FREEDOM TO TINKER
research and expert commentary on digital technologies in public life

No boundaries: Exfiltration of personal data by session-replay scripts

NOVEMBER 15, 2017 BY STEVEN ENGLEHARDT LEAVE A COMMENT

This is the first post in our "No Boundaries" series, in which we reveal how third-party scripts on websites have been extracting personal information in increasingly intrusive ways. [0]
by Steven Englehardt, Gunes Acar, and Arvind Narayanan

You may know that most websites have third-party analytics scripts that record which pages you visit and the searches you make. But lately, more and more sites use "session replay" scripts. These scripts record your keystrokes, mouse movements, and scrolling behavior, along with the entire contents of the pages you visit, and send them to third-party servers. Unlike typical analytics services that provide aggregate statistics, these scripts are intended for the recording and playback of individual browsing sessions, as if someone is looking over your shoulder.

The stated purpose of this data collection includes gathering insights into how users interact with websites and discovering broken or confusing pages. However the extent of data collected by these services far exceeds user expectations [1]; text typed into forms is collected before the user submits the form, and precise mouse movements are saved, all without any visual indication to the user. This data can't reasonably be expected to be kept anonymous. In fact, some companies allow publishers to **explicitly link** recordings to a user's real identity.

For this study we analyzed seven of the top session replay companies (based on their relative popularity in our measurements [2]). The services studied are Yandex, FullStory, Hotjar, UserReplay, Smartlook, Clicktale, and SessionCam. We found these services in use on 482 of the Alexa top 50,000 sites.

This video shows the **"*co-browse*" feature** of one company, where the publisher can watch user sessions live.

What can go wrong? In short, a lot.

Collection of page content by third-party replay scripts may cause sensitive information such as medical conditions, credit card details and other personal information displayed on a page to leak to the third-party as part of the recording. This may expose users to identity theft, online scams, and other unwanted behavior. The same is true for the collection of user inputs during checkout and registration processes.

The replay services offer a combination of manual and automatic redaction tools that allow publishers to exclude sensitive information from recordings. However, in order for leaks to be avoided, publishers would need to diligently check and scrub all pages which display or accept user information. For dynamically generated sites, this process would involve inspecting the underlying web application's server-side code. Further, this process would need to be repeated every time a site is updated or the web application that powers the site is changed.

A thorough redaction process is actually a requirement for several of the recording services, which **explicitly forbid** the collection of user data. This negates the core premise of these session replay scripts, who market themselves as plug and play. For example, Hotjar's homepage **advertises**: "*Set up Hotjar with one script in a **matter of seconds***" and Smartlook's sign-up procedure features their script tag next to a timer with the tagline "*every minute you lose is a lot of video*".

To better understand the effectiveness of these redaction practices, we set up test pages and installed replay scripts from six of the seven companies [3]. From the results of these tests, as well as an analysis of a number of live sites, we highlight four types of vulnerabilities below:

1. Passwords are included in session recordings. All of the services studied attempt to prevent password leaks by automatically excluding password input fields from recordings. However, mobile-friendly login boxes that use text inputs to store unmasked passwords are not redacted by this rule, unless the publisher manually adds redaction tags to exclude them. We found at least one website where the password entered into a registration form leaked to SessionCam, even if the form is never submitted.

2. Sensitive user inputs are redacted in a partial and imperfect way. As users interact with a site they will provide sensitive data during account creation, while making a purchase, or while searching the site. Session recording scripts can use keystroke or input element loggers to collect this data.

All of the companies studied offer some mitigation through automated redaction, but the coverage offered varies greatly by provider. UserReplay and SessionCam replace all user input with an equivalent length masking text, while FullStory, Hotjar, and Smartlook exclude specific input fields by type. We

Freedom to Tinker is hosted by Princeton's Center for Information Technology Policy, a research center that studies digital technologies in public life. Here you'll find comment and analysis from the digital frontier, written by the Center's faculty, students, and friends.

