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

Gunes Acar, Steven Englehardt, and Arvind Narayanan
FTC PrivacyCon 2018 Talk Proposal

Authors

Gunes Acar

Steven Englehardt

Arvind Narayanan

Publication

No boundaries: Exfiltration of personal data by session-replay scripts
Steven Englehardt, Gunes Acar, and Arvind Narayanan
Ongoing work, blog post published on Freedom to Tinker

Available at:
https://freedom-to-tinker.com/2017/11/15/no-boundaries-exfiltration-of-personal-data-by-session-replay-scripts/

Talk Abstract

This talk will be based on our ongoing study of session replay scripts. These scripts record keystrokes, mouse movements, and scrolling behavior, along with the entire contents of the webpages visited, and send them to third-party servers. Unlike typical web analytics services that provide aggregate statistics, session replay scripts are intended for the recording and playback of individual browsing sessions, as if someone is looking over the visitor's shoulder.

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 be sent 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.

For this study we analyzed seven of the top session replay companies (based on their relative popularity in our measurements). 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.

Our methodology involves appending known values to the page and search for these values in the resulting network traffic. We attribute leaks to the organizations using the HTTP request's call stack.

The replay services offer a combination of manual and automatic tools that allow publishers to exclude or redact 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.

To better understand the effectiveness of these redaction practices, we set up test pages and installed replay scripts from six of the seven companies. In our talk we will highlight the vulnerabilities that we discovered in these experiments. These vulnerabilities include 1) leaking of users passwords to third-party session recording services, 2) imperfect redaction of sensitive user input, 3) incomplete exclusion of personally identifying information being displayed on the page.

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