Third-Party Web Tracking

Franziska (Franzi) Roesner
franzi@cs.washington.edu

CSE 484 Guest Lecture
October 16, 2013

Who am I?

- 5th (& final) year PhD student advised by Yoshi
- I've worked on:
  - Automobile security
  - Third-party web tracking
  - Permission granting in smartphones (etc.)
  - Securing embedded user interfaces
  - Security/privacy for augmented reality

Today

- Background on web security
- Understanding web tracking
- Measuring web tracking
- Building new web tracking defenses

Same-Origin Policy

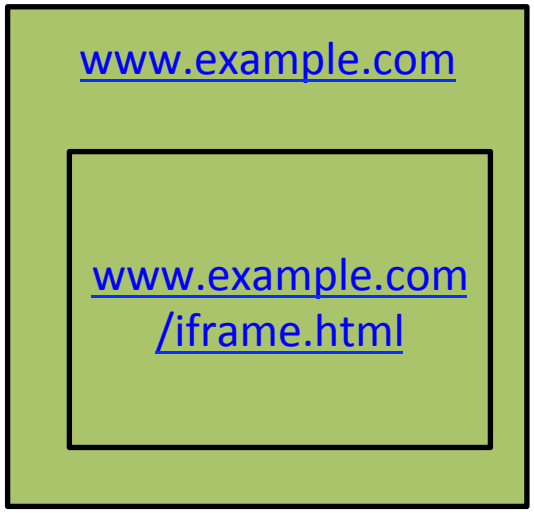
Website origin = (scheme, domain, port)

Compared URL	Outcome	Reason
http://www.example.com/dir/page.html	Success	Same protocol and host
http://www.example.com/dir2/other.html	Success	Same protocol and host
http://www.example.com:**81**/dir/other.html	Failure	Same protocol and host but different port
https://www.example.com/dir/other.html	Failure	Different protocol
http://**en.example.com**/dir/other.html	Failure	Different host
http://**example.com**/dir/other.html	Failure	Different host (exact match required)
http://**v2.www.example.com**/dir/other.html	Failure	Different host (exact match required)

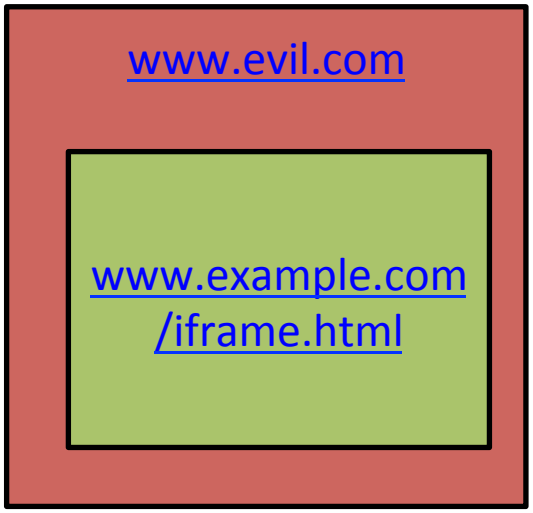
[Example thanks to Wikipedia.]

Same-Origin Policy (DOM)

- Only code from same origin can access HTML elements on another site (or in an iframe).



www.example.com (the parent) **can** access HTML elements in the iframe (and vice versa).



