AN INTRODUCTION TO WEB TRACKING

Fabrizio d'Amore

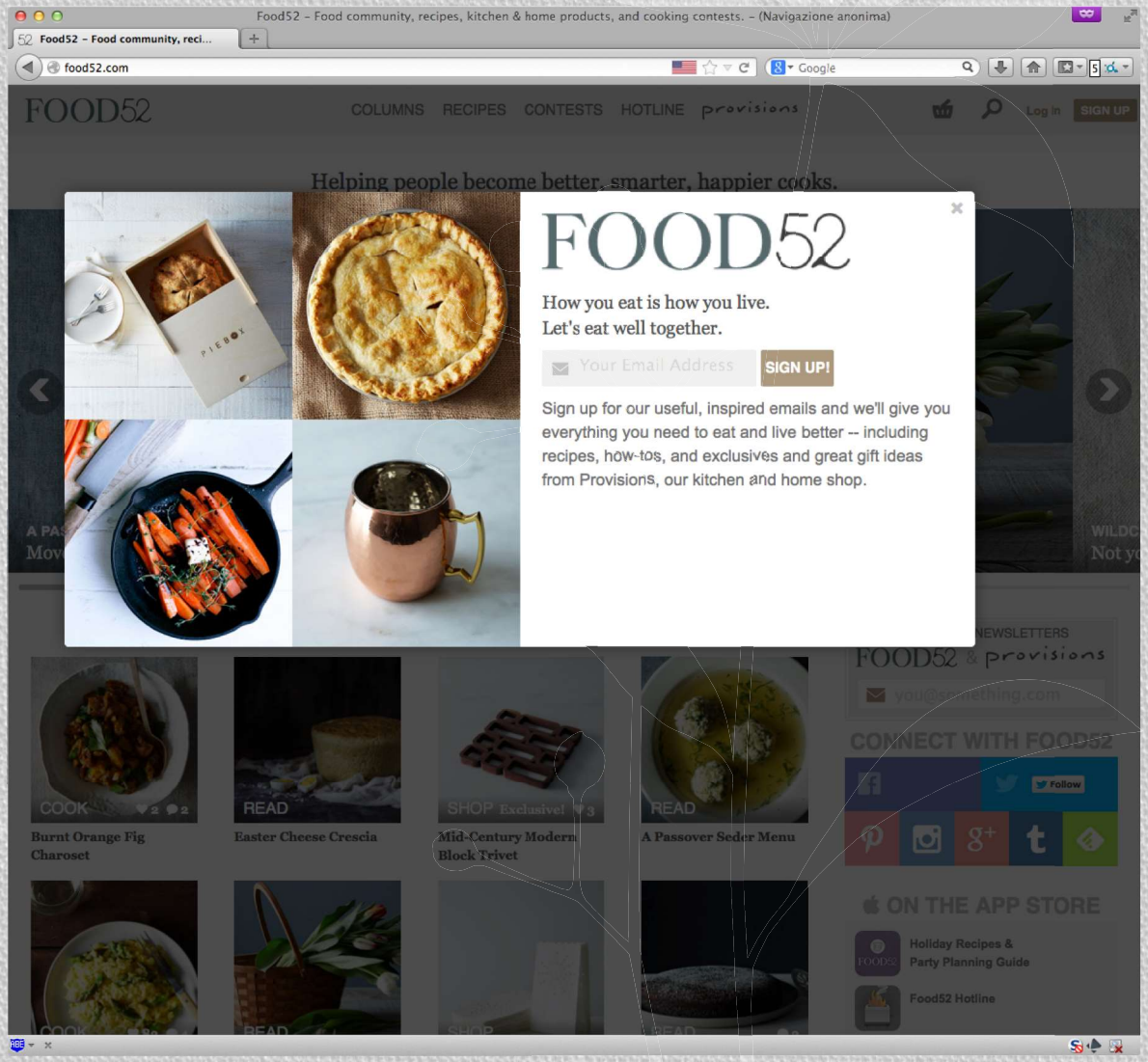
damore@dis.uniroma1.it

what Web tracking is

- collecting user's **browsing history** from a Web site, that can be
  - first party
    the Web site the user intended to visit
  - third party
    an unaffiliated/unknown/unexpected Web site the user did not intend to visit or was unaware of
- other information may be collected
  - names, username, dates, and other **personal data**
- data may be made object of **correlation** and/or **mining**
  - *data correlation* = determine possible statistical relationship between two sets of data
  - *data mining* = discovering patterns in (large) data sets, aiming at extracting information and transforming it into an understandable structure for further use

multiple contents

- due to HTML5 facilities many Web sites can delegate other (unknown or unexpected) sites for some contents
  - delegated sites can do the same
- multiple delegation and loss of control by 1st party
  - external contents may be invisible
  - example: test http://food52.com by NoScript
  - 3rd party sites may be able to access user info



multiple-parts contents

Figure 1. Third-party advertising, social, and video content on the New York Times website. Analytics content is not visible.