CITP CENTER FOR INFORMATION TECHNOLOGY POLICY

Search this website … Search

What We Discuss

AACS bitcoin CD Copy Protection censorship CITP Competition Computing in the Cloud Copyright Cross-Border Issues cybersecurity policy DMCA DRM Education Events Facebook FCC Government Government transparency Grokster Case Humor Innovation Policy Law Managing the Internet Media Misleading Terms NSA Online Communities Patents Peer-to-Peer Predictions Princeton Privacy Publishing Recommended Reading Secrecy Security Spam Super-DMCA surveillance Tech/Law/Policy Blogs Technology and Freedom Virtual Worlds Voting Wiretapping WPM

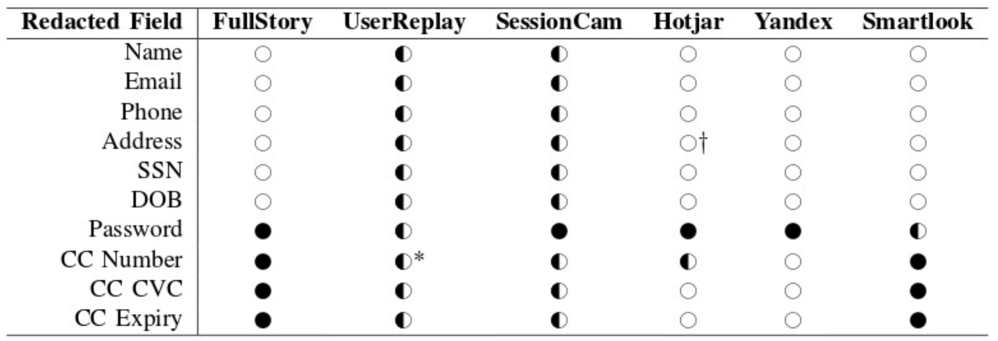
Contributors

Select Author...

Archives by Month

- **2017:** J F M A M J J A S O N D
- **2016:** J F M A M J J A S O N D
- **2015:** J F M A M J J A S O N D
- **2014:** J F M A M J J A S O N D
- **2013:** J F M A M J J A S O N D
- **2012:** J F M A M J J A S O N D
- **2011:** J F M A M J J A S O N D
- **2010:** J F M A M J J A S O N D
- **2009:** J F M A M J J A S O N D
- **2008:** J F M A M J J A S O N D
- **2007:** J F M A M J J A S O N D
- **2006:** J F M A M J J A S O N D

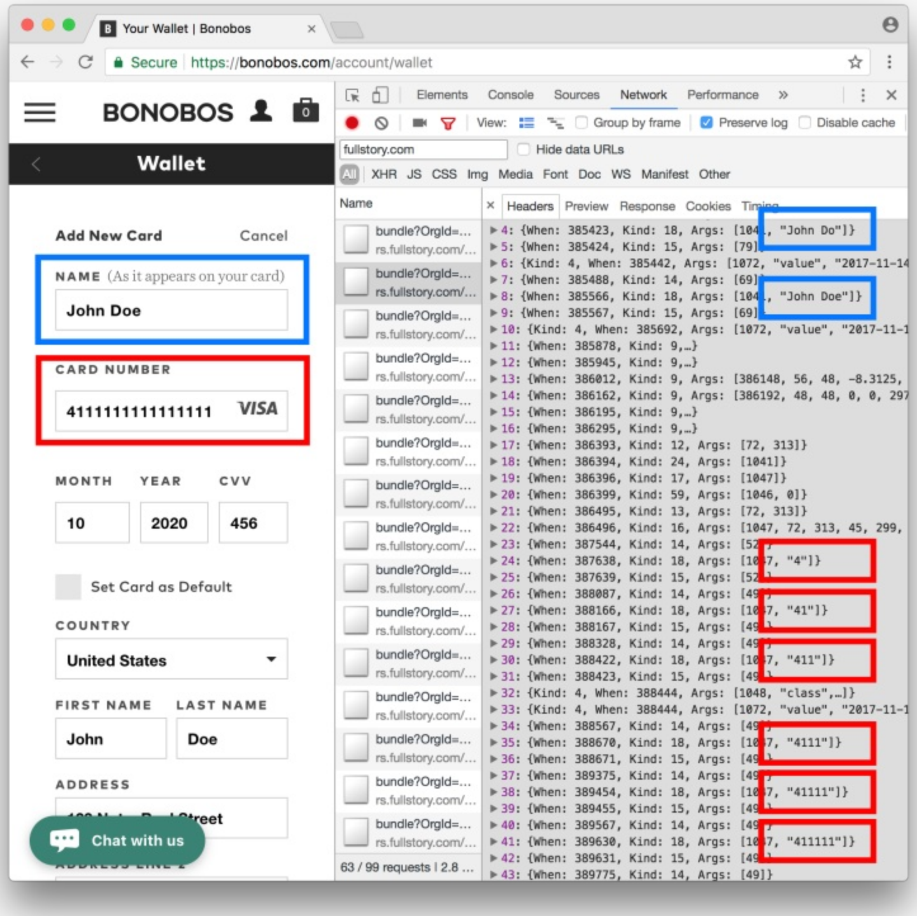
summarize the redaction of other fields in the table below.