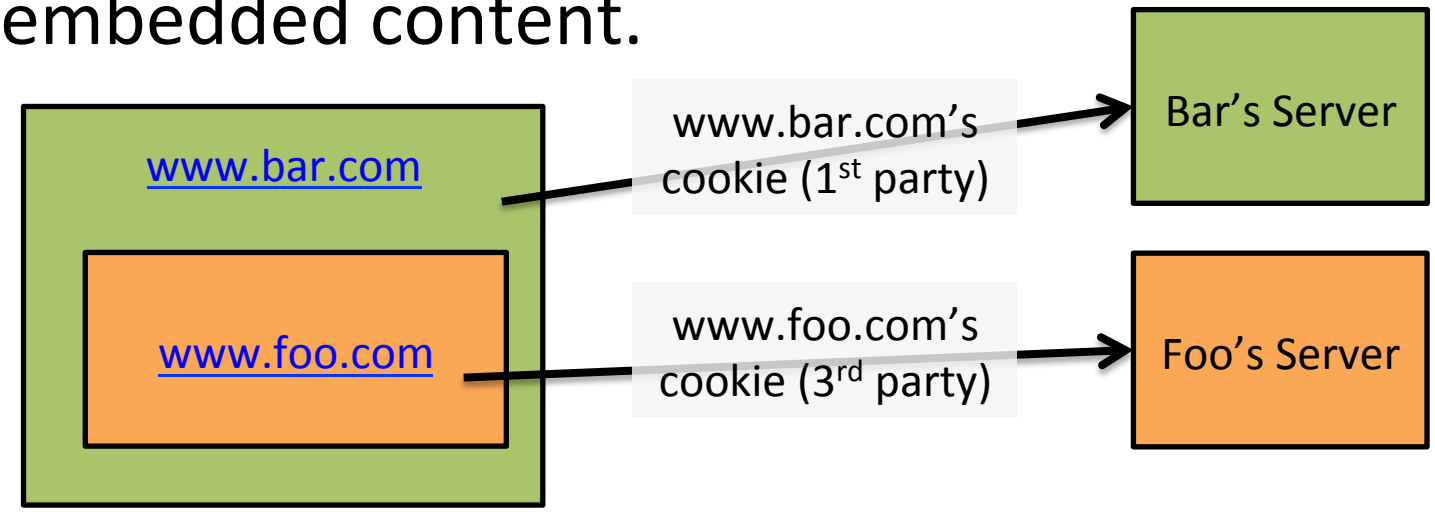
www.evil.com (the parent) **cannot** access HTML elements in the iframe (and vice versa).

Same-Origin Policy (Cookies)

- **For cookies:** Only code from same origin can read/write cookies associated with an origin.
  - Can be set via Javascript (`document.cookie=…`) or via `Set-Cookie` header in HTTP response.
  - Can narrow to subdomain/path (e.g., http://example.com can set cookie scoped to http://account.example.com/login.)

Same-Origin Policy (Cookies)

- Browsers automatically include cookies with HTTP requests.
- **First-party cookie:** belongs to top-level domain.
- **Third-party cookie:** belongs to domain of embedded content.



Same-Origin Policy (Scripts)

- When a website **includes a script**, that script runs in the context of the embedding website.

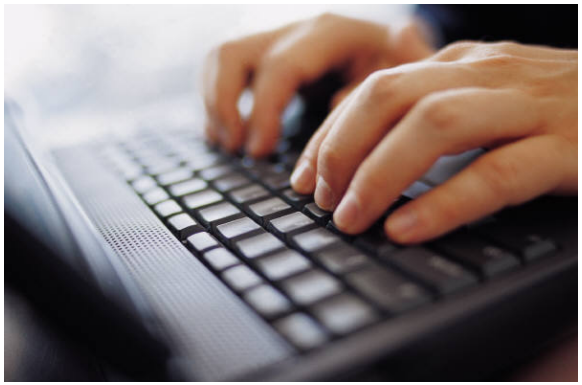
www.example.com

```
<head>
<script src="http://
otherdomain.com/
library.js"></script>
</head>
```

