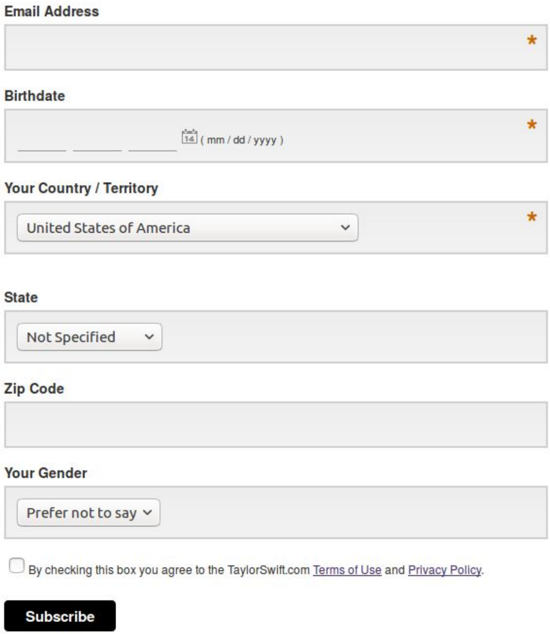
We can use web crawls to detect PII collection:

- **Crawl 50K sites with OpenWPM**
  - main page and 5 inner pages
- **Monitor access to PII sources**
  - Autofilled credentials
  - Mutation events to monitor form insertion
  - HTMLInputElement instrumentation to intercept access to form input fields
- **Search for PII in network traffic**
  - Request and response headers
  - POST payloads

OpenWPM: https://github.com/citp/OpenWPM

Challenge: Measurements require the automated submission of PII to sites

Mailing list sign-ups



Email Address *

Birthdate *

(mm / dd / yyyy)

Your Country / Territory *

United States of America

State

Not Specified

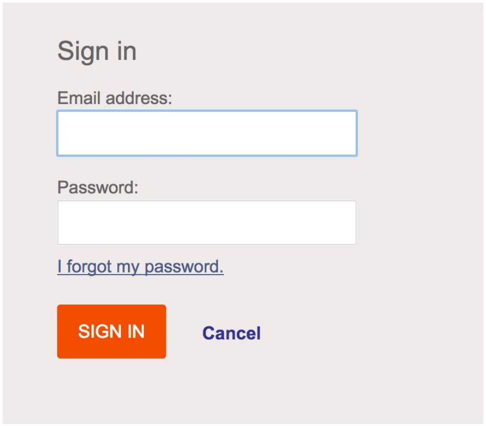
Zip Code

Your Gender

Prefer not to say

□ By checking this box you agree to the TaylorSwift.com Terms of Use and Privacy Policy.

Login Forms



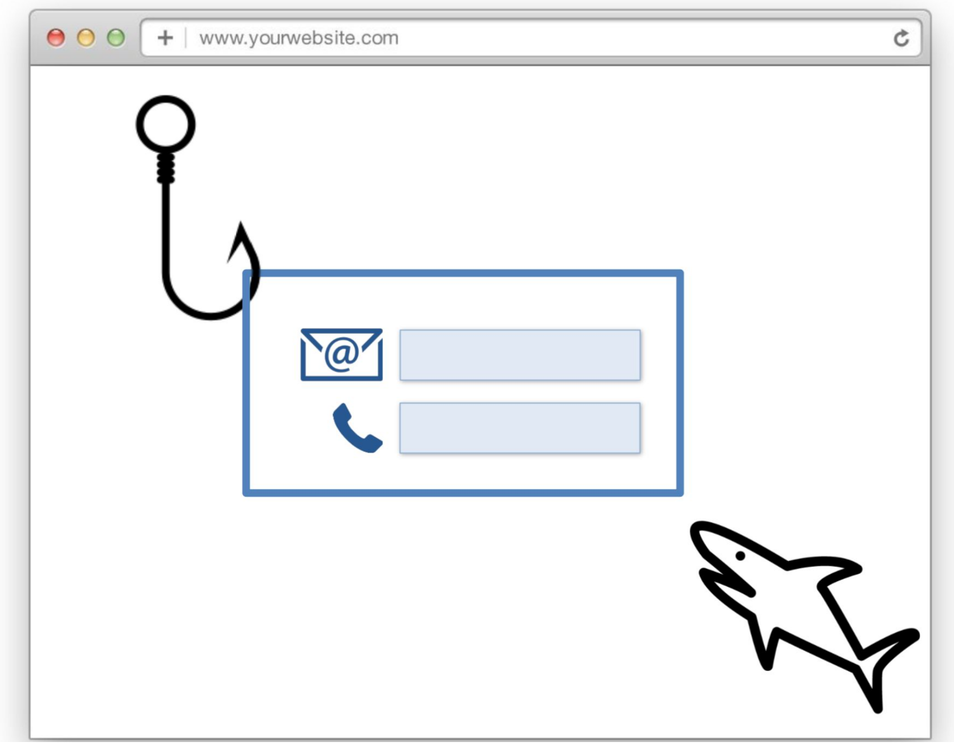
Sign in

Email address:

Password:

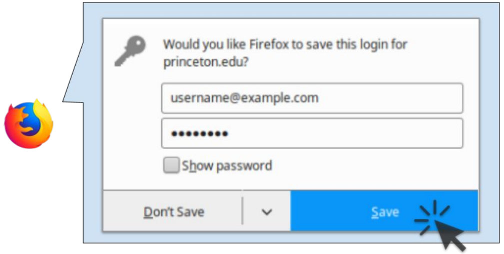
I forgot my password.

