OmniCrawl:
Comprehensive Measurement of Web Tracking with Real Desktop and Mobile Browsers

Darion Cassel¹, Su-Chin Lin², Alessio Buraggina³, William Wang⁴, Andrew Zhang⁵, Lujo Bauer¹, Hsu-Chun Hsiao², Limin Jia¹, and Timothy Libert⁶

¹ Carnegie Mellon University  
² National Taiwan University  
³ University of Miami  
⁴ University of Chicago  
⁵ University of Illinois Urbana-Champaign  
⁶ Google
Motivation: Mobile Browser Fingerprinting

Phones and computers: Complex configuration of **hardware** and **software**

**Browser fingerprinting:**
Tracker gathers enough characteristics to *uniquely identify* a browser across site visits
Motivation: Rise of Mobile Browsing

Mobile browsing overtook desktop browsing in 2016.

Mobile browsers’ sensor APIs used for tracking: Motion, Orientation, etc.

Few studies take the vantage point of mobile browsing.
Motivation: Rise of Mobile Browsing

Few studies take the vantage point of mobile browsing.

### Mobile browsing overtook desktop browsing in 2016

- Mobile browsers’ sensor APIs used for tracking: Motion, Orientation, etc.

#### RQ1: How do mainstream mobile and desktop browsers compare in terms of tracking and advertising?

#### RQ2: Is it ecologically valid to use emulated browsers in a web measurement study?
Motivation: Rise of Privacy-focused Browsers

- Many mobile browsers claim to enhance privacy

Firefox Focus:
  - “Block known [...] ad, analytics, and social trackers”

DuckDuckGo:
  - “stop advertisers from tracking you on the sites you visit”

Ghostery:
  - “block trackers”
  - “anonymizes your data to further protect your privacy”
Motivation: Rise of Privacy-focused Browsers

- Many mobile browsers claim to enhance privacy

Firefox Focus:

“Block known [ad, analytics, and social trackers]”

RQ3: How effective are privacy-focused browsers at blocking tracking and advertising?

RQ4: What are the strengths and weaknesses of individual privacy-focused browsers?
OmniCrawl Infrastructure Design

Goal: Synchronized crawling across multiple desktop and mobile browsers

1. URL
   - Command line
   - Main crawler: Appium

2. Request
   - Desktop browsers

3. Request
   - Mobile browsers

4. Response

5. Instrumented response

6. Browser API access logs

7. Done crawling

NA: 1x Windows 10, 1x Ubuntu 18.04

AS: 1x Windows 10, 1x Ubuntu 18.04

NA: 9x Motorola G5 Plus (Android 8.1)

AS: 9x ASUS ZenFone Max (Android 8.1)
Goal: Crawl popular (high-ranked) and lower-ranked websites

Tranco Ranking: Stability and manipulation resistance

Our 20K site dataset:

10,000 sites 1:1
10,000 sites 1:10 Uniform

Ensure dataset contains lesser-known sites

Rank 1 → Rank 10K → Rank 110K

Dataset Analysis Methodology

Goal: Measure third-party tracking-and-advertising and potential fingerprinting

10

(1) Request Type
1st-party | 3rd-party

(2) Request Classification
Tracking-and-advertising entities
Block lists and webxray

1st-party  |  3rd-party

(1) API Categories
Sensitive APIs: Audio, Automation, Canvas, etc.

(2) Fingerprinting Heuristics
For WebRTC, Audio, Font, Canvas, WebGL

Statistical testing methodology to establish significance
Summary of Key Results – RQ1

RQ1: How do mainstream mobile and desktop browsers compare in terms of tracking and advertising?

- Mobile-only & desktop-only 3rd-party TA entities → 0.45% of requests
  - Ecosystems of third-party tracking-and-advertising entities more homogenous than previously thought
  - Mobile → more accesses to APIs indicative of fingerprinting
    - +54% Screen API
    - +21% WebRTC FP-heuristics
Summary of Key Results – RQ2

RQ2: Is it ecologically valid to use emulated browsers in a web measurement study?

- OpenWPM-Mobile → significantly more requests and accesses
  - +6% 3rd-party tracking-and-advertising requests
  - +50% Plugin API accesses

- Emulated browsers may not be suitable for measuring third-party requests or browser APIs accesses
Summary of Key Results – RQ3

RQ3: How effective are privacy-focused browsers at blocking tracking and advertising?

- **Do not** uniformly reduce third-party tracking-and-advertising requests

- Most effective for **highly prevalent** entities:
  1. Google (-60%)
  2. Facebook (-66%)
  3. Adobe Systems (-56%)
RQ4: What are the strengths and weaknesses of individual privacy-focused browsers?

- **Generally** effective at reducing tracking-and-advertising... but

- **Vary widely** in effectiveness
  - e.g. 3TA requests: 13x
  - API accesses: -25% group

- **Deviate** from marketing claims
  - e.g. Tor “tracker blocking” unsupported
Recommendations for Studies and Vendors

- Future web measurement study design:
  - Avoid *emulated* browsers for mobile-specific measurements
  - Modify off-the-shelf browser *drivers* (esp. Selenium)

- Privacy-focused browser vendors:
  - Clarify concrete *implications* of marketing claims
  - Offer more *user control* over blocking comprehensiveness

More details in the paper & artifact available: [https://github.com/omnicrawl/](https://github.com/omnicrawl/)
OmniCrawl: Comprehensive Measurement of Web Tracking with Real Desktop and Mobile Browsers

Darion Cassel\textsuperscript{1}, Su-Chin Lin\textsuperscript{2}, Alessio Buraggina\textsuperscript{3}, William Wang\textsuperscript{4}, Andrew Zhang\textsuperscript{5}, Lujo Bauer\textsuperscript{1}, Hsu-Chun Hsiao\textsuperscript{2}, Limin Jia\textsuperscript{1}, and Timothy Libert\textsuperscript{6}

\textsuperscript{1} Carnegie Mellon University \quad \textsuperscript{2} National Taiwan University \quad \textsuperscript{3} University of Miami
\textsuperscript{4} University of Chicago \quad \textsuperscript{5} University of Illinois Urbana-Champaign \quad \textsuperscript{6} Google