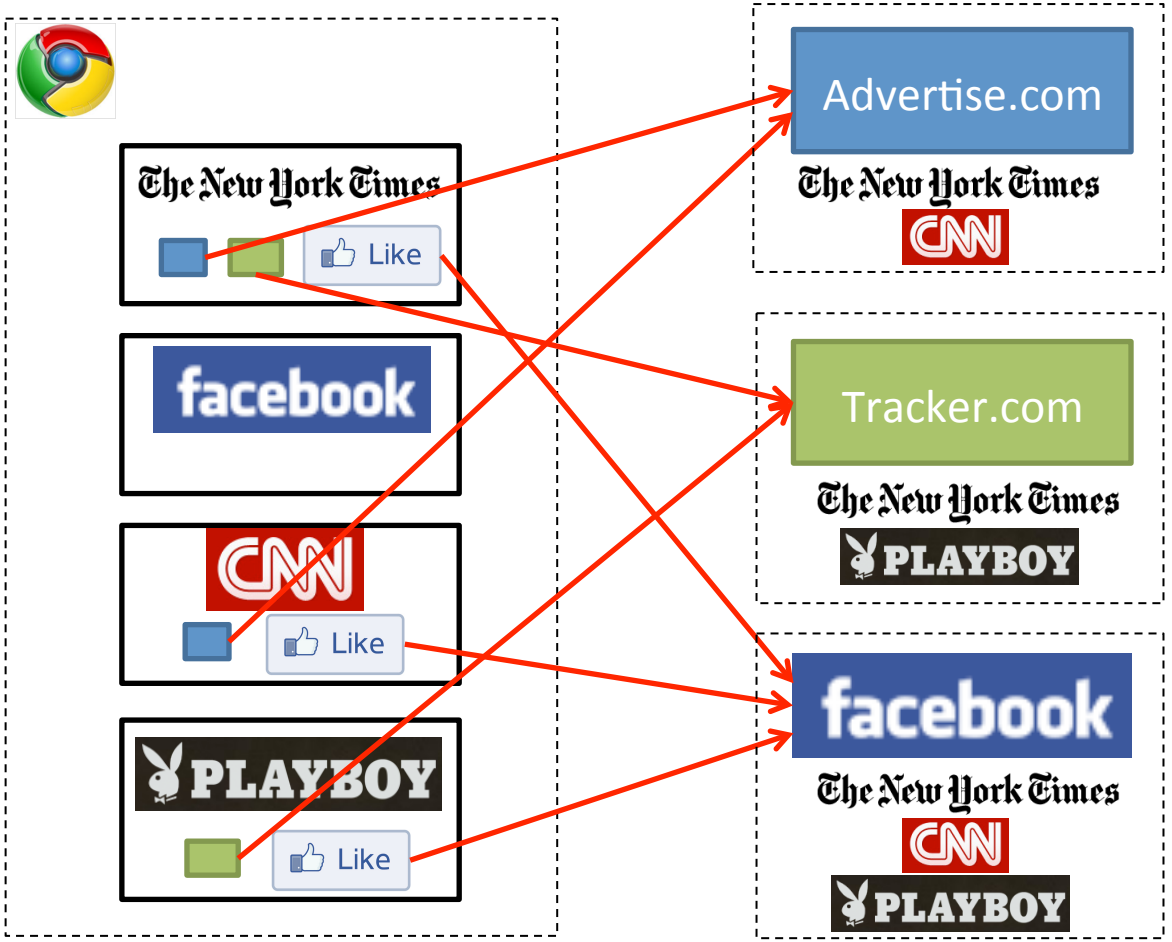
The code from http://otherdomain.com **can** access HTML elements and cookies on www.example.com.

- If code in the script sets a cookie, under what origin will it be set?

Third-Party Web Tracking



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



(Hypothetical tracking relationships only.)

Tracking is Complicated

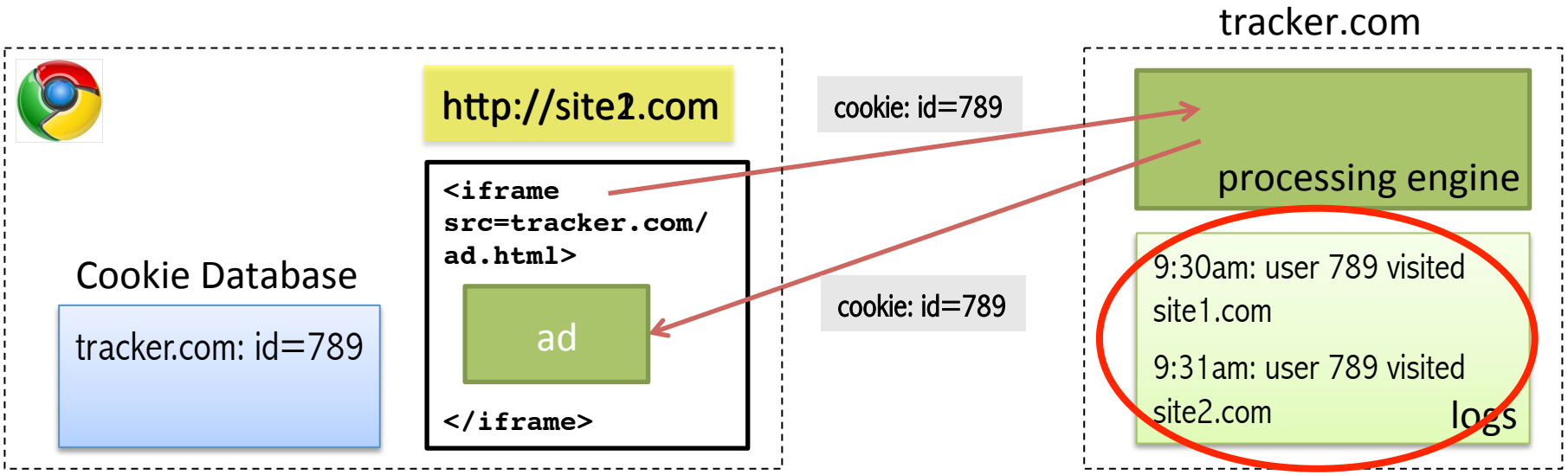
- Much discussion of tracking, but limited understanding of how it actually works.
- Our goals:
  - Understand the tracking ecosystem.
    - How is tracking actually done in the wild?
    - What kinds of browsing profiles do trackers compile?
    - How effective are defenses available to users?
  - Address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