Injecting PII into the web: bait technique



Third parties collect PII for tracking

Autofill abuse



Would you like Firefox to save this login for princeton.edu?

username@example.com

•••••••••

□ Show password

Don't Save | Save

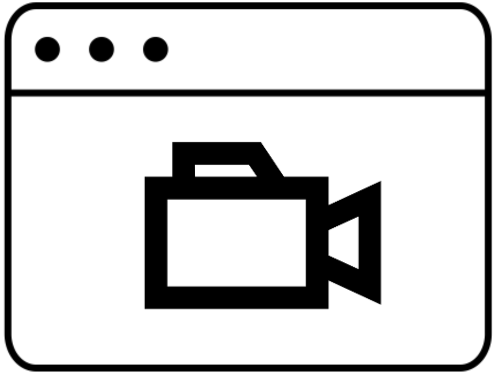
Social Login

Login with Facebook

Firstparty.com will receive: your public profile and email address

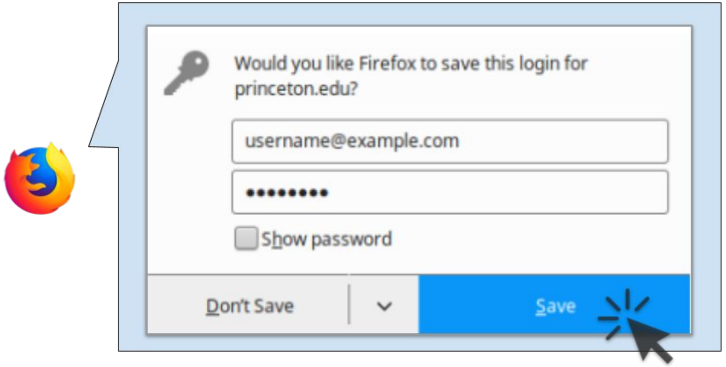
Continue

Session Recording



"No boundaries: Exfiltration of personal data by session-replay scripts" (freedom-to-tinker.com)
"No boundaries for user identities: Web trackers exploit browser login managers" (freedom-to-tinker.com)
"No boundaries for Facebook data: third-party trackers abuse Facebook Login" (freedom-to-tinker.com)

Login manager abuse for web tracking



Would you like Firefox to save this login for princeton.edu?

username@example.com

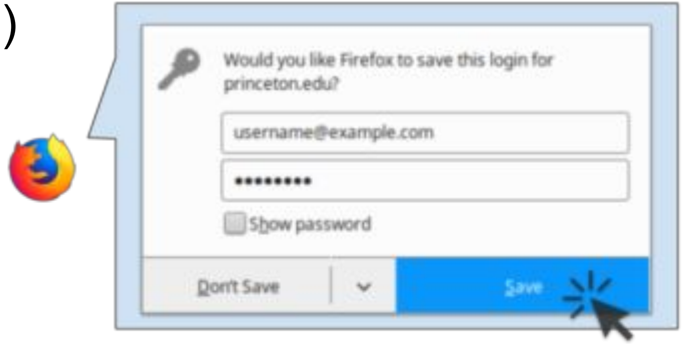
•••••••••

□ Show password

Acar, Englehardt and Narayanan, "No boundaries for user identities: Web trackers exploit browser login managers" (Freedom to Tinker)

Built-in login managers

- Remembers username & passwords (opt-in)
- Autofills login forms
- Different than CC and address autofill



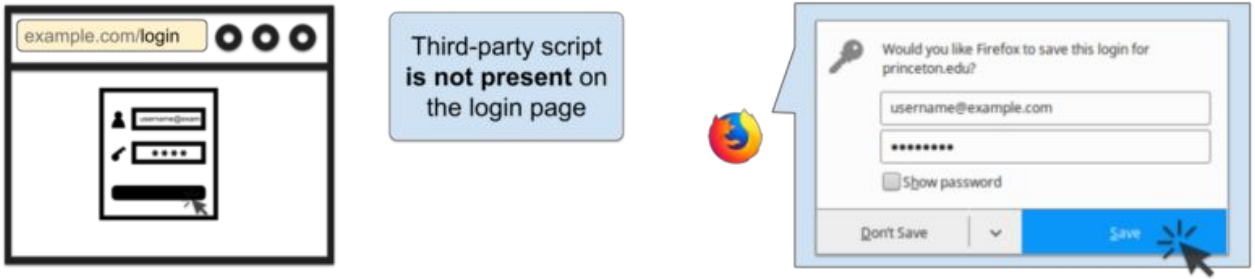
Would you like Firefox to save this login for princeton.edu?

username@example.com

•••••••••

Acar, Englehardt and Narayanan, "No boundaries for user identities: Web trackers exploit browser login managers" (Freedom to Tinker)

User submits a login or registration form, clicks "Save" to store the credentials.



Third-party script **is not present** on the login page

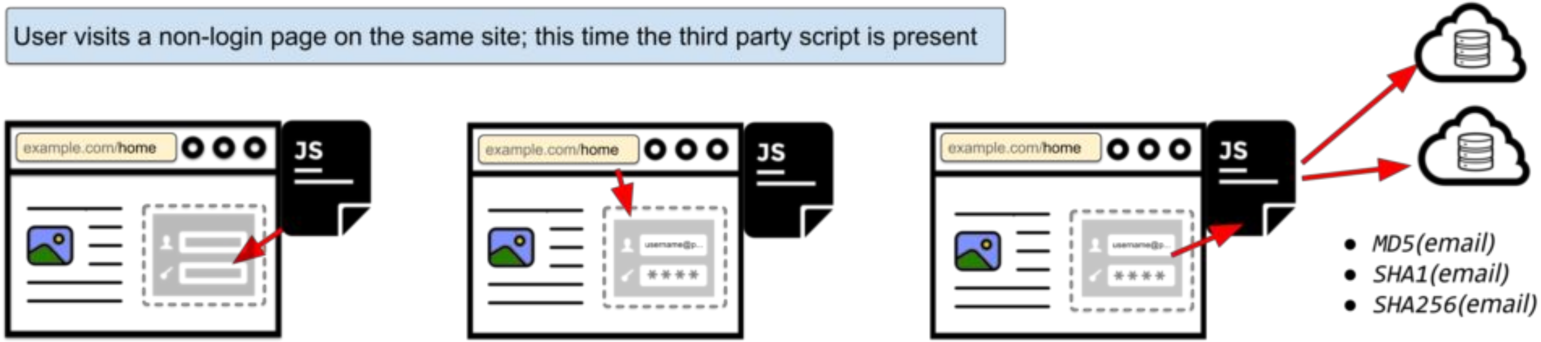
Would you like Firefox to save this login for princeton.edu?

username@example.com

•••••••••

□ Show password

User visits a non-login page on the same site; this time the third party script is present



