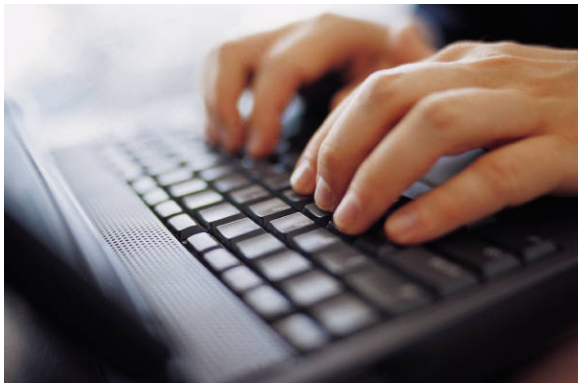
Detecting and Defending Against Third-Party Tracking on the Web

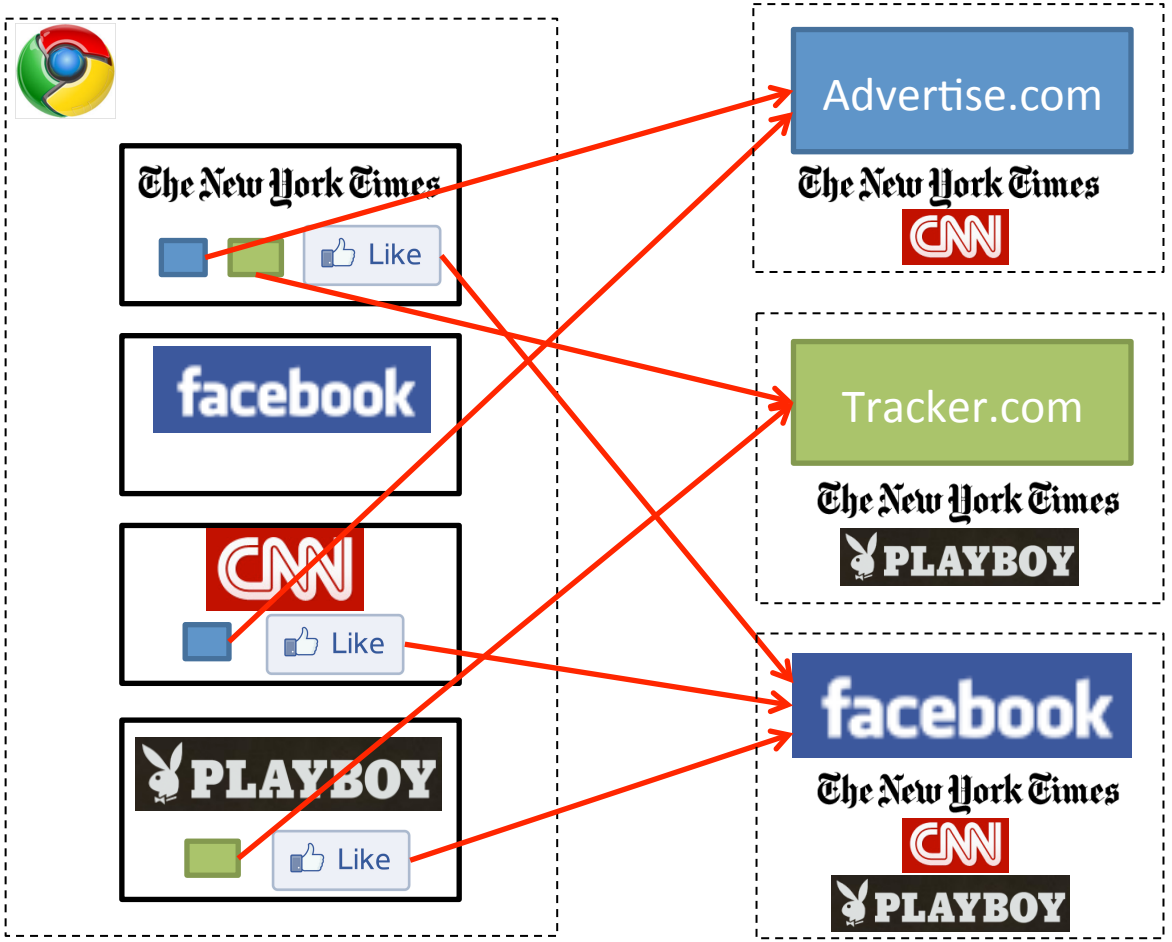
Franziska Roesner, Tadayoshi Kohno, David Wetherall

April 26, 2012
NSDI

Third-Party Web Tracking



Bigger browsing profiles
= increased value for trackers
= reduced privacy for users



(Hypothetical tracking relationships only.)

Tracking is Complicated

- Much discussion of tracking, but limited understanding of how it actually works.
- Our goals:
  - Understand the tracking ecosystem.
    - How is tracking actually done in the wild?
    - What kinds of browsing profiles do trackers compile?
    - How effective are defenses available to users?
  - Address gaps with new defense (ShareMeNot).