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

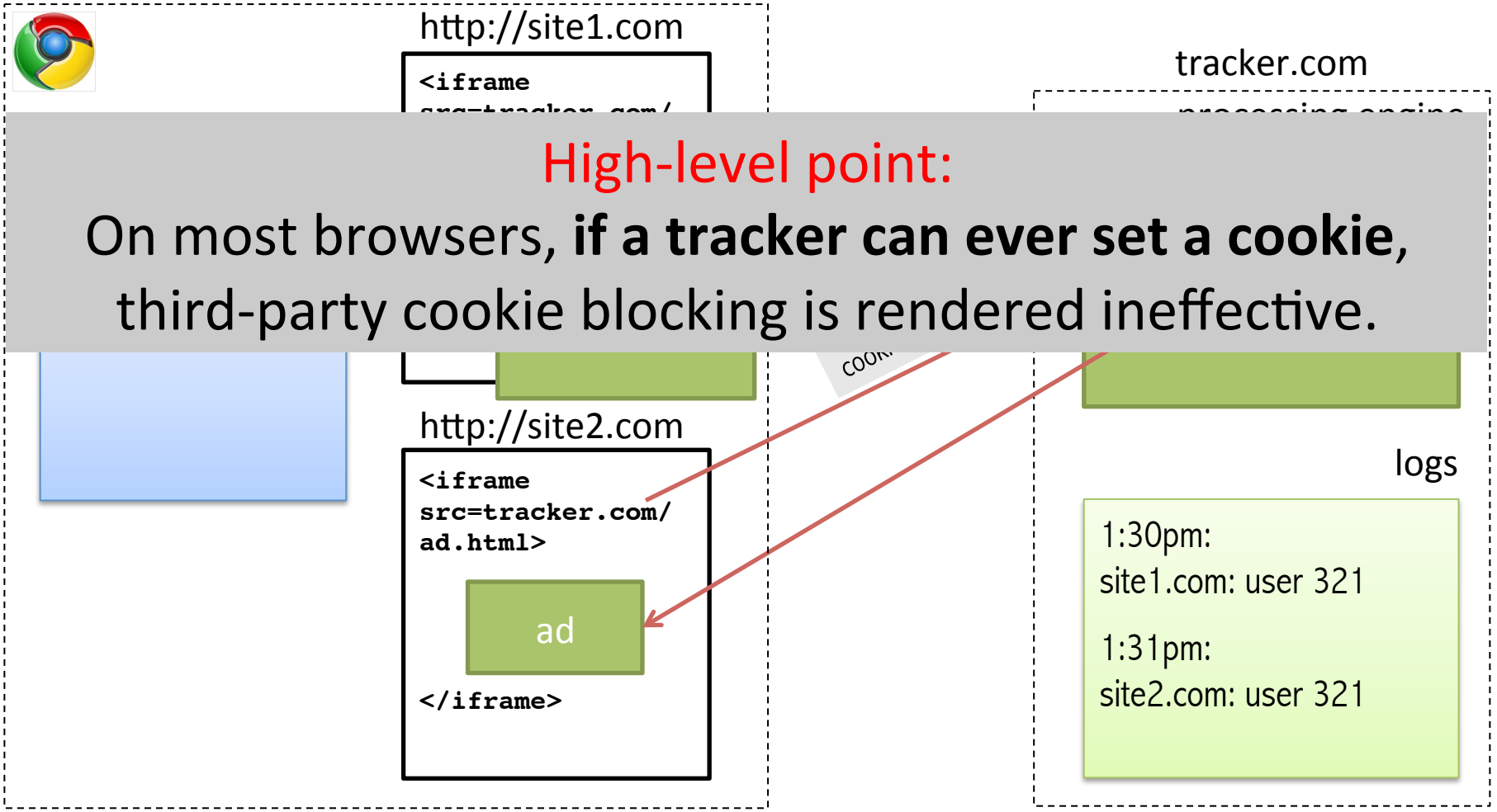
Evolution: Third-party cookie blocking

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

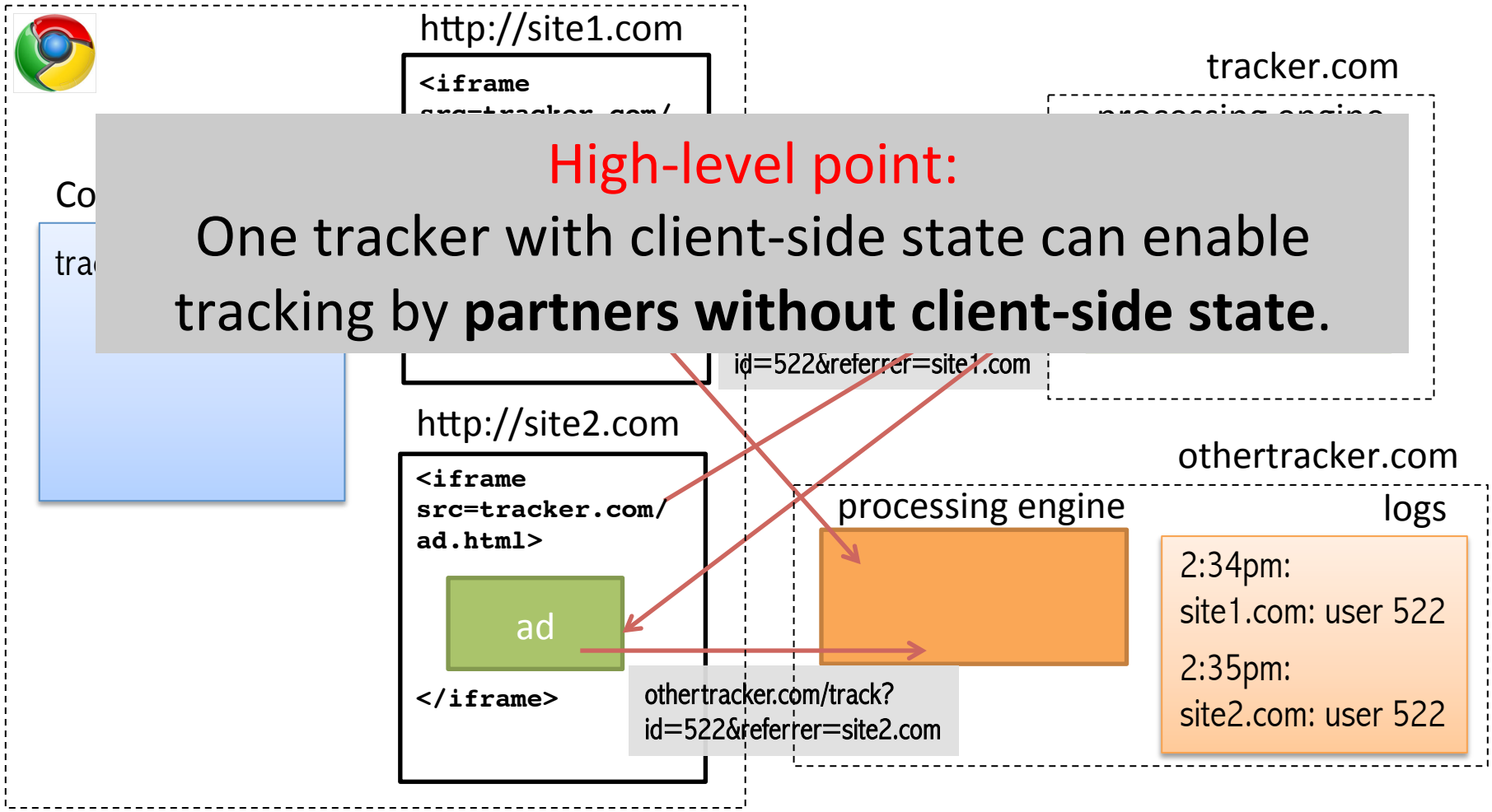
Name	Scope	User Visits Directly?	Overview
N/A	Within-Site	Yes	Site does its own on-site analytics.
Analytics	Within-Site	No	Site uses third-party analytics engine (e.g., Google Analytics).
Vanilla	Cross-Site	No	Site embeds third-party tracker that uses third-party storage (e.g., Doubleclick).
Forced	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