Redacted Field	FullStory	UserReplay	SessionCam	Hotjar	Yandex	Smartlook
Name	○	◐	◐	○	○	○
Email	○	◐	◐	○	○	○
Phone	○	◐	◐	○	○	○
Address	○	◐	◐	○†	○	○
SSN	○	◐	◐	○	○	○
DOB	○	◐	◐	○	○	○
Password	●	◐	●	●	●	◐
CC Number	●	◐*	◐	◐	○	●
CC CVC	●	◐	◐	○	○	●
CC Expiry	●	◐	◐	○	○	●

Summary of the automated redaction features for form inputs enabled by default from each company. **Filled circle:** Data is excluded; **Half-filled circle:** equivalent length masking; **Empty circle:** Data is sent in the clear
* UserReplay sends the last 4 digits of the credit card field in plain text
† Hotjar masks the street address portion of the address field.

Automated redaction is imperfect; fields are redacted by input element type or heuristics, which may not always match the implementation used by publishers. For example, FullStory redacts credit card fields with the `autocomplete` attribute set to `cc-number`, but will collect any credit card numbers included in forms without this attribute.



Your Wallet | Bonobos

Secure | https://bonobos.com/account/wallet

BONOBOS

Wallet

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

COUNTRY: United States

FIRST NAME: John — LAST NAME: Doe

ADDRESS: 100 North Pool Street

Chat with us

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/...

Headers Preview Response Cookies Timing

```
▶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]}
```

63 / 99 requests | 2.8 ...

The account page of the clothing store Bonobos leaks full credit card details to FullStory. The screenshot of Chrome's network inspector shows the leaked data being sent letter-by-letter as it is typed. The user's full credit card number, expiration, CVV number, name, and billing address are leaked on this page. Email address and gift card numbers are among the other types of data leaked on Bonobos site.

To supplement automated redaction, several of the session recording companies, including **Smartlook**, **Yandex**, **FullStory**, **SessionCam**, and **Hotjar** allow sites to further specify inputs elements to be excluded from the recording. To effectively deploy these mitigations a publisher will need to actively audit every input element to determine if it contains personal data. This is complicated, error prone and costly, especially as a site or the underlying web application code changes over time. For instance, the financial service site fidelity.com has several redaction rules for Clicktale that involve **nested tables and child elements referenced by their index**. In the next section we further explore these challenges.

A safer approach would be to mask or redact all inputs by default, as is done by UserReplay and SessionCam, and allow whitelisting of known-safe values. Even fully masked inputs provide imperfect protection. For example, the masking used by UserReplay and Smartlook leaks the length of the user's password

- **2005:** J F M A M J J A S O N D
- **2004:** J F M A M J J A S O N D
- **2003:** J F M A M J J A S O N D
- **2002:** J F M A M J J A S O N D