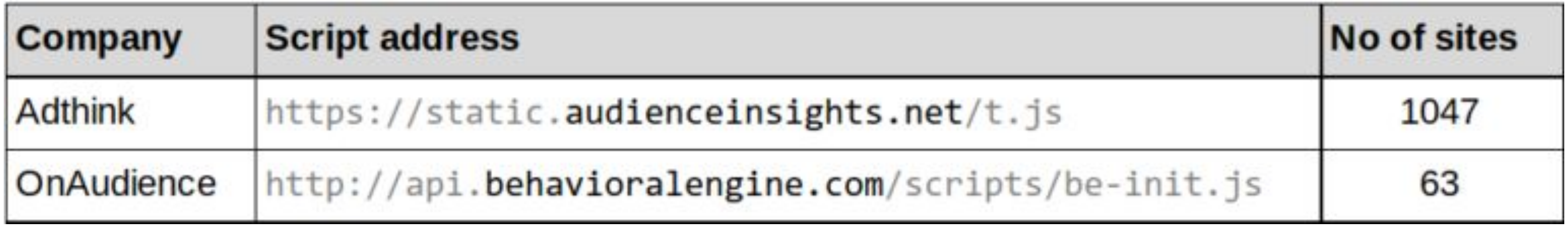
- *MD5(email)*
- *SHA1(email)*
- *SHA256(email)*

1. Third-party script injects an invisible login form

2. Login manager fills in user's email and password

3. The script reads the email address from the form and sends it hashes to third-party servers

Findings



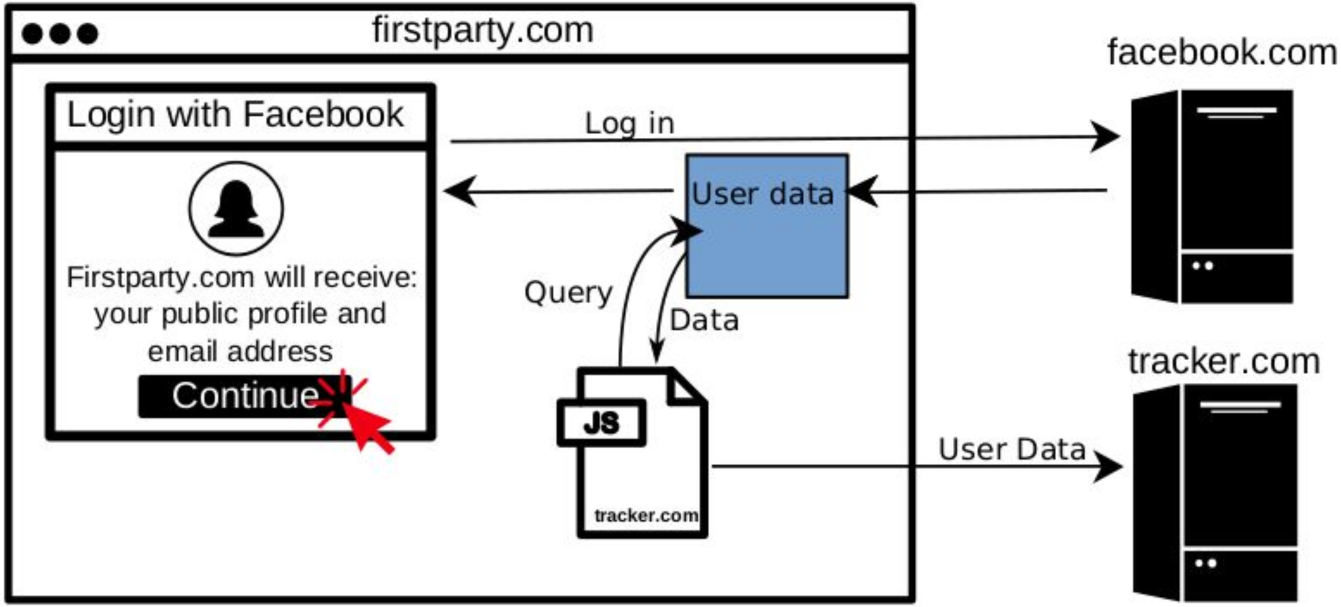
Company	Script address	No of sites
Adthink	https://static.audienceinsights.net/t.js	1047
OnAudience	http://api.behavioralengine.com/scripts/be-init.js	63

Social Login abuse for web tracking

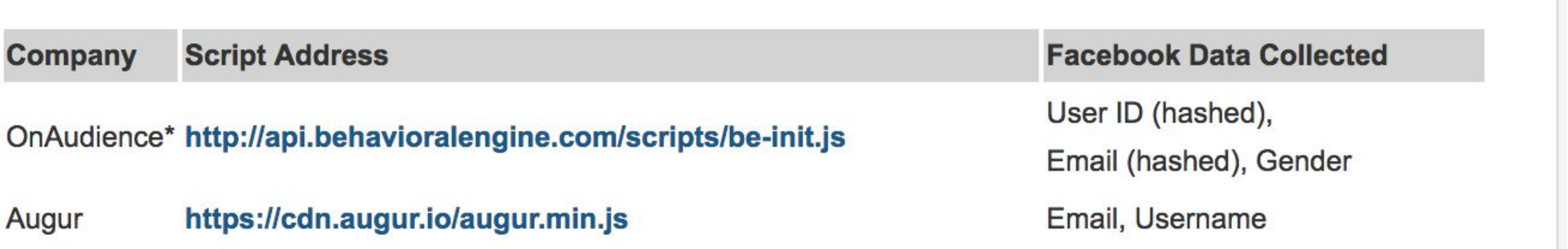
Login with Facebook

Firstparty.com will receive: your public profile and email address

Englehardt, Acar, and Narayanan, “No boundaries for Facebook data: third-party trackers abuse Facebook Login” (Freedom to Tinker)


firstparty.com
Login with Facebook
Firstparty.com will receive: your public profile and email address
Continue
Log in
User data
Query
Data
JS
tracker.com
User Data
facebook.com
tracker.com

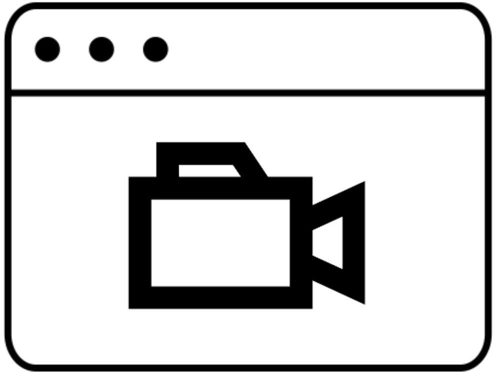
Findings



Company	Script Address	Facebook Data Collected
OnAudience*	http://api.behavioralengine.com/scripts/be-init.js	User ID (hashed), Email (hashed), Gender
Augur	https://cdn.augur.io/augur.min.js	Email, Username