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

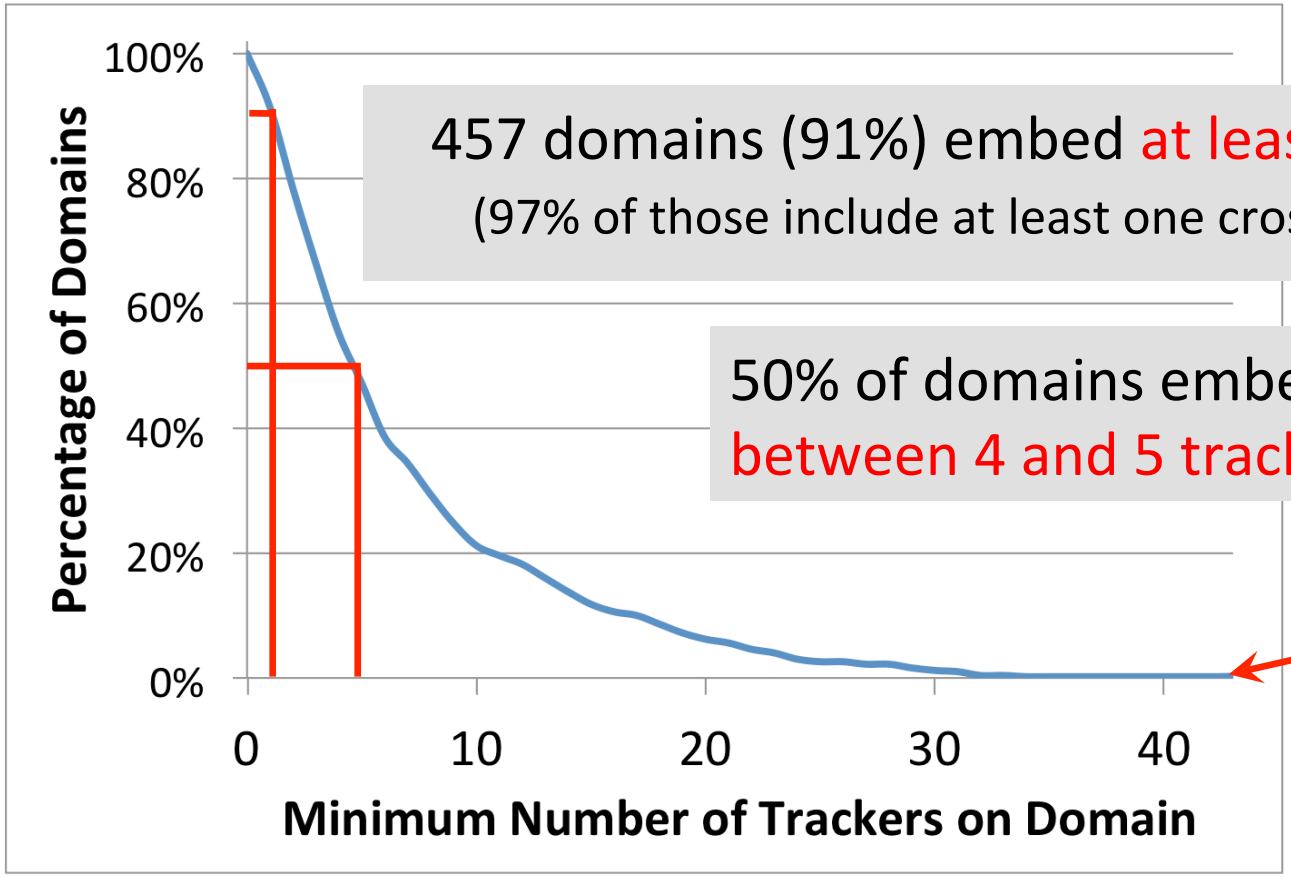
Anonymous

Measurement Study

- Tool: TrackingTracker Firefox add-on that crawls the web and automatically categorizes trackers.
- 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)