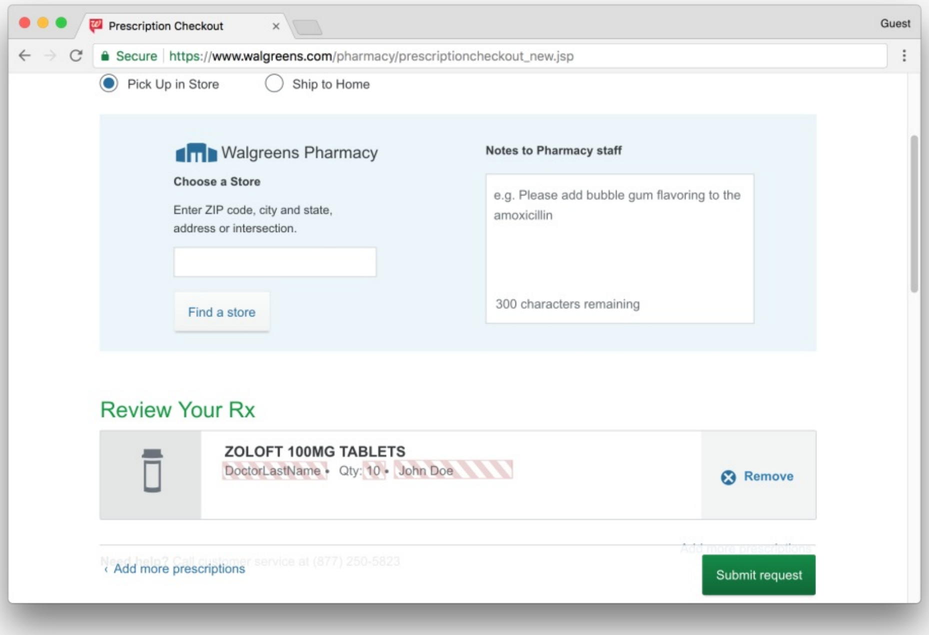
author log in

3. Manual redaction of personally identifying information displayed on a page is a fundamentally insecure model. In addition to collecting user inputs, the session recording companies also collect rendered page content. Unlike user input recording, none of the companies appear to provide automated redaction of displayed content by default; all displayed content in our tests ended up leaking.

Instead, session recording companies expect sites to manually label all personally identifying information included in a rendered page. Sensitive user data has a number of avenues to end up in recordings, and small leaks over several pages can lead to a large accumulation of personal data in a single session recording.

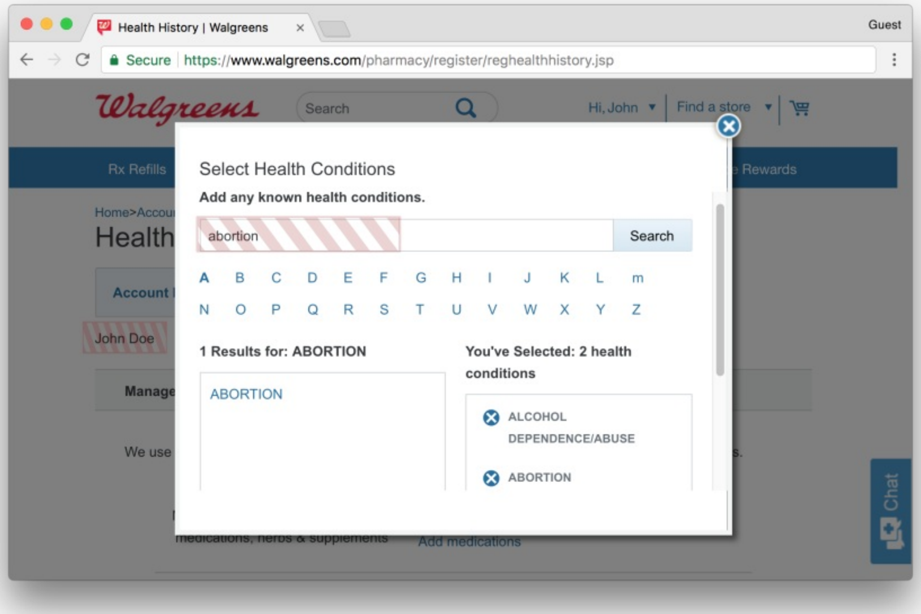
For recordings to be completely free of personal information, a site’s web application developers would need to work with the site’s marketing and analytics teams to iteratively scrub personally identifying information from recordings as it’s discovered. Any change to the site design, such as a change in the class attribute of an element containing sensitive information or a decision to load private data into a different type of element requires a review of the redaction rules.