...as well as several others grabbing user ID

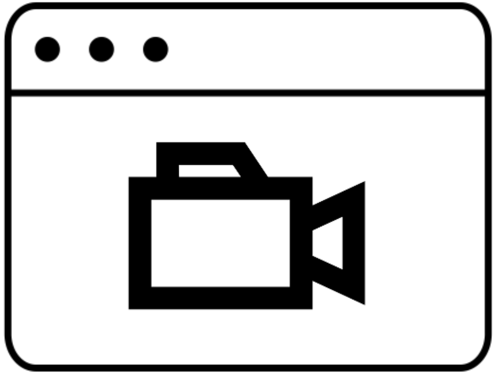
Session recording scripts scoop up sensitive information



“No boundaries: Exfiltration of personal data by session-replay scripts” (freedom-to-tinker.com)

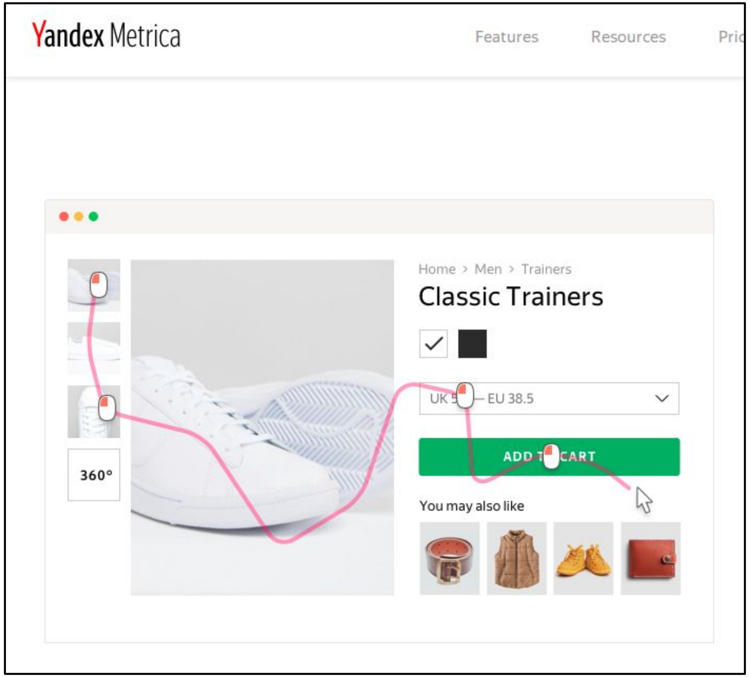
What are session recording scripts?

- Session recording scripts create a “video” of all of a user’s actions on a site.
  - Key presses
  - Mouse clicks, mouse movements
  - Scrolling behavior...
- Publishers can later review the videos.



Englehardt, Acar, and Narayanan, “No boundaries: Exfiltration of personal data by session-replay scripts” (Freedom to Tinker)

Why use session recording scripts?



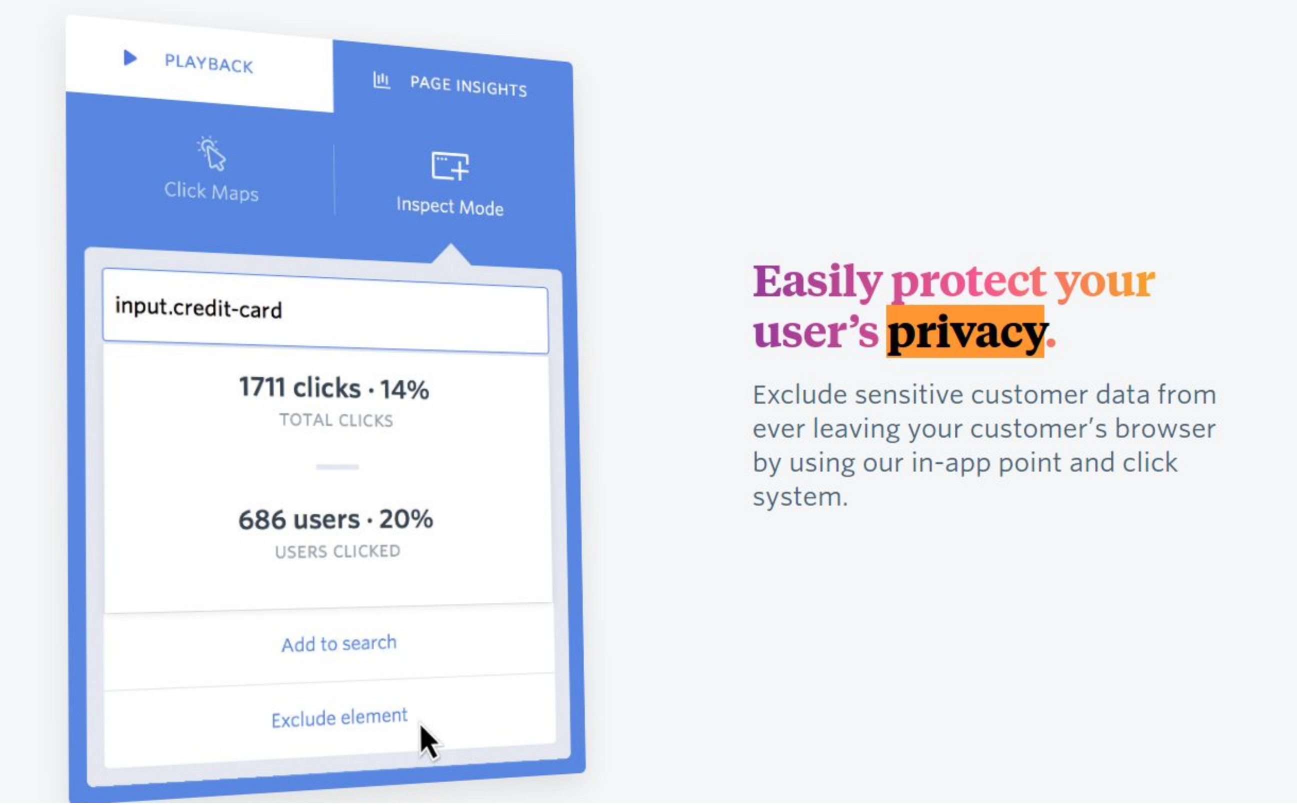
Answer questions like:

- Who are my most valuable customers?
- Who added items to the cart but didn’t convert?
- Where do users leave the onboarding flow?
- Where are users frustrated?