CSP

- Content Security Policy (CSP) is a new (2012) emerging standard for making 1° party able to control contents that are delegated to 3° parties
  - W3C candidate Recommendation (September 2016 https://www.w3.org/TR/CSP3/) CSP Level 3
  - originally developed by Mozilla
- CSP provides a standard HTTP header that allows website owners to declare approved sources of content that browsers should be allowed to load on that page
  - covered types are JavaScript, CSS, HTML frames, fonts, images and embeddable objects (Java applets, ActiveX, audio/video files etc.)

CSP implementation

- header names (partial list)
  - **`Content-Security-Policy`**. Standard header name proposed by the W3C document.
    - Google Chrome: version >= 25. Firefox: version >= 23
  - **`X-WebKit-CSP`**. Experimental header introduced into Google Chrome and other WebKit-based browsers (Safari) in 2011
  - **`X-Content-Security-Policy`**. Experimental header introduced in Gecko 2 based browsers (Firefox 4 to Firefox 22, Thunderbird 3.3, SeaMonkey 2.1)
- **`Content-Security-Policy`** header declares a list of white-listed sources for site contents: compliant browsers enforce the policy
- anytime a requested resource or script execution violates the policy, the browser will fire a POST request to the value specified in **`report-uri`** containing details of the violation
  - CSP reports are standard JSON structures and can be captured either by application's own API or public CSP report receivers

policy directives

- **`script-src`**: white-listed sources for scripts
- **`connect-src`**: limits the origins to which you can connect (via XMLHttpRequest (AJAX), WebSockets, and EventSource)
- **`font-src`**: specifies the origins that can serve web fonts
  - Google's Web Fonts could be enabled via **`font-src`**
- **`frame-src`**: lists the origins that can be embedded as frames
  - **`frame-src https://youtube.com`** would enable embedding YouTube videos, but no other origins
- **`img-src`**: defines the origins from which images can be loaded
- **`media-src`**: restricts the origins allowed to deliver video and audio
- **`object-src`**: allows control over Flash, applets and other plugins
- **`style-src`**: is `script-src`'s counterpart for stylesheets

1° vs. 3° party tracking

- same goals
- 3° party tracking technologies are a superset of 1° party ones
- 3° party tracking is more subtle and is often perceived as an abuse
- 3° party tracker is in some cases 1° party for same user (e.g., social networks)

WE DON'T LOSE GENERALITY BY CONSIDERING 3° PARTY TRACKING

leaked data

- location, interests, purchases, employment status, sexual orientation, financial challenges, medical conditions, and more
- examining individual page loads is often adequate to draw many conclusions about a user; analyzing patterns of activity allows yet more inferences
- when 1st party embeds 3rd party content 3rd party Web site is made aware of 1st party page URL
  - HTTP `referrer`
  - if 3rd party executes some script it can learn the page title from `document.title`
- in 2011 Epic Marketplace (advertising network) had publicly exposed its interest segment data, offering a rare glimpse of what third-party trackers seek to learn about
  - user segments included menopause, getting pregnant, repairing bad credit, and debt relief
- many examples in the recent literature

identifying browsing history

- 3rd party is also a 1st party
  - e.g., social networks
- 1st party sells the user's identity
- 1st party unintentionally provides identity
  - e.g., data in URL or in the document title
    - user profile not in e.g. http://example.com/self/ but in e.g. http://example.com/userid
    - user first name in title ("Welcome Fabrizio")
    - see Mayer and Mitchell tables [2012] (next slide)
- 3rd party uses a security exploit
  - e.g., a cross-site security vulnerability on 1st party website to learn user's identity
- re-identification: 3rd party could match pseudonymous browsing histories against identified datasets to re-identify them
  - e.g., compare browsing activity to the times and locations of links publicly shared by Twitter users

unintentional identity info leakage

Table I
THIRD PARTIES RECEIVING USERNAME AND ID ON 185 POPULAR SITES.