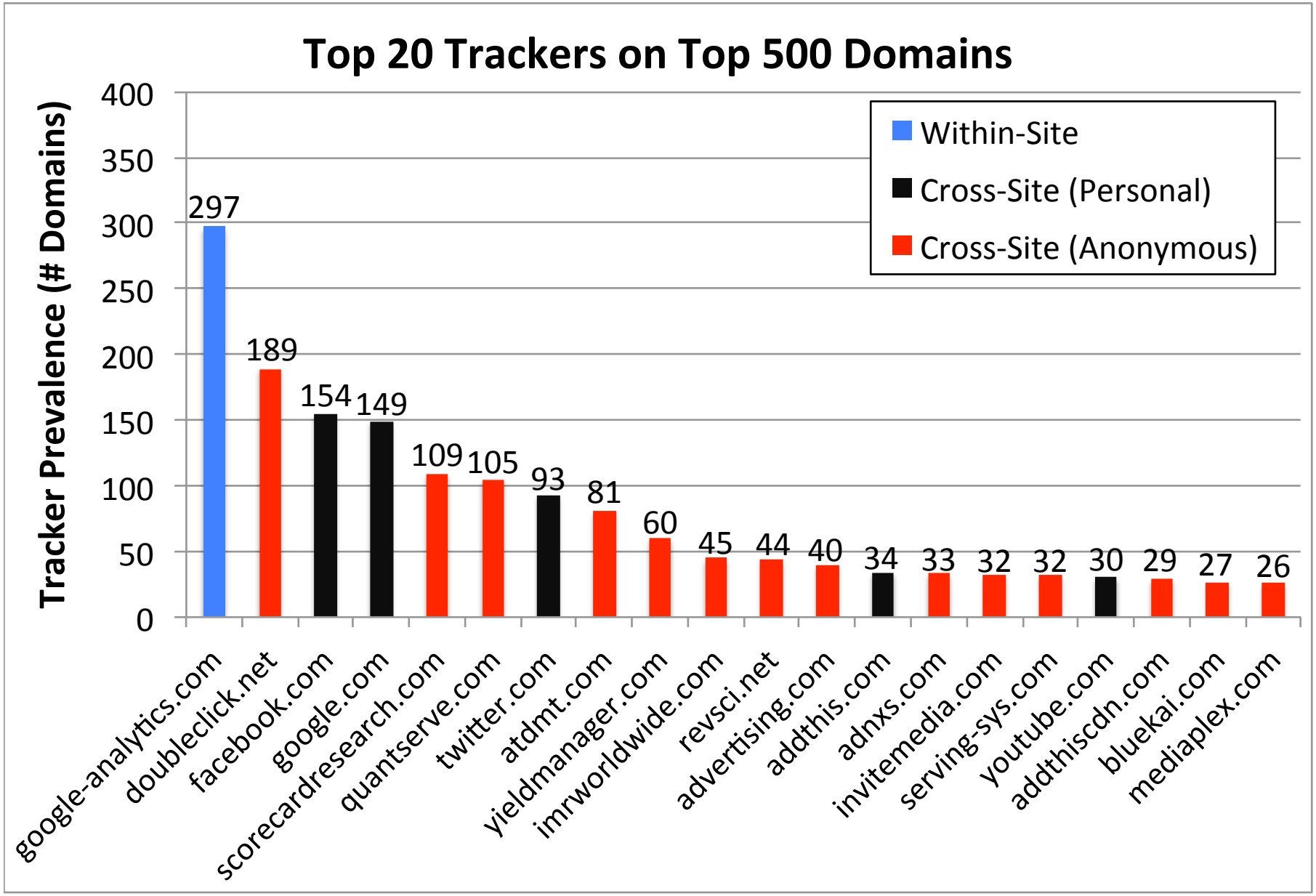
- 524 unique trackers on 500 domains

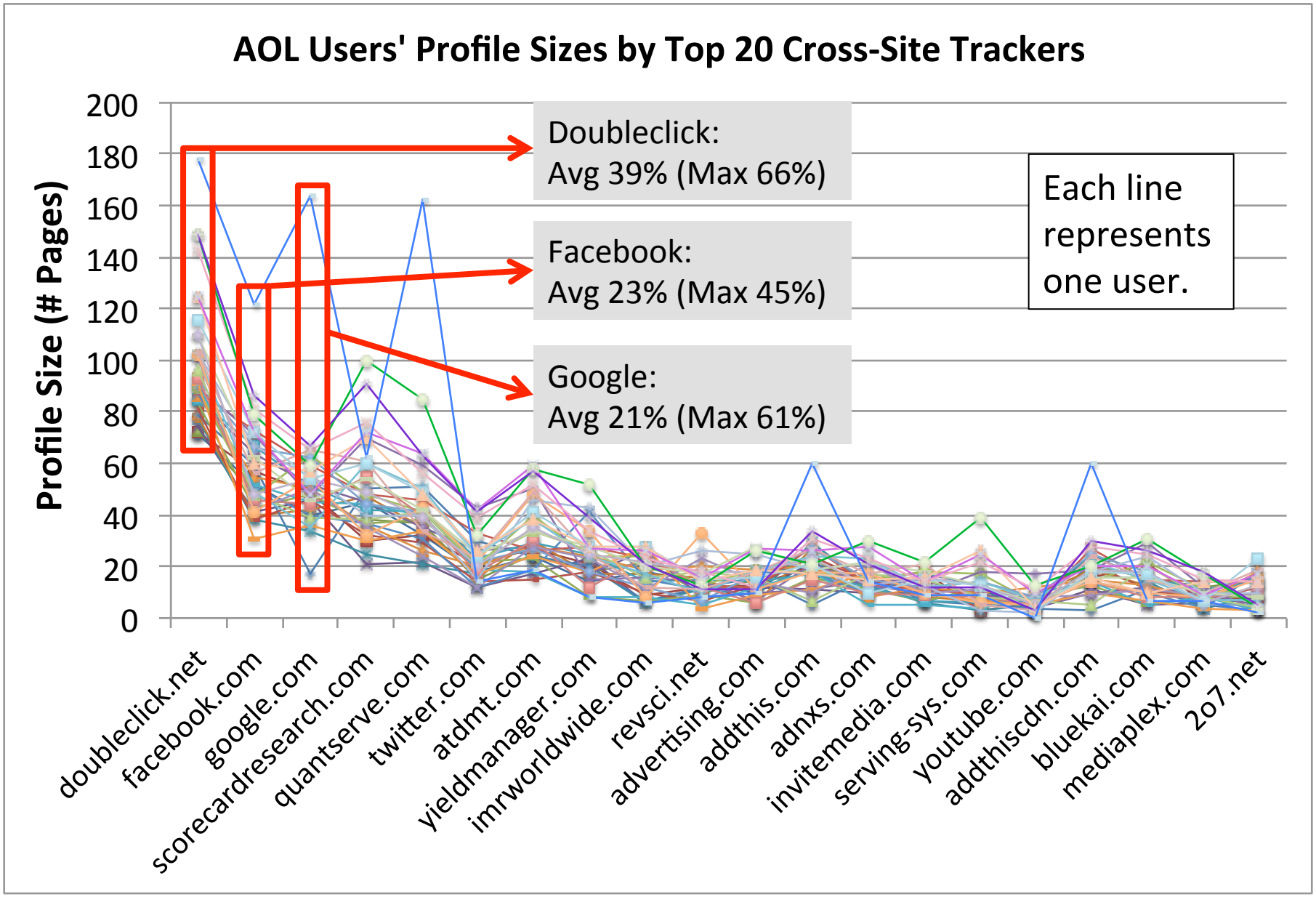


457 domains (91%) embed at least one tracker.
(97% of those include at least one cross-site tracker.)

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

Building New Systems

1. **ShareMeNot:** defense vs. personal tracking
   - Still allows social media widgets to be used.

http://sharemenot.cs.washington.edu

2. **TrackingObserver:** a platform for web tracking detection, measurement, prevention
   - Dynamic detection improves on state-of-the-art blacklist methods.

www.franziroesner.com

franzi@cs.washington.edu