The problem: recordings require **a ton** of data

- Full page source and text
- Mouse movements & clicks
- Keypresses

Companies support redaction



input.credit-card

1711 clicks · 14%
TOTAL CLICKS

686 users · 20%
USERS CLICKED

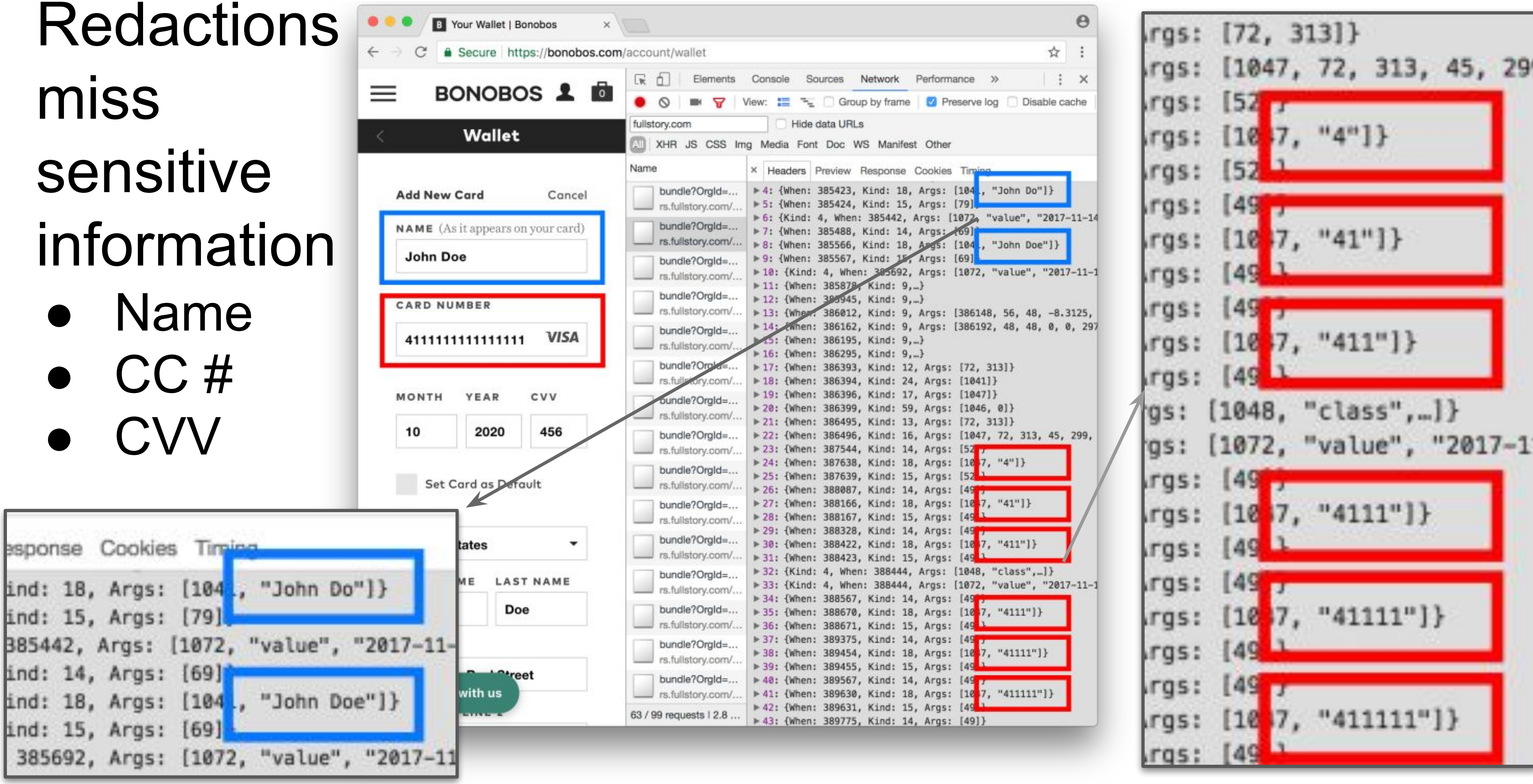
Easily protect your user's privacy.

Exclude sensitive customer data from ever leaving your customer's browser by using our in-app point and click system.

How can things go wrong?

Redactions miss sensitive information

- Name
- CC #
- CVV



Your Wallet | Bonobos
Secure https://bonobos.com/account/wallet
BONOBOS
Wallet
Elements Console Sources Network Performance »
View: □ Group by frame ☑ Preserve log □ Disable cache
fullstory.com □ Hide data URLs
All XHR JS CSS Img Media Font Doc WS Manifest Other
Name
bundle?OrgId=... rs.fullstory.com/...
63 / 99 requests | 2.8 ...
Headers Preview Response Cookies Timing