Third-Party PS+1	Websites Leaking Username or ID
scorecardresearch.com	81 (44%)
google-analytics.com	78 (42%)
quantserve.com	63 (34%)
doubleclick.net	62 (34%)
facebook.com	45 (24%)

Table II
POPULAR WEBSITES LEAKING USERNAME AND ID.

First-Party PS+1	Third-Party PS+1s Receiving Username or ID
rottentomatoes.com	83
cafemom.com	59
lyricsmode.com	54
ivillage.com	53
livejournal.com	53

understanding harms

- actor: who causes harm to a comsumer (assume M different actors)
  - actor might be an authorized employee, malicious employee, competitor, acquirer, hacker, or government agency
- means of access: enabling the actor to use tracking data (N means)
  - data might be voluntarily transferred, sold, stolen, misplaced, or accidentally distributed
- action that harms the consumer (P actions)
  - publication, a less favorable offer, denial of a benefit, or termination of employment
- particular harm that is inflicted (Q types of harm)
  - might be physical, psychological, or economic

HIGH NUMBER OF COMBINATIONS: MxNxPxQ

users do not want to be tracked

- a 2009 representative U.S. phone survey by Turow et al. found that 87% of respondents would not want advertising based on tracking
- in an unrepresentative 2010 survey of Amazon Mechanical Turk users by McDonald and Cranor, only 45% of respondents wanted to be shown any ads that had been tailored to their interests
- a December 2010 USA Today/Gallup poll reported 67% of respondents thought behavioral targeting should be outright illegal
- in a mid-2011 representative U.S. online survey by TRUSTe and Harris Interactive, 85% of respondents said they would not consent to tracking for ad targeting, and 78% said they would not consent to tracking for website analytics
- a 2012 representative telephone survey by Pew Research found that 68% of respondents were "not okay" with behavioral advertising

STILL MUCH TO UNDERSTAND:

DISAGGREGATING ON TRACKED DATA AND PAYING THE USERS

several policies can be adopted

- stakeholders: EU / USA policymakers, academics/researchers, browsers vendors, 3° party sites, advertisers, etc.
- all agreeing on the importance of the topic, disagreeing on
  - **what should consumers be able to control?**
    - policymakers and advocates believe consumers should have control over the collection of web tracking information
    - online advertising trade groups have argued that control should only extend to specific uses of data
  - **what should the default be?**
    - EU policymakers believe no tracking should be the default
    - advertising trade groups have argued tracking should be the default
  - **who should design the choice mechanism?**
    - advertising trade groups would like to control choice mechanism design
    - policymakers and advocates believe the browser vendors should retain design responsibility

tracking technologies

- stateless vs. stateful
- active vs. passive
  - can be detected?
- cookies based vs. non-cookies

cookies

- stateful
- active
- cookie-based

based on persistent 3° party cookies (already discussed)

supercookies

- stateful, active, cookie and non-cookie based
- not handled by browsers
- websites can encode a globally unique pseudonymous device identifier into any stateful web technology so long as it persists at least $\log_2 n$ bits, where n is the number of Internet-connected devices (presently roughly 5 billion, requiring 33 bits).
- Table provides a list of commonly deployed stateful web technologies and notes which have been observed in use for third-party web tracking.
  - The evercookie library provides a reference implementation for many of these tracking techniques

(a) "Supercookies"

HTTP authentication† [84]
HTTP caching ("cache cookies")
cache control
ETags* ("ETag cookies") [85]
Last-Modified [85] (e.g. [86])
cache content
resource (e.g. JavaScript, HTML, CSS, or media)*
status code
redirect location (e.g. [87])
hits and misses (e.g. [88])
TLS/SSL session ID [89]
browsing history††
userData storage (Internet Explorer only)*
HTML5 storage (session, local, and global)*
HTML5 protocol handlers†
HTML5 content handlers†
W3C geolocation API permission†
`window.name` property* (session only)
HTTP strict transport security [90]
plug-in storage* (e.g. Flash local shared objects, or "Flash cookies")
DNS cache

∗ Observed in use by a third-party website.
† User intervention required.
†† Largely inaccessible in newer browsers, but see [88], [91].