Outline

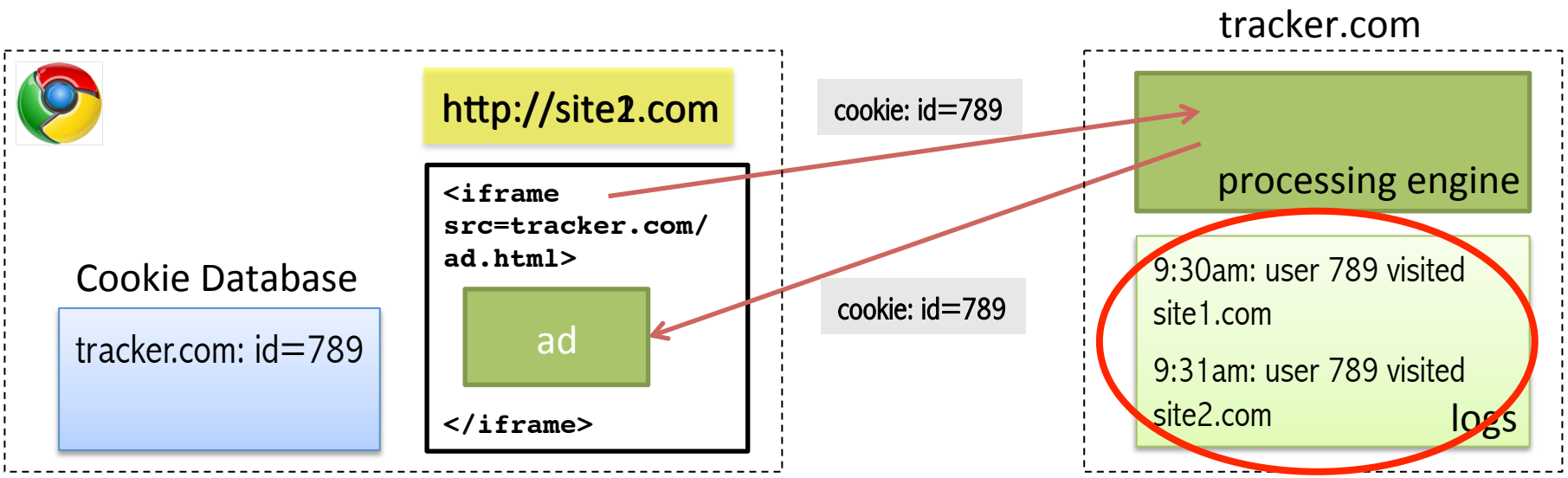
- How Tracking Works
  - Tracking Mechanisms
  - Tracking Taxonomy
- Measurements
- Defenses

Mechanisms Required By Trackers

- Ability to store user identity in the browser
  - Browser cookies
  - HTML5 LocalStorage and Flash cookies (LSOs)
  - Not considering more exotic storage mechanisms or approximate fingerprinting
- Ability to communicate visited page and user identity back to tracker
  - Identity: Cookies attached to requests
  - Visited page: HTTP referrers
  - Both: scripts that embed information in URLs

Tracking: The Simple Version

- **Within-Site**: First-party cookies are used to track repeat visits to a site.
- **Cross-Site**: Third-party cookies are used by trackers included in other sites to create profiles.



Our Tracking Taxonomy

Name	Scope	User Visits Directly?	Overview
N/A	Within-Site	Yes	Site does its own on-site analytics.
Analytics	Within-Site	No	Site uses third-party analytics engine (e.g., Google Analytics).
Vanilla	Cross-Site	No	Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).
Forced	Cross-Site	Yes (forced)	Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).
Referred	Cross-Site	No	Tracker relies on another cross-site tracker to leak unique identifier values.
Personal	Cross-Site	Yes	Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).