Add New Card Cancel

NAME (As it appears on your card)
John Doe

CARD NUMBER
4111111111111111 VISA

MONTH 10
YEAR 2020
CVV 456

□ Set Card as Default

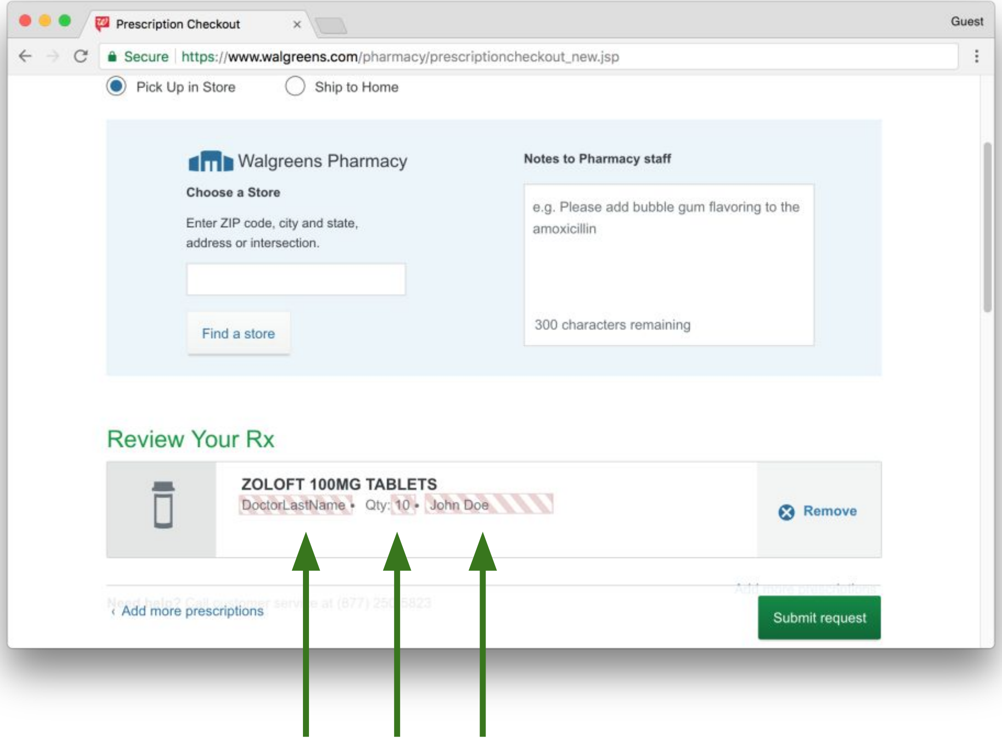
tates

LAST NAME
Doe

with us

```
▶4: {When: 385423, Kind: 18, Args: [1041, "John Do"]}
▶5: {When: 385424, Kind: 15, Args: [79]}
▶6: {Kind: 4, When: 385442, Args: [1072, "value", "2017-11-14
▶7: {When: 385488, Kind: 14, Args: [69]}
▶8: {When: 385566, Kind: 18, Args: [1041, "John Doe"]}
▶9: {When: 385567, Kind: 15, Args: [69]}
▶10: {Kind: 4, When: 385692, Args: [1072, "value", "2017-11-1
▶11: {When: 385878, Kind: 9,…}
▶12: {When: 385945, Kind: 9,…}
▶13: {When: 386012, Kind: 9, Args: [386148, 56, 48, -8.3125,
▶14: {When: 386162, Kind: 9, Args: [386192, 48, 48, 0, 0, 297
▶15: {When: 386195, Kind: 9,…}
▶16: {When: 386295, Kind: 9,…}
▶17: {When: 386393, Kind: 12, Args: [72, 313]}
▶18: {When: 386394, Kind: 24, Args: [1041]}
▶19: {When: 386396, Kind: 17, Args: [1047]}
▶20: {When: 386399, Kind: 59, Args: [1046, 0]}
▶21: {When: 386495, Kind: 13, Args: [72, 313]}
▶22: {When: 386496, Kind: 16, Args: [1047, 72, 313, 45, 299,
▶23: {When: 387544, Kind: 14, Args: [52]}
▶24: {When: 387638, Kind: 18, Args: [1047, "4"]}
▶25: {When: 387639, Kind: 15, Args: [52]}
▶26: {When: 388087, Kind: 14, Args: [49]}
▶27: {When: 388166, Kind: 18, Args: [1047, "41"]}
▶28: {When: 388167, Kind: 15, Args: [49]}
▶29: {When: 388328, Kind: 14, Args: [49]}
▶30: {When: 388422, Kind: 18, Args: [1047, "411"]}
▶31: {When: 388423, Kind: 15, Args: [49]}
▶32: {Kind: 4, When: 388444, Args: [1048, "class",…]}
▶33: {Kind: 4, When: 388444, Args: [1072, "value", "2017-11-1
▶34: {When: 388567, Kind: 14, Args: [49]}
▶35: {When: 388670, Kind: 18, Args: [1047, "4111"]}
▶36: {When: 388671, Kind: 15, Args: [49]}
▶37: {When: 389375, Kind: 14, Args: [49]}
▶38: {When: 389454, Kind: 18, Args: [1047, "41111"]}
▶39: {When: 389455, Kind: 15, Args: [49]}
▶40: {When: 389567, Kind: 14, Args: [49]}
▶41: {When: 389630, Kind: 18, Args: [1047, "411111"]}
▶42: {When: 389631, Kind: 15, Args: [49]}
▶43: {When: 389775, Kind: 14, Args: [49]}
```

Walgreens misses fields during redaction



Prescription Checkout
Guest
Secure
https://www.walgreens.com/pharmacy/prescriptioncheckout_new.jsp

● Pick Up in Store ○ Ship to Home

Walgreens Pharmacy

Choose a Store

Enter ZIP code, city and state, address or intersection.