supercookies examples

- in mid-2011 KISSmetrics (third-party analytics service) was using cookies, Flash cookies, ETag cookies, cache cookies, userData, and HTML5 local storage; the non-cookie tracking technologies were used to recreate a cookie if deleted

- Microsoft was using an ETag cookie and a cache cookie in connection with its script for syncing an advertising identifier across web properties [2011]

Flash local shared objects

- they are supercookie-like
- from Wikipedia: "*A Local Shared Object (LSO) is a collection of cookie-like data stored as a file on a user's computer. LSOs are used by all versions of Adobe Flash Player and Version 6 and above of Macromedia's now-obsolete Flash MX Player*"
- LSOs can be blocked by defining, site by site, the behavior of Flash
  - no browser setting
- there is a Global Storage Setting Panel available at http://www.macromedia.com/support/documentation/en/flashplayer/help/settings_manager03.html
  - unsatisfactory solution that might make it failing legitimate sites

fingerprinting

- stateless, active or passive, not cookie based
- websites may be able to learn properties about browsers that, taken together, form a unique or nearly unique identifier
  - some properties require active discovery through a script or plug-in; other properties can be passively learned from network traffic
- Passive fingerprinting is particularly problematic since it cannot be detected with web measurement. Further research is needed to understand how effective passive fingerprinting is and what steps websites can take to scrub passive fingerprinting data from their logs. A recent study of Hotmail and Bing users by Yen et al. [2012] suggests passive fingerprinting may be sufficient to track many stationary browsers
- check the Panopticlick project page from Eff: https://panopticlick.eff.org/

active vs. passive fingerprinting

(b) Active "Fingerprinting"
operating system
CPU type
user agent
time zone
clock skew
display settings
installed fonts
installed plugins
enabled plugins
supported MIME types
cookies enabled
third-party cookies enabled

(c) Passive "Fingerprinting"
IP address
operating system
user agent
language
HTTP accept headers

user choice: Opt-Out Cookies

- several problems with this approach
  - it requires manual updating: to opt out of new third parties, a user has to install new cookies
  - cookies expire, so a user has to periodically renew opt-out cookies
  - users may clear their cookies, inadvertently removing their opt-out preferences
  - opt-out cookies are fragile; it is easy for a third party to improperly set or delete an opt-out cookie
  - opt-out cookies scale poorly; each third-party PS+1 requires a network roundtrip, resulting in a sluggish user experience when changing many preferences
- browser extensions for persisting opt-out cookies, such as TACO or Google Keep My Opt Outs, largely mitigate these issues at the cost of usability

user choice: AdChoices Icon

- online advertising companies have begun to insert an “AdChoices” icon (13x13px) and text (10pt) into display ads to increase user awareness of behavioral targeting and existing self-regulatory choice mechanisms
- clicking the icon provides additional information about how the ad was targeted and, in many cases, a link to landing page where the user can set opt-out cookies

Figure 2. Evolution of the AdChoices icon.

(a) Proposed icon and text [104] (actual size at 115 DPI).

(b) Implemented icon and text [105] (actual size at 115 DPI).

user choice: blocking and DNT

- blocking consists of adopting browser extension that use a block list, either available as a subscription for a browser extension or wrapped in a configurable browser extension
  - examples: AdBlock Plus, AdBlock Edge, Ghostery
- Do Not Track (DNT) uses a combination of technology and policy to provide consumer choice over web tracking
  - it is simply an HTTP header, DNT, that signals a user's preference about web tracking
  - associated technologies have been proposed that would allow a website to request exceptions and signal its own tracking status. Firefox, Internet Explorer, Safari, and Opera presently support a Do Not Track opt-out preference (sending the DNT: 1 header). Google has pledged to add the feature to Chrome

readings assignments

- Jonathan R. Mayer and John C. Mitchell. *Third-Party Web Tracking: Policy and Technology*. https://cyberlaw.stanford.edu/files/publication/files/trackingsurvey12.pdf
  - *these slides come from this survey*
- Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, Claudia Diaz. *The Web Never Forgets: Persistent Tracking Mechanisms in the Wild*. Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (https://www.dropbox.com/s/u2wl9686h9n0fp5/p674-acar.pdf?dl=0)
- Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, Giovanni Vigna. *Cookieless Monster: Exploring the Ecosystem of Web-based Device Fingerprinting*. https://lirias.kuleuven.be/bitstream/123456789/393661/1/