Evolution: Embedding analytics libraries

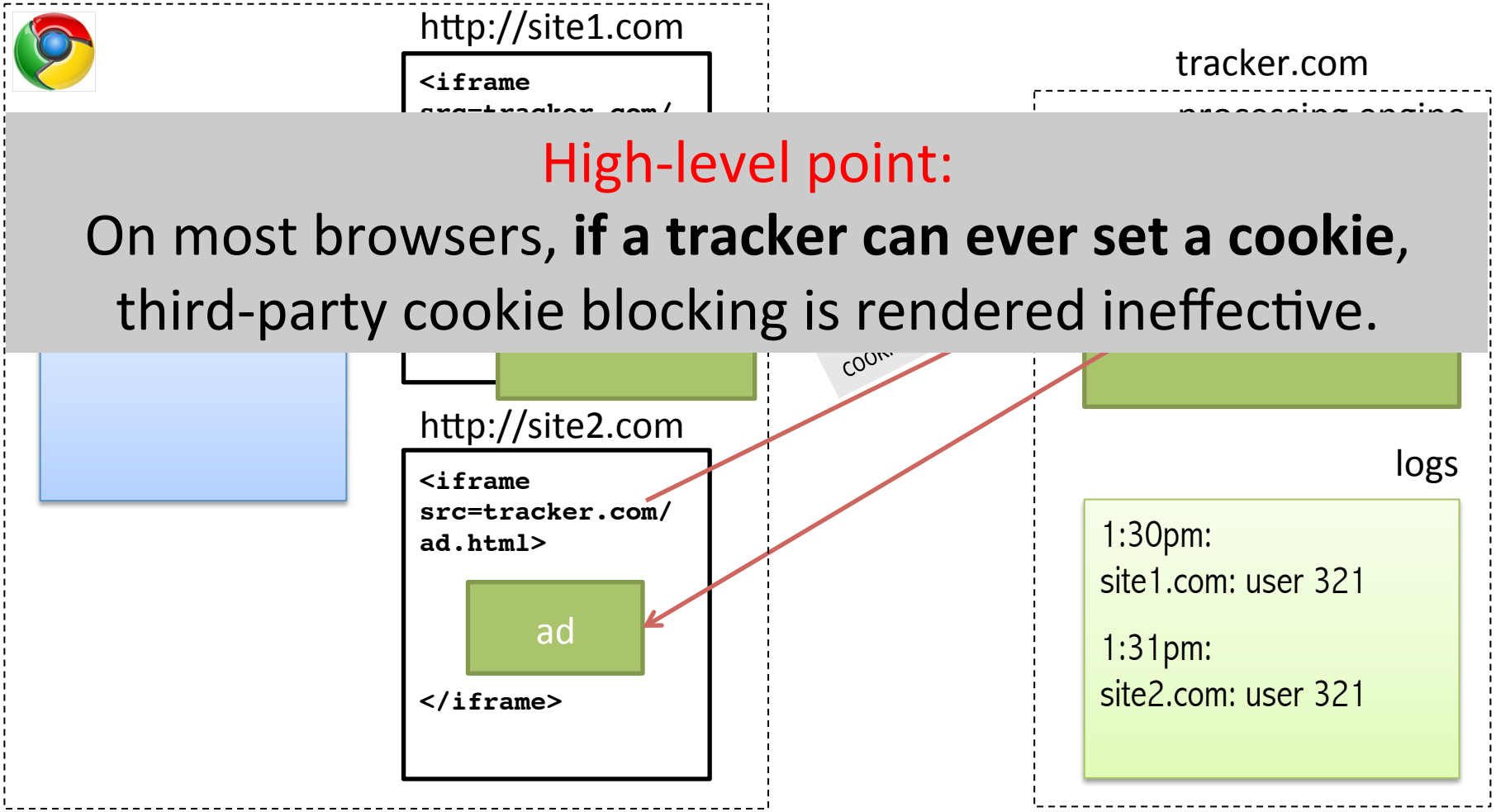
Evolution: Third-party cookie blocking

Quirks of Third-Party Cookie Blocking

- Option blocks the **setting** of third-party cookies: all browsers
- Option blocks the **sending** of third-party cookies: only Firefox

- Result: Once a third-party cookie is somehow set, it can be used (in most browsers).