Notes to Pharmacy staff

e.g. Please add bubble gum flavoring to the amoxicillin

300 characters remaining

Review Your Rx

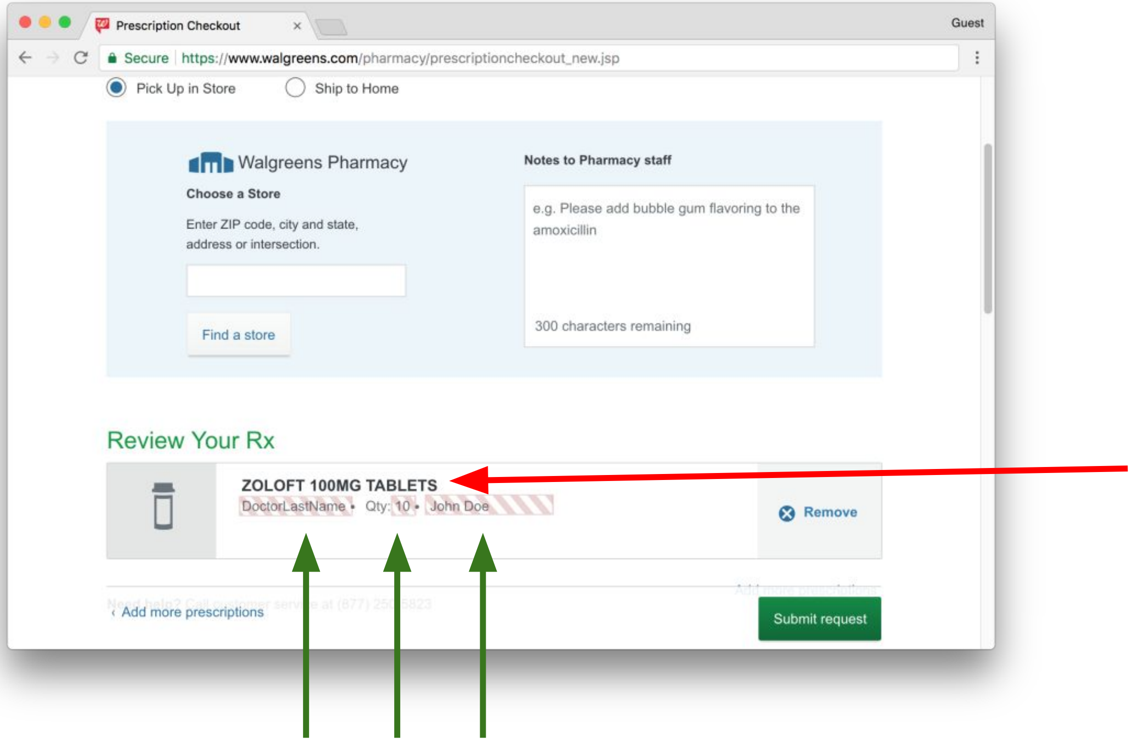
ZOLOFT 100MG TABLETS

DoctorLastName • Qty: 10 • John Doe

Remove
Need help? Call customer service at (877) 250-5823
Add more prescriptions
‹ Add more prescriptions
Submit request

Walgreens makes thorough use of redaction

Walgreens misses fields during redaction



Prescription Checkout

Guest

Secure https://www.walgreens.com/pharmacy/prescriptioncheckout_new.jsp

● Pick Up in Store ○ Ship to Home

Walgreens Pharmacy

Choose a Store

Enter ZIP code, city and state, address or intersection.

Find a store

Notes to Pharmacy staff

e.g. Please add bubble gum flavoring to the amoxicillin

300 characters remaining

Review Your Rx

ZOLOFT 100MG TABLETS

DoctorLastName • Qty: 10 • John Doe

Remove

Need help? Call customer service at (877) 250-5823

‹ Add more prescriptions

Add more prescriptions

Submit request

But prescription information is missed!

(the user’s full name was not redacted on the previous page)

Walgreens makes thorough use of redaction

Session recordings are widespread

- 14+ analytics company offer recording services
  - Present on 99,174 of the top 1 million sites

- Evidence of recording on 7,918 sites.
  - Likely a lower bound as recording scripts sample users

Session recording present on ~1 - 10% of the top 1 million sites. We found several severe PII leaks after manually reviewing ~30 sites.

→ **How many more leaks are out there?**

What can we do?

1. Just keep measuring?

Will public backlash be enough? (Probably not)

2. Try to plug holes in browsers?

Limit autofill? → Sure

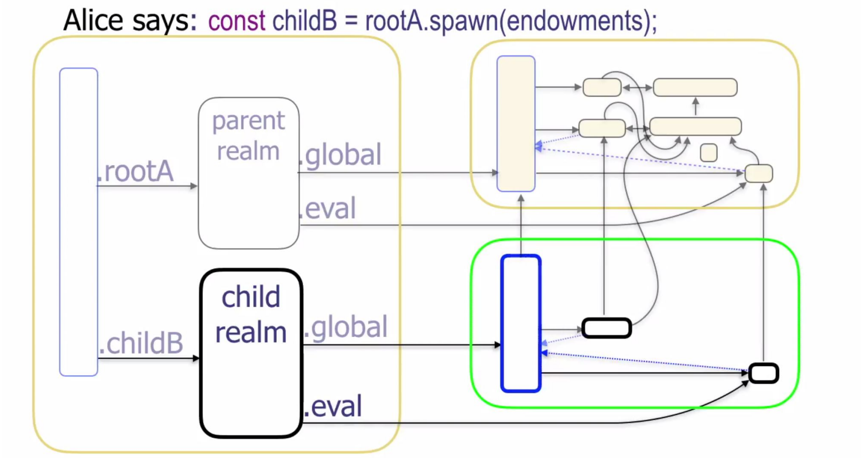
Limit social login sniffing? → How?

3. Push for regulation?

Hopeful in Europe, but what about the rest of the world?

Possible direction: Better JS confinement

Frozen Realms



(https://github.com/tc39/proposal-frozen-realms)

COWL

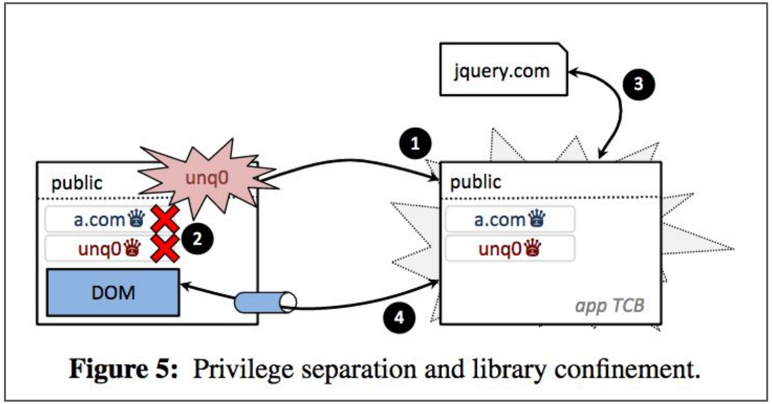
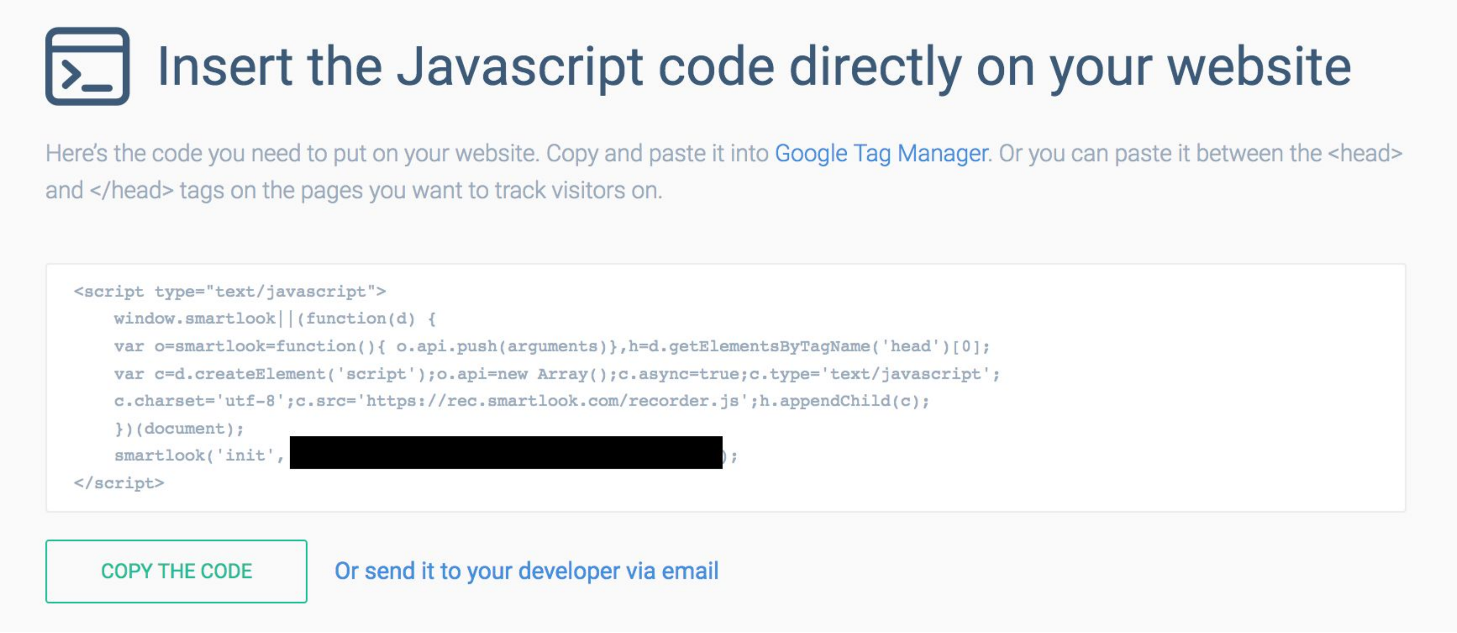


