Announcements

• Join CS241  https://courseware.stanford.edu
• Project 1 out tonight
• Lab v0.1 out tomorrow morning
  – VM (1.7 Go)
  – Used to test project as well
• Lab section Friday 4:15-5:05
  – Introduction (all of them)
  – Browser security : 1 - 4
User Privacy on the Web

Arvind Narayanan
What is privacy?

How is it different from security?

Why does it matter?
CS has a narrow view of privacy

Privacy = confidentiality?

Dummy cameras

“Oh my God, my TiVo thinks I’m gay”

Everyone-but-X access control

Users come with their own mental models
Web Browsers and Privacy
Tracking

Trackers from 10 different domains!
Source: EFF
Sticky tracking

Violation of same-origin policy
(blocking 3rd party cookies no longer works)

ad.hi5.com = ad.yieldmanager.com

Flash cookies

Browser fingerprinting
# Partial List of Trackers

<table>
<thead>
<tr>
<th>24/7 Real Media</th>
<th>33Across</th>
<th>Acerno</th>
<th>Acxiom Relevance-X</th>
<th>AdAdvisor</th>
<th>AdBrite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adfly</td>
<td>AdInterax (Yahoo!)</td>
<td>AdJuggler</td>
<td>AdShuffle</td>
<td>ADTECH (AOL)</td>
<td>Advertising.com (AOL)</td>
</tr>
<tr>
<td>Aggregate Knowledge</td>
<td>Akamai</td>
<td>AlmondNet</td>
<td>Atlas (Microsoft)</td>
<td>AudienceScience</td>
<td>Bizo</td>
</tr>
<tr>
<td>Blue Kai</td>
<td>BlueLithium (Yahoo!)</td>
<td>Bluestreak</td>
<td>BrightRoll</td>
<td>BTBuckets</td>
<td>Burst Media</td>
</tr>
<tr>
<td>Casale Media</td>
<td>Chitika</td>
<td>ChoiceStream</td>
<td>ClickTale</td>
<td>Collective Media</td>
<td>comScore VoiceFive</td>
</tr>
<tr>
<td>Coremetrics</td>
<td>Cossette</td>
<td>Criteo</td>
<td>Effective Measure</td>
<td>Eloqua</td>
<td>Eyeblast</td>
</tr>
<tr>
<td>eXelate</td>
<td>EyeWonder</td>
<td>e-planning</td>
<td>Facilitate Digital</td>
<td>FetchBack</td>
<td>FlashTalking</td>
</tr>
<tr>
<td>Fox Audience Network</td>
<td>FreeWheel</td>
<td>Google</td>
<td>Hurra</td>
<td>interCLICK</td>
<td>Lotame</td>
</tr>
<tr>
<td>Navegg</td>
<td>NextAction</td>
<td>NexTag</td>
<td>Mediaplex (ValueClick Media)</td>
<td>Media 6 Degrees</td>
<td>Media Math</td>
</tr>
<tr>
<td>Microsoft</td>
<td>MindSet Media</td>
<td>Nielsen Online</td>
<td>nugg.ad</td>
<td>Omninure</td>
<td>OpenX</td>
</tr>
<tr>
<td>Outbrain</td>
<td>PointRoll</td>
<td>PrecisionClick</td>
<td>Pulse 360</td>
<td>Quantcast</td>
<td>Quilgo (AOL)</td>
</tr>
<tr>
<td>richrelevance</td>
<td>Right Media (Yahoo!)</td>
<td>Rocket Fuel</td>
<td>Safecount *</td>
<td>ScanScout</td>
<td>Smart Adserver</td>
</tr>
<tr>
<td>Snoobi</td>
<td>Specific Media</td>
<td>TACODA (AOL)</td>
<td>Tattoo Media</td>
<td>Tealium</td>
<td>TradeDoubler</td>
</tr>
<tr>
<td>Traffic Marketplace</td>
<td>Tribal Fusion / Exponential</td>
<td>TruEffect</td>
<td>Tumri</td>
<td>Turn</td>
<td>Undertone Networks / Zedo</td>
</tr>
<tr>
<td>ValueClick Media</td>
<td>Vizu</td>
<td>Weborama</td>
<td>WebTrends</td>
<td>Yahoo</td>
<td>[x+1]</td>
</tr>
</tbody>
</table>
Everything has a Fingerprint
Most entropy comes from version numbers of fonts, plugins etc.