As a case study, we examine the pharmacy section of Walgreens.com, which embeds FullStory. Walgreens makes extensive use of manual redaction for both displayed and input data. Despite this, we find that sensitive information including medical conditions and prescriptions are leaked to FullStory alongside the names of users.



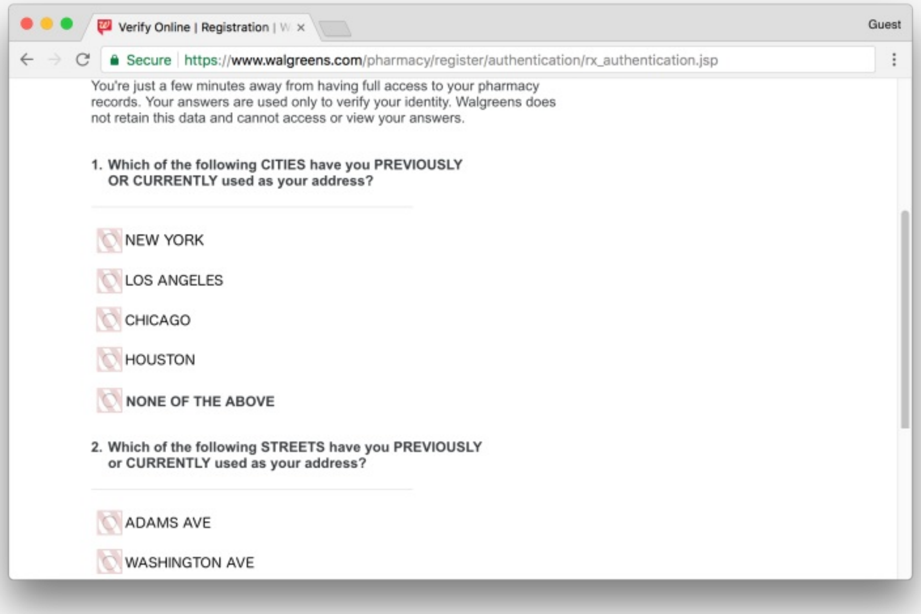
Prescription Checkout
Guest
Secure | https://www.walgreens.com/pharmacy/prescriptioncheckout_new.jsp
● Pick Up in Store
○ Ship to Home
Walgreens Pharmacy
Choose a Store
Enter ZIP code, city and state, address or intersection.
Find a store
Notes to Pharmacy staff
e.g. Please add bubble gum flavoring to 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

The above image shows a prescription request for the anti-depressant drug, Zoloft. During the process of creating the request, the name of the prescribed drug is leaked to FullStory [4]. Manual redaction was used to exclude the user’s name, their doctor’s name, and the quantity of medicine from the recording (marked in the image by a striped overlay). However, the user’s full name was leaked earlier in the process (not shown in this image), which allows anyone with access to the recording to associate this prescription with the user’s real identity.



Health History | Walgreens
Guest
Secure | https://www.walgreens.com/pharmacy/register/reghealthhistory.jsp
Walgreens
Search
Hi, John
Find a store
Rx Refills
Rewards
Home>Accou
Health
Account
John Doe
Manage
We use
medications, herbs & supplements
Add medications
Chat
Select Health Conditions
Add any known health conditions.
abortion
Search
A B C D E F G H I J K L m
N O P Q R S T U V W X Y Z
1 Results for: ABORTION
ABORTION
You've Selected: 2 health conditions
ALCOHOL DEPENDENCE/ABUSE
ABORTION

Walgreens allows users to enter their “Health History”, which can include other prescriptions and health conditions that may be relevant to prescription requests. During this process, most of the user’s personal and health information are excluded from FullStory’s recording through manual redaction. However, the process leaks the selected medicine and health conditions, the latter of which is shown above.



Verify Online | Registration | W

Guest

Secure https://www.walgreens.com/pharmacy/register/authentication/rx_authentication.jsp

You're just a few minutes away from having full access to your pharmacy records. Your answers are used only to verify your identity. Walgreens does not retain this data and cannot access or view your answers.

