Privacy Diffusion on the Web: A Longitudinal Perspective

Balachander Krishnamurthy
AT&T Labs–Research
http://www.research.att.com/~bala/papers

Craig Wills
Worcester Polytechnic Institute
http://www.cs.wpi.edu/~cew
Privacy

- Privacy worries stem from nature of the information disseminated, what data collectors *might* do with it (not just today)

- Goal is to allow standard network activity while preserving desired privacy

- Various daily interactions on the Web (commerce, email, search...): some of which require supply of private information

- Sites use many techniques to track users (1x1 pixel Web bugs, cookies)

- Aggregators track across sites (dclk, googlesyndication, tacoda)

- Measure of dissemination of user-related information across *unrelated* sites: *privacy footprint*
First-party vs. Thirdy-Party nodes

Examine connections between first-party visible (servers explicitly visited) and hidden third-party (visited as by-product) nodes

Visible Nodes
- www.accuweather.com
- www.nationalreview.com
- www.cnn.com
- www.americanexpress.com
- online.wsj.com
- www.amazon.com
- www.target.com

Hidden Nodes
- a248.e.akamai.net
- i.a.cnn.net
- m.2mdn.net
- m.doubleclick.net
- cnn.122.207.net
- americanexpress.122.207.net
- dowjones.122.207.net
- g-images.amazon.com

Third-party nodes may be CDNs, ad sites, and aggregators
1. Ad Networks: First-party sites (publishers) arrange with ad networks to place ads on their pages via images or javascript code.
   E.g., Google’s Adsense (googlesyndication.com, doubleclick.net), AOL (advertising.com, tacoda.net), Yahoo!(yieldmanager.net)

2. Analytics companies: measure traffic, characterize users by downloading a JavaScript file and send back information in a URL.
   E.g., google-analytics.com (urchin.js), 2o7.net (Omniture), atdmt.com (Microsoft/aquantive), quantserve.com (Quantcast)

3. CDNs: Serve images, rarely JavaScript. e.g., akamai.net, yimg.com

Privacy leaks to all of them.
Mechanics of data collection

- Visible nodes: Popular 1200 Web sites in dozen Alexa categories

- Extracted hidden nodes corresponding to each visible node via a Firefox extension that fetches objects and records request/response

- Examined cookies, JavaScript, identifying URLs (those with ? = &)

- Also narrowed examination to consumer and fiduciary sites: subset of sites that raise more privacy concerns.

Two visible nodes are *associated* if accessing them results in accessing the same hidden node.

Association can be due to several reasons:

1. **server**: Identical server name (www.google-analytics.com)

2. **domain**: Aggregated by merging hidden nodes with same 2nd-level domain names. E.g. timecom.112.2o7.net and msnbcom.112.2o7.net

3. **adns**: Aggregated by merging hidden nodes that share the same ADNS (authoritative DNS server). E.g. doubleclick.net and ebayobjects.com have the same ADNS.
- DNS for third-party servers may be provided by sites like ultradns.net

- CDNs are increasingly used to serve content for third party servers (e.g., JavaScript or images with cookies)

- We check ADNS of 3rd-party and 1st-party servers—if they differ and the ADNS server is not that of a known CDN or DNS service, we use the 3rd-party server as the domain

- e.g. pixel.quantserve.com’s ADNS is akamai, so domain is quantserve.com, but w88.go.com’s domain is omniture.com, based on its ADNS.
Privacy footprint: longitudinal study

- Footprint shows the number and diversity of 3rd-party sites visited as a result of a user visiting first party sites.

- We examine the penetration of the top 3rd-party domains that aggregate information about user’s movements on the Web.

- Multiple 3rd-parties may track users on a given first-party site and so this is examined as well.

- Finally, we examine the role of economic acquisitions of aggregator companies that buy others and increase their tracking ability.
Some domains showed up in later epochs (quantserve, google-analytics)

Top line shows the combined impact of the top-10 domains at each epoch – going from 40% to nearly 70%.
Manner of tracking

Initially just 3d-party cookies, but now 1st-party cookies and JavaScript: so we examined traces of requested objects, cookies and JavaScript downloaded.