Forced Tracking



Our Tracking Taxonomy

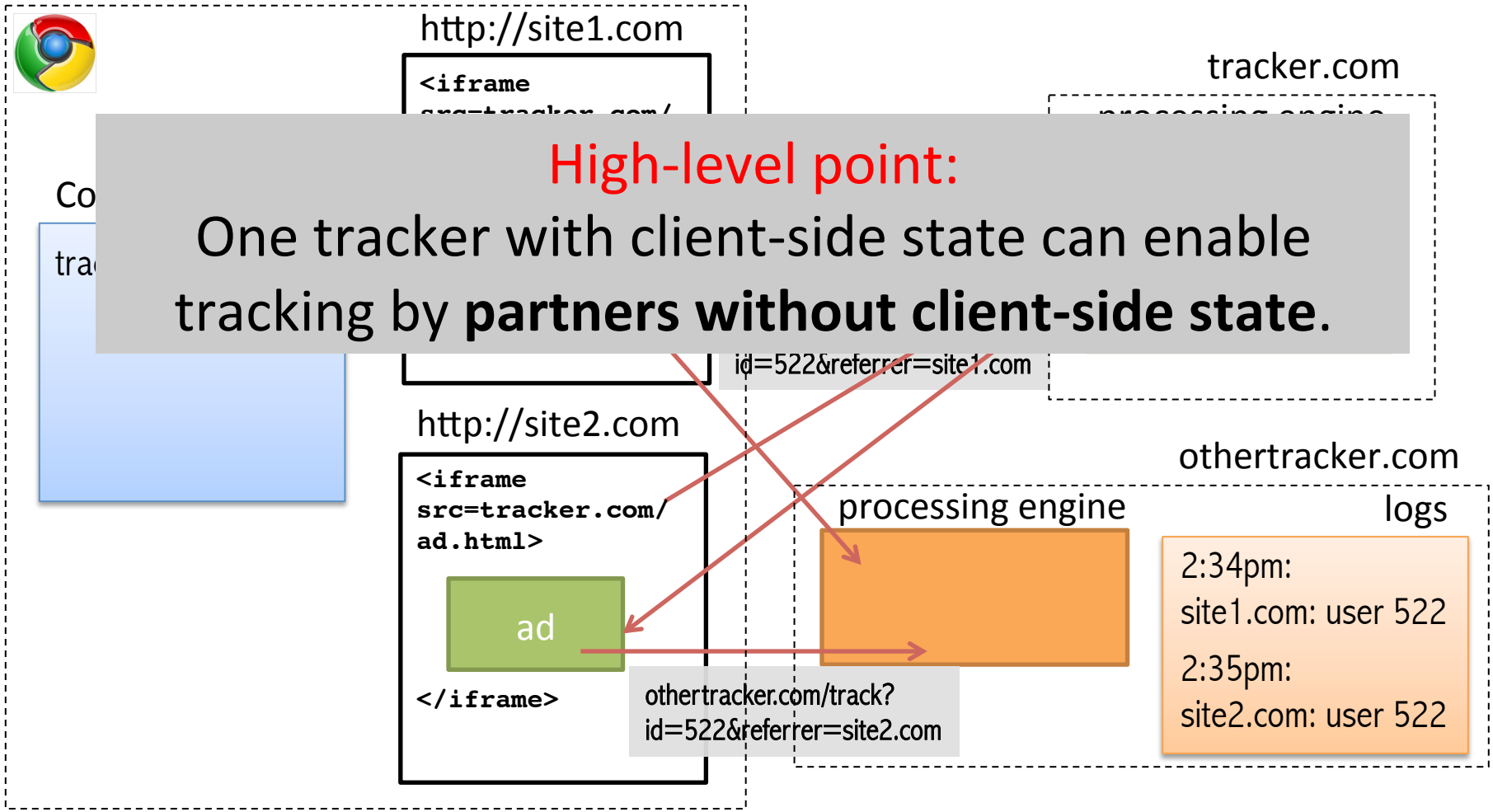
Type (Name)	Scope	User Visits Directly?	Overview
N/A	Within-Site	Yes	Site does its own on-site analytics.
Analytics	Within-Site	No	Site uses third-party analytics engine (e.g., Google Analytics).
Vanilla	Cross-Site	No	Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).
Forced	Cross-Site	Yes (forced)	Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).
Referred	Cross-Site	No	Tracker relies on another cross-site tracker to leak unique identifier values.
Personal	Cross-Site	Yes	Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).

Evolution: Embedding analytics libraries

Evolution: Third-party cookie blocking

Evolution: Complex ad networks

Referred Tracking



Our Tracking Taxonomy

Type (Name)	Scope	User Visits Directly?	Overview
N/A	Within-Site	Yes	Site does its own on-site analytics.
Analytics	Within-Site	No	Site uses third-party analytics engine (e.g., Google Analytics).
Vanilla	Cross-Site	No	Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).
Forced	Cross-Site	Yes (forced)	Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).
Referred	Cross-Site	No	Tracker relies on another cross-site tracker to leak unique identifier values.
Personal	Cross-Site	Yes	Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).

Evolution: Embedding analytics libraries

Evolution: Third-party cookie blocking

Evolution: Complex ad networks

Evolution: Social networks

Personal Tracking

- Just loading these buttons (not clicking on them) enables tracking.
- Users visit these sites directly.
- This tracking is often not anonymous (linked to accounts).

Our Tracking Taxonomy

Type (Name)	Scope	User Visits Directly?	Overview
N/A	Within-Site	Yes	Site does its own on-site analytics.
Analytics	Within-Site	No	Site uses third-party analytics engine (e.g., Google Analytics).
Vanilla	Cross-Site	No	Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).
Forced	Cross-Site	Yes (forced)	Site embeds third-party tracker that forced the user to visit directly (e.g., via popup).
Referred	Cross-Site	No	Tracker relies on another cross-site tracker to leak unique identifier values.
Personal	Cross-Site	Yes	Site embeds third-party tracker that the user otherwise visits directly (e.g., Facebook).

Evolution: Embedding analytics libraries

Evolution: Third-party cookie blocking

Evolution: Complex ad networks

Evolution: Social networks

Anonymous

Outline

- How Tracking Works
  - Tracking mechanisms
  - Tracking taxonomy
- Measurements
- Defenses