User-agent string alone: 10 bits of entropy

84% of fingerprints unique

With Flash or Java, 94% unique
Panopticlick example

Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US; rv:1.9.1.7) Gecko/20091221 Firefox/3.5.7

(Already unique among ~1M browsers tested 😞)
Panopticclick example

Plugin 0: Adobe Acrobat; Adobe Acrobat Plug-In Version 7.00 for Netscape; nppdf32.dll; (Acrobat Portable Document Format; application/pdf; pdf) (Acrobat Forms Data Format; application/vnd.fdf; fdf) (XML Version of Acrobat Forms Data Format; application/vnd.adobe.xfdf; xfdf) (Acrobat XML Data Package; application/vnd.adobe.xdp+xml; xdp) (Adobe FormFlow99 Data File; application/vnd.adobe.xfd+xml; xfd). Plugin 1: Adobe Acrobat; Adobe PDF Plug-In For Firefox and Netscape; nppdf32.dll; (Acrobat Portable Document Format; application/pdf; pdf) (Acrobat PDF in XML Format; application/vnd.adobe.pdfxml; pdfxml) (Adobe PDF in XML Format; application/vnd.adobe.x-mars; mars) (Acrobat Forms Data Format; application/vnd.adobe.xfdf; xfdf) (Acrobat XML Data Package; application/vnd.adobe.xdp+xml; xdp) (Adobe FormFlow99 Data File; application/vnd.adobe.xfd+xml; xfd). Plugin 2: Google Update; Google Update; npGoogleOneClick8.dll; (application/x-vnd.google.oneclickctrl.8; ). Plugin 3: Microsoft® Windows Media Player Firefox Plug-in; npmswmp; np-mswmp.dll; (np-mswmp; application/x-ms-wmp; *) (; application/asx; *) (; video/x-ms-asf-plugin; *) (; application/x-mplayer2; *) (; video/x-ms-asf; asf,asx,* ) (; video/x-ms-wm; wm,* ) (; audio/x-ms-wma; wma,* ) (; audio/x-ms-wax; wax,* ) (; video/x-ms-wmv; wmv,* ) (; video/x-ms-wvx; wvx,* ). Plugin 4: Move Media Player; npmnqmp 07103010; npmnqmp07103010.dll; (npmnqmp; application/x-vnd.moveplayer.qm; qmx,qpl) (npmnqmp; application/x-vnd.moveplay2.qm; ) (npmnqmp; application/x-vnd.movenetworks.qm; ). Plugin 5: Mozilla Default Plug-in; Default Plug-in; npnl32.dll; (Mozilla Default Plug-in; *; *). Plugin 6: Shockwave Flash; Shockwave Flash 10.0 r32; NPSWF32.dll; (Adobe Flash movie; application/x-shockwave-flash; swf) (FutureSplash movie; application/futuresplash; spl). Plugin 7: Windows Genuine Advantage; 1.7.0059.0; npLegitCheckPlugin.dll; (npLegitCheckPlugin; application/WGA-plugin; *).
Panopticlick example

Identity, Pseudonymity and Anonymity

Pseudonymity = Anonymity + Linking
History stealing and privacy

CSS :visited property
  – getComputedStyle()
  – cache timing
  – server hit

Possible malicious uses:
  • History as cookies
  • Better phishing
Identity stealing

• All social networking sites have “group” functionality: users can join groups.
• Users typically join multiple groups – some of which are public.
• Group affiliations act as fingerprint
• Predictable group-specific URLs exist

http://www.facebook.com/group.php?gid=[groupID]&v=info&ref=nf+
https://www.xing.com/net/[groupID]/forums+

Wondracek, Holz, Kirda, Kruegel
Users Fingerprinted by top $N$ Groups
History Stealing Bandwidth
Identity Stealing 1.0

Pre-social web
Site-specific identities

Social network affiliations
Footprints in the social network

Bonneau, Boneh, Narayanan, Rydstedt (in progress)
Identity Stealing 2.0

Social web: federated identity
Footprints on individual pages

Social web: sharing
Footprints in the social network;
identities on a.com and b.com irrelevant

Bottom line: server side fixes not feasible
Fixing history stealing

There are two sites that can observe a click

Link from site S $\rightarrow$ url U will appear visited only if 
$U$ was visited from a page in $S$

(Some sites whitelisted)

Jackson, Bortz, Boneh, Mitchell
Fixing history stealing (Mozilla)

Hodgepodge of heuristics
Private Browsing

• Local attacker
  – Should not be able to figure out which sites were visited

• Web attacker
  – Public and private mode unlinkable
  – Different private sessions unlinkable
  – Private browsing undetectable

• All browsers fail in one way or another

• Usage experiment

Aggarwal, Burzstein, Jackson, Boneh
Categories of changes

• Initiated by website, no user interaction
  – e.g. adding visited pages to history
• Initiated by website, interaction required
  – e.g. adding a password
• Initiated by user (e.g. bookmark)
• Generic changes such as installing a patch
Is state set in private mode accessible in public mode

<table>
<thead>
<tr>
<th></th>
<th>FF</th>
<th>Safari</th>
<th>Chrome</th>
<th>IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Cookies</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>HTML5 Local storage</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Bookmarks</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Password database</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Form autocompletion</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>User approved SSL self-signed cert</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Downloaded items list</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>n/a</td>
</tr>
<tr>
<td>Search box search terms</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Browser’s web cache</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Client certs</td>
<td>yes</td>
<td>n/a</td>
<td>n/a</td>
<td>yes</td>
</tr>
<tr>
<td>Custom protocol handlers</td>
<td>yes</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Examples of violations