1. Which of the following CITIES have you PREVIOUSLY OR CURRENTLY used as your address?

○ NEW YORK

○ LOS ANGELES

○ CHICAGO

○ HOUSTON

○ NONE OF THE ABOVE

2. Which of the following STREETS have you PREVIOUSLY or CURRENTLY used as your address?

○ ADAMS AVE

○ WASHINGTON AVE

During account signup, Walgreens requires a user to verify their identity by asking a standard set of identity verification questions. The selection options for these questions, which may reveal the user's personal information, are displayed on the page and are transferred to FullStory. Additionally, the mouse tracking feature of FullStory will likely reveal the user's selection, even though the radio button selection is redacted. The inclusion of this data in recordings directly contradicts the statement at the top of the page: "Walgreens does not retain this data and cannot access or view your answers".

We do not present the above examples to point fingers at a certain website. Instead, we aim to show that the redaction process can fail even for a large publisher with a strong, legal incentive to protect user data. We observed similar personal information leaks on other websites, including on the checkout pages of Lenovo [5]. Sites with less resources or less expertise are even more likely to fail.

4. Recording services may fail to protect user data. Recording services increase the exposure to data breaches, as personal data will inevitably end up in recordings. These services must handle recording data with the same security practices with which a publisher would be expected to handle user data.

We provide a specific example of how recording services can fail to do so. Once a session recording is complete, publishers can review it using a dashboard provided by the recording service. The publisher dashboards for Yandex, Hotjar, and Smartlook all deliver playbacks within an HTTP page, even for recordings which take place on HTTPS pages. This allows an active man-in-the-middle to injecting a script into the playback page and extract all of the recording data. Worse yet, Yandex and Hotjar deliver the publisher page content over HTTP — data that was previously protected by HTTPS is now vulnerable to passive network surveillance.

The vulnerabilities we highlight above are inherent to full-page session recording. That's not to say the specific examples can't be fixed — indeed, the publishers we examined can patch their leaks of user data and passwords. The recording services can all use HTTPS during playbacks. But as long as the security of user data relies on publishers fully redacting their sites, these underlying vulnerabilities will continue to exist.

Does tracking protection help?

Two commonly used ad-blocking lists **EasyList** and **EasyPrivacy** do not block FullStory, Smartlook, or UserReplay scripts. EasyPrivacy has filter rules that block Yandex, Hotjar, ClickTale and SessionCam.

At least one of the five companies we studied (UserReplay) allows publishers to disable data collection from users who have Do Not Track (DNT) set in their browsers. We scanned the configuration settings of the Alexa top 1 million publishers using UserReplay on their homepages, and found that none of them chose to honor the DNT signal.

Improving user experience is a critical task for publishers. However it shouldn't come at the expense of user privacy.

End notes:

[0] We use the term 'exfiltrate' in this series to refer to the third-party data collection that we study. The term 'leakage' is sometimes used, but we eschew it, because it suggests an accidental collection resulting from a bug. Rather, our research suggests that while not necessarily malicious, the collection of sensitive personal data by the third parties that we study is inherent in their operation and is well known to most if not all of these entities. Further, there is an element of furtiveness; these data flows are not public knowledge and neither publishers nor third parties are not transparent about them.

[1] A **recent analysis** of the company Navistone, completed by Hill and Mattu for Gizmodo, explores how data collection prior to form submission exceeds user expectations. In this study, we show how analytics companies collect far more user data with minimal disclosure to the user. In fact, some services suggest

the first party sites simply **include a disclaimer** in their site's privacy policy or terms of service.

[2] We used **OpenWPM** to crawl the Alexa top 50,000 sites, visiting the homepage and 5 additional internal pages on each site. We use a two-step approach to detect analytics services which collect page content.