Measurement Tool: TrackingTracker

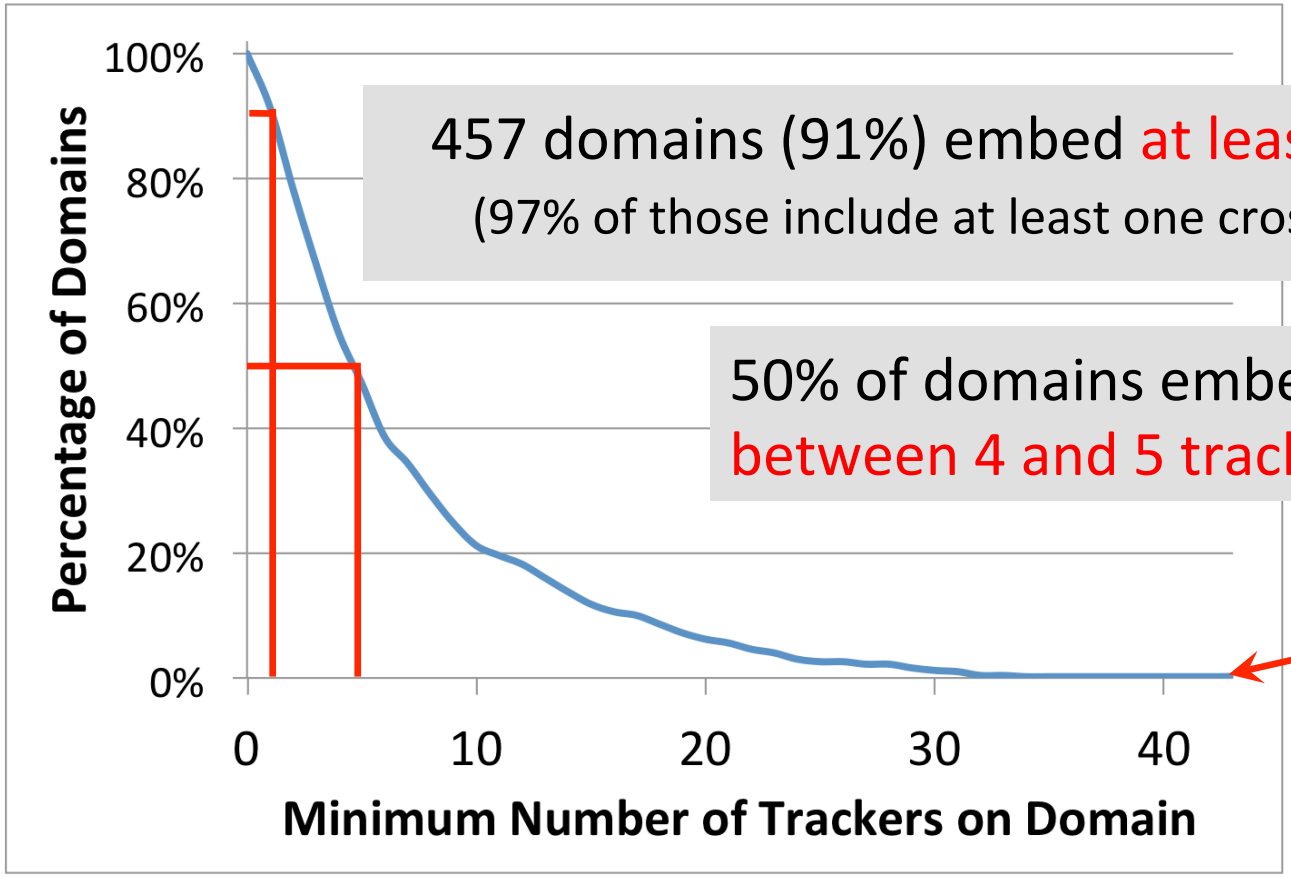
- Firefox add-on
- Based on taxonomy of client-side mechanisms
- Crawls the web, automatically categorizing trackers
- Monitors:
  - Third-party requests
  - Cookies, HTML5 LocalStorage, Flash LSOs
    (considers state that changes across clean measurement runs)
  - Identifier leaks

Measurement Study

- 3 data sets
  - Alexa Top 500
    - 5 pages per domain: main page and up to 4 links
  - Alexa Non-Top 500
    - Sites ranked #501, #601, #701, etc.
    - 5 pages per domain: main page and up to 4 links
  - AOL search logs
    - 300 unique queries for 35 random users

Tracking Prevalence (Top 500)