Figure 5: Privilege separation and library confinement.

(https://www.usenix.org/node/186158)

→ Problem: Requires the cooperation of sites ←

Possible direction: Better JS confinement



Insert the Javascript code directly on your website

Here's the code you need to put on your website. Copy and paste it into Google Tag Manager. Or you can paste it between the <head> and </head> tags on the pages you want to track visitors on.

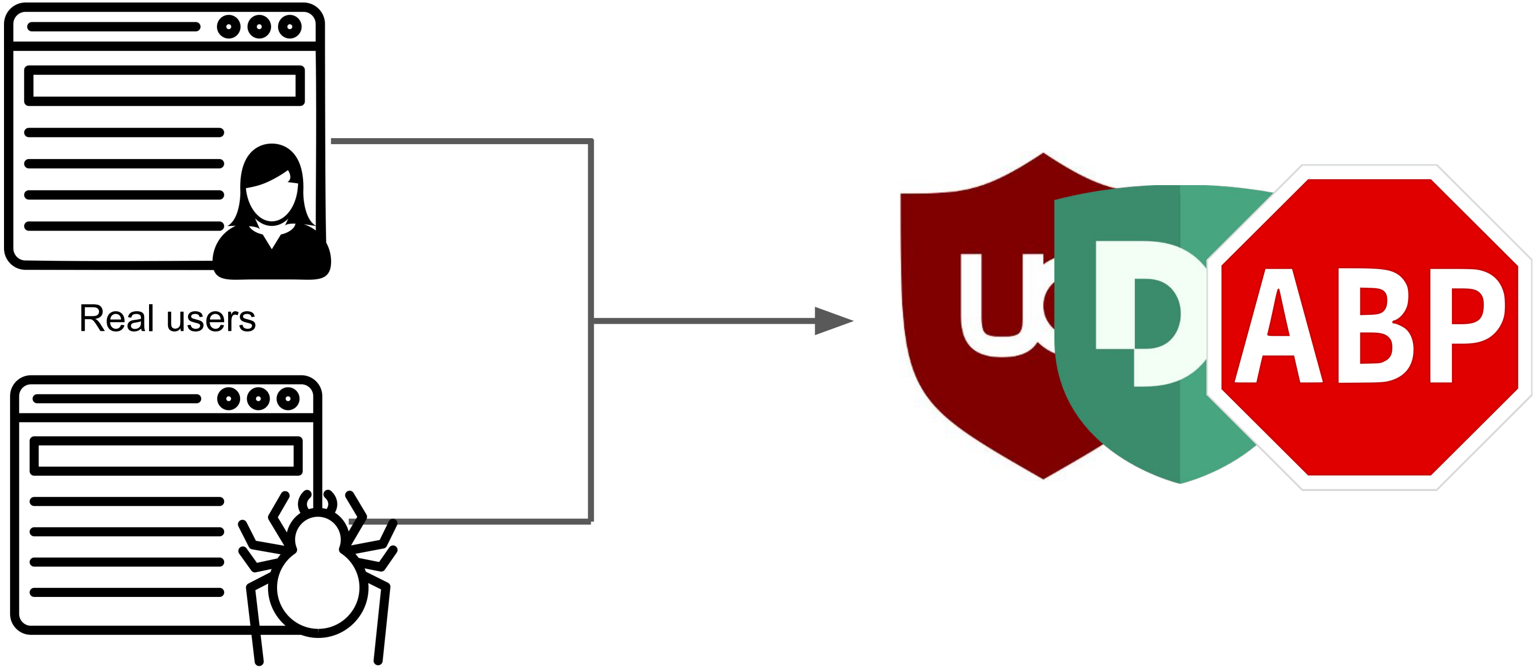
```
<script type="text/javascript">
    window.smartlook||(function(d) {
    var o=smartlook=function(){ o.api.push(arguments)},h=d.getElementsByTagName('head')[0];
    var c=d.createElement('script');o.api=new Array();c.async=true;c.type='text/javascript';
    c.charset='utf-8';c.src='https://rec.smartlook.com/recorder.js';h.appendChild(c);
    })(document);
    smartlook('init', [redacted]);
</script>
```

→ Problem: Requires the cooperation of sites ←

Possible direction: Measurement + Blocking

Detect invasive scripts

Build better blocklists



Real users

Crawlers

Possible direction: Measurement + Blocking

Problems:

- Broken sites
- Obfuscation
- User privacy concerns

How can we stop invasive web tracking?

1. **Just keep measuring?**
2. **Try to plug holes in browsers?**
3. **Push for regulation?**
4. **Work on confinement solutions?**
5. **Detect and block invasive scripts?**

Research: https://freedom-to-tinker.com/tag/noboundaries/

Me: https://senglehardt.com | senglehardt@mozilla.com