• Custom protocol handlers

• SSL client certs

• Self-signed certs
Private Browsing – Rates of Use

The chart shows the rates of use for Safari, Firefox 3.5+, Google Chrome, and IE 8, categorized into gift sites (blue bars) and adult sites (pink bars). The highest rate is observed for Safari, followed by Firefox 3.5+, Google Chrome, and finally IE 8.
Plugins and Extensions

- Plugins on, extensions off by default
- Both plugins and extensions on
- Plugins on, most extensions functionality off

Firefox top 40 extensions: over ½ violate privacy

Chrome: better security model for extensions
fine-grained privileges and isolation
Facebook

Plan: “identity layer” for the web

Q1. How would it work?

Q2. What are the privacy risks?
Peeking through SSL

• SSL communications reveal length of packets

• AJAX-rich pages have lots and lots of interactions with the server

• These interactions are very specific to the internal state of the page

Chen, Wang, Wang, Zhang
Peeking through SSL

Vulnerabilities in an online tax application
Making public data “more public”

Fastest DeadPool Ever? IRSeek Shuts Down
by Michael Arrington on Dec 3, 2007

Apparently IRC participants don’t really like the idea of their conversations being indexed and searchable. Which is exactly what we said when we first wrote about IRSeek four days ago. The site has now shut down. A blog post explains why:

we have disabled the site...we were under the impression that users in public chat-rooms are aware that their conversations are, by definition, public. Since people
Making public data “more public”

Massive Scrape of Twitter’s Friend Graph

UPDATE:

We’ve posted several Twitter datasets on Infochimps. Take a look and build something cool!

UPDATE:

We’ve taken the data down for the moment, at Twitter’s request. STAY CALM. They want to support research on the twitter graph, but feel that since this is users’ data there should be terms of use in place. We’ve taken the data down while those terms are formulated. I pass along from @ev: “Thank you for your patience and cooperation.”
Making public data “more public”

Spokeo

Claim: Your personal information may be viewable through the Spokeo web site.

TRUE

Example: [Collected via e-mail, March 2010]

This is so scary to me. My address along with a picture of my home is showing on this site.

GO to this website: http://www.spokeo.com/ and type in your name. If you find ANYTHING with your info on it, go to "Privacy" at the bottom of the page and follow the instructions to remove your information. Some of the information they have listed may not be correct, but if your address, phone number and a picture of your home comes up, that’s cause for concern!

---

Holy Cow Family and Friends!

I just received a link for a crazy website. A friend emailed it to me saying that the website has all the personal info on her family and that maybe I should check it out to make sure I am not on
Data sifted from Facebook wiped after legal threats

12:36 31 March 2010 by Jim Giles

Legal threats from Facebook have led to the destruction of a social science dataset about to be released to researchers.

Lawyers from the social networking site contacted Pete Warden, an entrepreneur based in Boulder, Colorado, in February after he announced plans to release data he had collected from the public profiles of 210 million Facebook users.

Warden says that Facebook threatened legal action if he did not delete the data. He duly destroyed all the records, saying he did not have the funds to contest a lawsuit.
Making public data “more public”

Infochimps MySpace Data Sale Draws Heat

A move by Austin-based Infochimps to offer bulk data access to data from MySpace, is drawing widespread heat over user privacy issues, spurring MySpace to issue a statement today saying it is not selling its user data. Infochimps announced last week that it entered into an agreement with MySpace to sell datasets based on MySpace's real time updates. MySpace said today that it is not selling its data, and that Infochimps is simply providing pre-packaged versions of its data for some third-party developers, who are unable to process real-time data. MySpace said the data shared in the sets is already public, and that users can limit their stream privacy settings to prevent inclusion in the data. MySpace also said that the data has already been available--for free--to developers to use, in such services as Google's real time search.
Making public data “more public”

WARNING: Google Buzz Has A Huge Privacy Flaw (GOOG) — There is a huge privacy flaw in Google’s new Twitter/Facebook competitor, Google Buzz. — When you first go into Google Buzz, it automatically sets you up with followers and people to follow. — A Google spokesperson tells us ...

Discussion: Google Blogoscoped, CNET News, Twittercism, 1001 Noisy Cameras, Network World, eWeek, AccMan, Gizmodo Australia, Social Business, ReadWriteWeb, Between the Lines, Giles Bowkett, TechCrunch, Datamation, Google Mobile Blog, Collaborative Thinking, Techdirt, bit-tech.net and Simon Willison's Weblog
Making public data “more public”

- Search
- Aggregation
- Real-time
- Linkage
Thank you