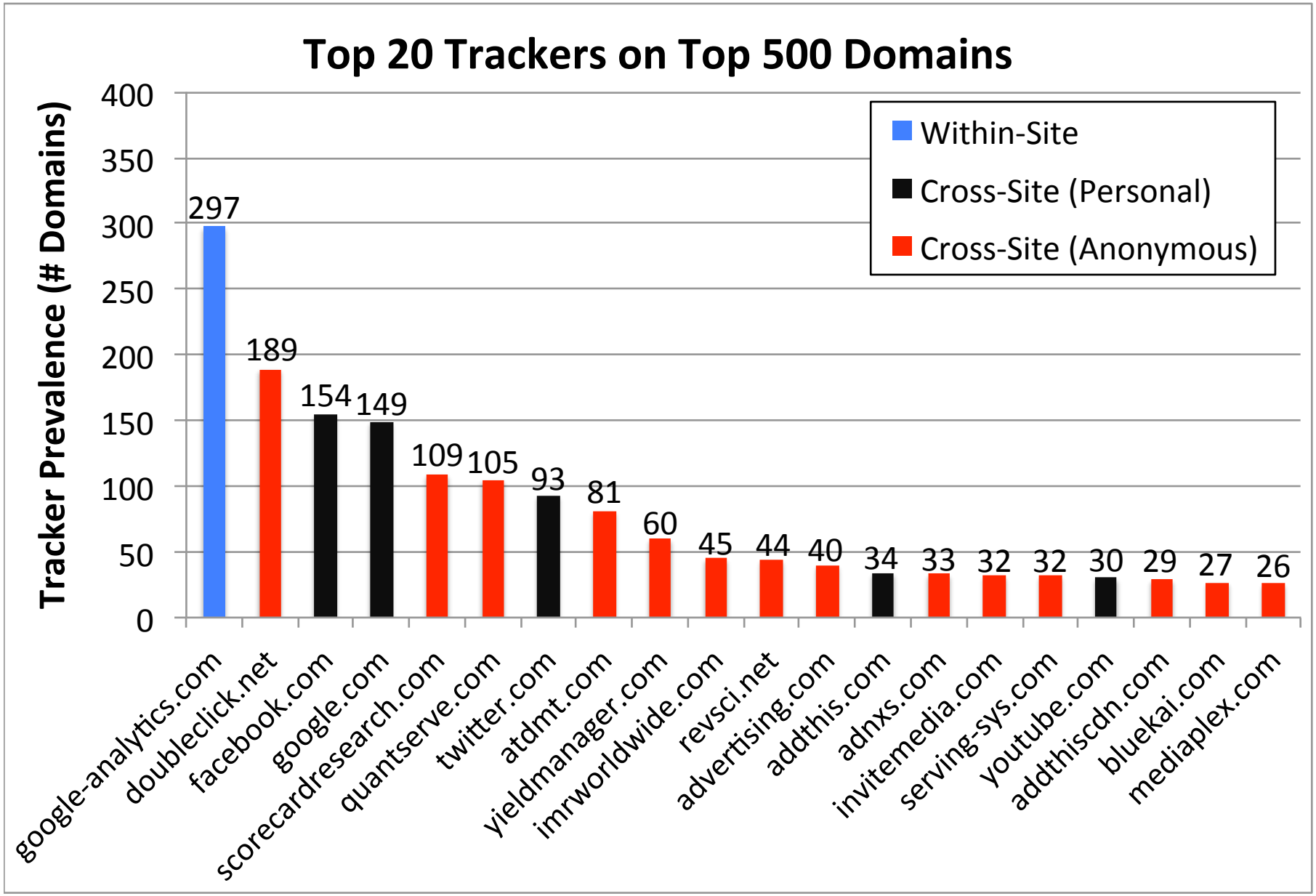
- 524 unique trackers on 500 domains

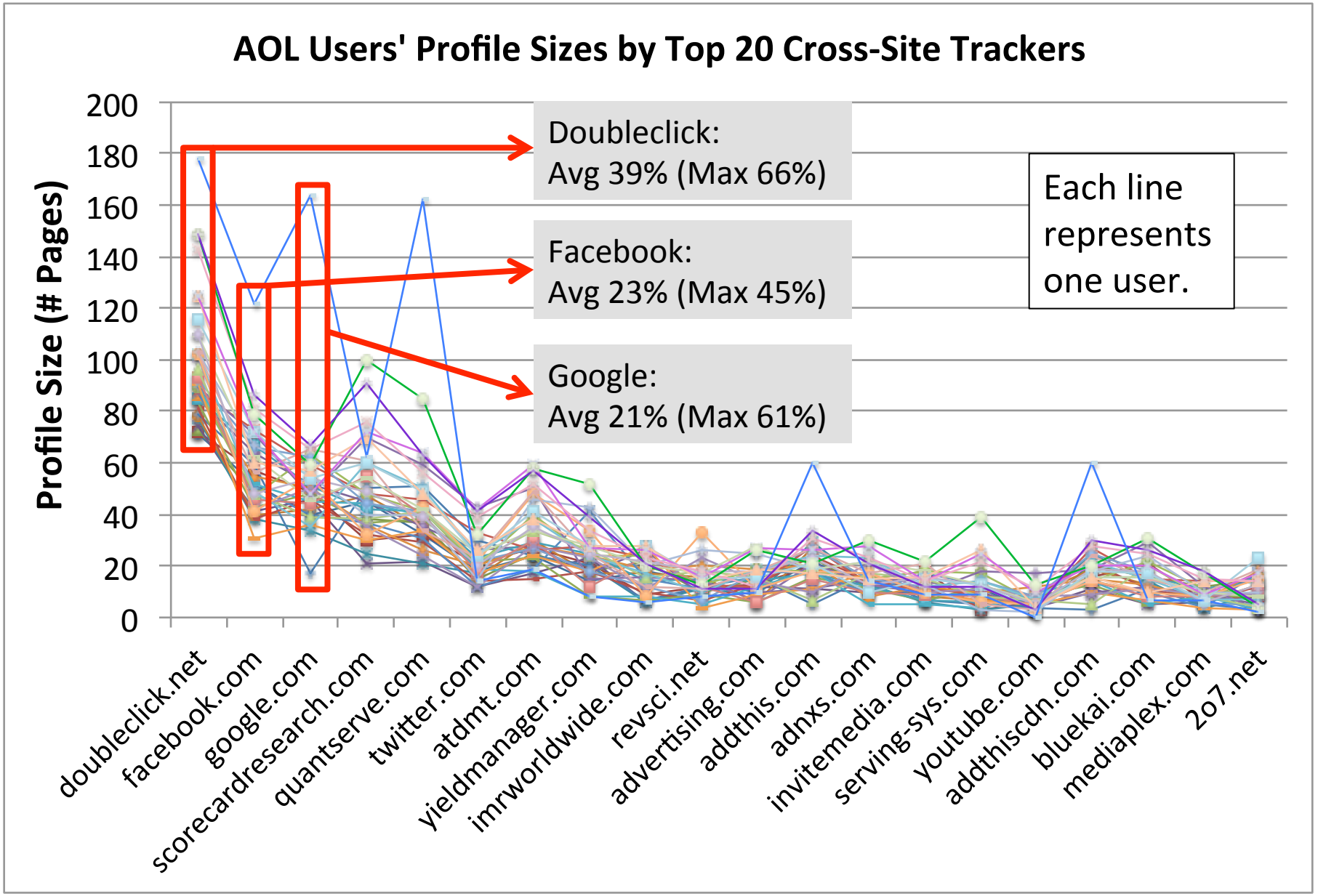


457 domains (91%) embed at least one tracker.
(97% of those include at least one cross-site tracker.)

50% of domains embed between 4 and 5 trackers.

One domain includes 43 trackers.


Top 20 Trackers on Top 500 Domains
Tracker Prevalence (# Domains)
400
350
300
250
200
150
100
50
0
Within-Site
Cross-Site (Personal)
Cross-Site (Anonymous)
297
189
154
149
109
105
93
81
60
45
44
40
34
33
32
32
30
29
27
26
google-analytics.com
doubleclick.net
facebook.com
google.com
scorecardresearch.com
quantserve.com
twitter.com
atdmt.com
yieldmanager.com
imrworldwide.com
revsci.net
advertising.com
addthis.com
adnxs.com
invitemedia.com
serving-sys.com
youtube.com
addthiscdn.com
bluekai.com
mediaplex.com


AOL Users' Profile Sizes by Top 20 Cross-Site Trackers
Profile Size (# Pages)
200
180
160
140
120
100
80
60
40
20
0
Doubleclick:
Avg 39% (Max 66%)
Facebook:
Avg 23% (Max 45%)
Google:
Avg 21% (Max 61%)
Each line represents one user.
doubleclick.net
facebook.com
google.com
scorecardresearch.com
quantserve.com
twitter.com
atdmt.com
yieldmanager.com
imrworldwide.com
revsci.net
advertising.com
addthis.com
adnxs.com
invitemedia.com
serving-sys.com
youtube.com
addthiscdn.com
bluekai.com
mediaplex.com
2o7.net