Four categories of 3d-party domains:

1. Only set 3d-party cookies, no JS (dclk, atdmt, 2o7.net)

2. Use JS with state saved in 1st-party cookies (google-analytics: urchin.js examines 1st-party cookies, forces retrieval via an identifying URL to send information to 3d-party server)

3. Both 3d-party cookies and JS to set 1st-party cookies (quantserve)

4. 3d-party cookies and JS not used to set 1st-party cookies but serve ad URLs with tracking information (adbrite, adbureau)
## Acquisitions of Third-Party Domains By Families

<table>
<thead>
<tr>
<th>Family</th>
<th>Acquired</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOL</td>
<td>advertising.com</td>
<td>Jun’04</td>
</tr>
<tr>
<td></td>
<td>tacoda.net</td>
<td>Jul’07</td>
</tr>
<tr>
<td></td>
<td>adsonar.com</td>
<td>Dec’07</td>
</tr>
<tr>
<td>Doubleclick</td>
<td>falkag.net</td>
<td>Mar’06</td>
</tr>
<tr>
<td>Google</td>
<td>youtube.com</td>
<td>Oct’06</td>
</tr>
<tr>
<td></td>
<td>doubleclick.net</td>
<td>Mar’07</td>
</tr>
<tr>
<td></td>
<td>feedburner.com</td>
<td>Jun’07</td>
</tr>
<tr>
<td>Microsoft</td>
<td>aquantive.com (atdmt.com)</td>
<td>May’07</td>
</tr>
<tr>
<td>Omniture</td>
<td>offermatica.com</td>
<td>Sep’07</td>
</tr>
<tr>
<td></td>
<td>hitbox.com</td>
<td>Oct’07</td>
</tr>
<tr>
<td>Valueclick</td>
<td>mediaplex.com</td>
<td>Oct’01</td>
</tr>
<tr>
<td></td>
<td>fastclick.net</td>
<td>Sep’05</td>
</tr>
<tr>
<td>Yahoo</td>
<td>overture.com</td>
<td>Dec’03</td>
</tr>
<tr>
<td></td>
<td>yieldmanager.com</td>
<td>Apr’07</td>
</tr>
<tr>
<td></td>
<td>adrevolver.com</td>
<td>Oct’07</td>
</tr>
</tbody>
</table>
Sep’08 Google family reach: 60%—highest among all third parties by far.
Primarily 2o7.net domain and then acquisitions—reach of 28%
Reach of 22% in Sep’08, growth from buying Aquantive (atdmt.com).
Other families: Yahoo: 15%, AOL: 14% in Sep’08
Depth of Tracking has also increased

Users are being tracked by two or more third-party entities.

- In Oct’05, 24% of 1200 popular Web sites contained more than one of the top 3rd-party domains.
- In Sep’08 this figure had risen to 52% (34% with more than two).
- It is not enough just to block a single tracking entity.
Examined 127 consumer sites’ longitudinal privacy leakage.

E.g., apple, blockbuster, buy, ebay, expedia, gap, hilton, ikea, kayak, netflix, oldnavy, target, sears

Steadily increasing node associations:
Oct ’05 58%, Apr ’06 66%, Oct ’06 66%, Feb ’08 74%, Jul ’08 77%

Top aggregators:
doubleclick.net, 2o7.net, google-analytics.com, yieldmanager.com, atdmt.com, advertising.com, akamai.net, tacoda.net, specificclick.net, offermatica.com
Google family is largest starting in ’08 but Omniture appears to be strong. Top-10 domains account for nearly 80%.
Top-10 3d-Party families in Fiduciary sites over time

81 sites in 9 categories:
credit financial insurance medical mortgage shopping subscription travel utility

Top-10 domains account for over 60%.
3rd-party aggregators are using *1st-party* cookies to track users via 3rd-party JavaScript - nearly 60%. Can’t reject all 1st-party cookies.

3rd-party JavaScript served by 1st-party server: cannot auto block - over 30%.

17.5% have 3rd-party objects “hidden” in seemingly 1st-party servers (Omniture’s JS on abc.go.com: ident URL for w88.go.com, ADNS shows it is in Omniture)
Recent privacy issues

- Recent: IE 8.0’s proposed InPrivate Browsing and “line of sight” blocking

- Goes after specific .js files being downloaded when users visit different sites. A start but superficial/inadequate

- On average 41% of 3rd-party domains accessed are in the top-10 domain set and half of these set cookies. InPrivate Blocking extended to do a transitive closure of third party site accesses could reduce leakage.

- cuil.com’s simple privacy policy

- Search information now stored “only” for 9 months to please European regulators.

- Chrome: URL completion leaks *any* URLs to Google by default

- Specific Media (175M individual profiles)
New privacy concerns

- Notion of “Collateral privacy damage”
- Privacy of other users are violated as a result of data/access given by a user
- E.g., email communication leads to social graph formation
- Posted/shared personal information can be applied to relatives
Conclusion

- We have examined longitudinal leakage of privacy on the Web
- We have explored manners of aggregation and extent
- Economic acquisition has reduced number of players and increased individual aggregator’s visibility footprint
- We are examining leakage of PII next (ACM SIGCOMM WOSN, August, 2009)