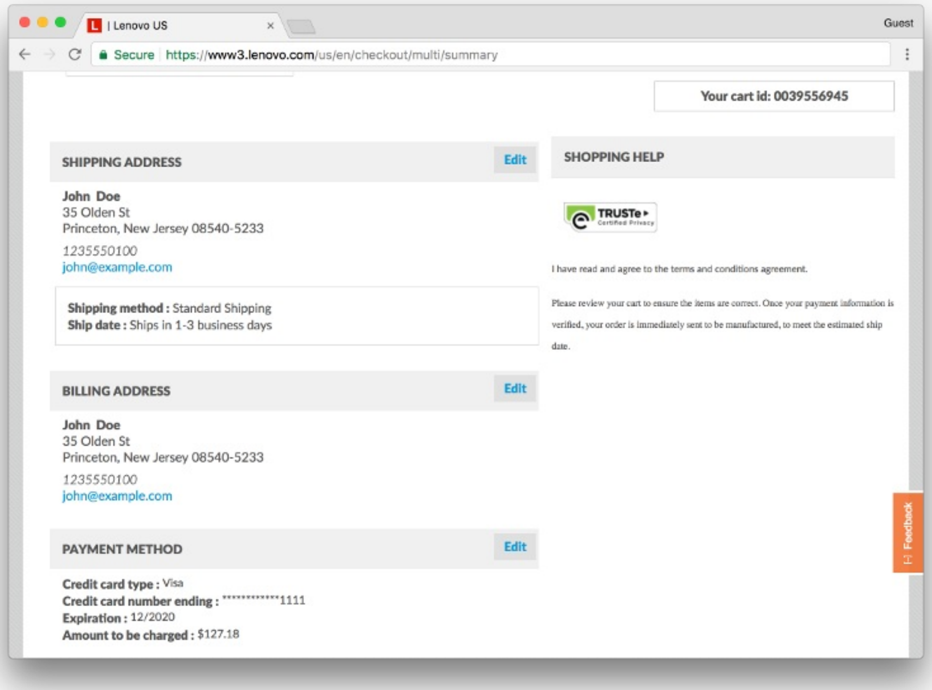
First, we inject a unique value into the HTML of the page and search for evidence of that value being sent to a third party in the page traffic. To detect values that may be encoded or hashed we use a detection methodology similar to **previous work on email tracking**. After filtering out leak recipients, we isolate pages on which at least one third party receives a large amount of data during the visit, but for which we do not detect a unique ID. On these sites, we perform a follow-up crawl which injects a 200KB chunk of data into the page and check if we observe a corresponding bump in the size of the data sent to the third party.

We found 482 sites on which either the unique marker was leaked to a collection endpoint from one of the services or on which we observed a data collection increase roughly equivalent to the compressed length of the injected chunk. We believe this value is a lower bound since many of the recording services offer the ability to **sample page visits**, which is compounded by our two-step methodology.

[3] One company (Clicktale) was excluded because we were unable to make the practical arrangements to analyze script's functionality at scale.

[4] FullStory's **terms and conditions** explicitly classify health or medical information, or any other information covered by HIPAA as sensitive data and asks customers to "not provide any Sensitive Data to FullStory."

[5] Lenovo.com is another example of a site which leaks user data in session recordings.



https://www3.lenovo.com/us/en/checkout/multi/summary

Your cart id: 0039556945

SHIPPING ADDRESS

John Doe
35 Olden St
Princeton, New Jersey 08540-5233
1235550100
john@example.com

Shipping method : Standard Shipping
Ship date : Ships in 1-3 business days

SHOPPING HELP

I have read and agree to the terms and conditions agreement.

Please review your cart to ensure the items are correct. Once your payment information is verified, your order is immediately sent to be manufactured, to meet the estimated ship date.

BILLING ADDRESS

John Doe
35 Olden St
Princeton, New Jersey 08540-5233
1235550100
john@example.com

PAYMENT METHOD

Credit card type : Visa
Credit card number ending : ************1111
Expiration : 12/2020
Amount to be charged : $127.18

On the final page of Lenovo's checkout procedure, the user's billing, shipping, and payment information is included in the text of the page. This information is thus included in the page source collected by FullStory as part of the recording process.

[6] We used the default scripts available to new accounts for 5 of the 6 providers. For UserReplay, we used a script taken from a live site and verified that the configuration options match the most common options found on the web.

FILED UNDER: PRIVACY   TAGGED WITH: ANALYTICS, DATA PRIVACY, PRIVACY, WEB PRIVACY, WPM

Speak Your Mind

Name

Email

Post Comment

Return to top of page