LocalStorage and Flash Cookies

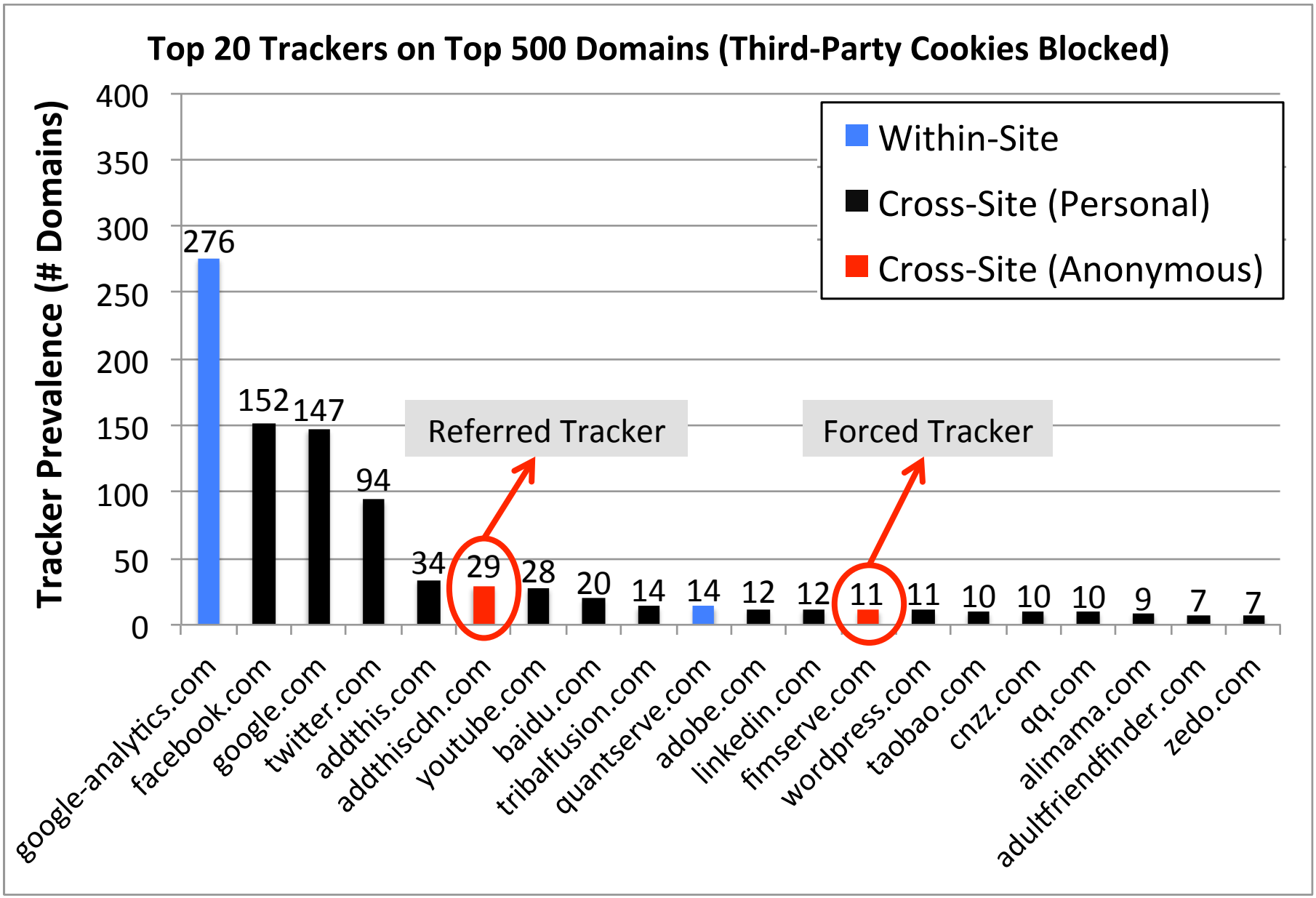
- Surprisingly little use of these mechanisms!
- Of 524 trackers on Alexa Top 500:
  - Only 5 set unique identifiers in LocalStorage
  - 35 set unique identifiers in Flash cookies
- Respawning:
  - LS → Cookie: 1 case; Cookie → LS: 3 cases
  - Flash→ Cookie: 6 cases; Cookie → Flash: 7 cases

Outline

- How Tracking Works
  - Tracking mechanisms
  - Tracking taxonomy
- Measurements
- Defenses

Defenses to Reduce Tracking

- We explore several in the paper:
  - Third-party cookie blocking 
  - Do Not Track header
  - Popup blocking
  - Clearing client-side state
  - Disabling JavaScript
  - Private browsing mode


Top 20 Trackers on Top 500 Domains (Third-Party Cookies Blocked)
Tracker Prevalence (# Domains)
400
350
300
250
200
150
100
50
0
Within-Site
Cross-Site (Personal)
Cross-Site (Anonymous)
Referred Tracker
Forced Tracker
276
152
147
94
34
29
28
20
14
14
12
12
11
11
10
10
10
9
7
7
google-analytics.com
facebook.com
google.com
twitter.com
addthis.com
addthiscdn.com
youtube.com
baidu.com
tribalfusion.com
quantserve.com
adobe.com
linkedin.com
fimserve.com
wordpress.com
taobao.com
cnzz.com
qq.com
alimama.com
adultfriendfinder.com
zedo.com

Personal Tracking Revisited

- Most popular, based on measurements:
  - Facebook, Google, Twitter, AddThis, YouTube, LinkedIn, Digg, Stumbleupon
- Third-party cookie blocking is ineffective.
- Existing browser extension solutions remove the buttons (undesirable to some users).
- Can we reduce tracking but allow use?

ShareMeNot

http://sharemenot.cs.washington.edu

- A browser extension that protects against tracking from third-party social media buttons while still allowing them to be used.
- Firefox version: removes cookies from relevant requests until user clicks button.
  - Similar: Priv3 Firefox add-on
- Chrome version: replace buttons with local stand-in button until user click.

Effectiveness of ShareMeNot (Top 500)

Tracker	Without ShareMeNot	With ShareMeNot
Facebook	154	9
Google	149	15
Twitter	93	0
AddThis	34	0
YouTube	30	0
LinkedIn	22	0
Digg	8	0
Stumbleupon	6	0

Summary

- Introduced taxonomy of tracking behavior for any client-side identifiers.
  - *Analytics, Vanilla, Forced, Referred, Personal*
- Studied tracking in the wild with browser measurements.
  - Revealed rich tracking ecosystem.
  - Results can assist informed broader discussions.
- Developed ShareMeNot, a new privacy-enhancing defense for personal tracking.

http://sharemenot.cs.washington.edu/