Towards Privacy-Friendly Online Advertising

Julien Freudiger, Nevena Vratonjic, and Jean-Pierre Hubaux

May 2009, W2SP
Internet Economy

Online advertising is at center of Internet economy

– Immediate and personalized
– Enables **Behavioral** targeting

**Annual $ Revenue — 1997 through 2008**

Source: Interactive Advertising Bureau Internet Advertising revenue report, 2008
Benefits

• For users
  – Relevance of ads
  – Sponsored services

• For websites
  – Generate profit from ads
  – New business models
Privacy Concerns

• Track user activities online
  – Interests (visited websites, search terms)
  – Conversations (email)
  – Friends (social networks)
• Privacy footprint (Krishnamurthy and Wills)
  – 72% of web servers share at least one advertiser
  – 3 third-party domains contacted on average per accessed web site
Privacy/traceability Trade-off

Provide a way to control amount of information shared
Outline

1. Online Advertising
   – Privacy Implications
   – Existing Solutions

2. Proposed Solution
   – Privacy friendly Cookie management
   – User centric

3. Preliminary Evaluation
   – Firefox Extension
Online Advertising

Users $U$  Visible servers $S$  Hidden servers $D$

$u \rightarrow s_1$: www.nytimes.com
$s_1 \rightarrow u$: index.html
$u \rightarrow d_1$: ads.com, **TP-cookie**
$d_1 \rightarrow u$: ads

$u \rightarrow s_2$: www.google.com
$s_1 \rightarrow u$: index.html
$u \rightarrow d_1$: ads.com, **TP-cookie**
$d_1 \rightarrow u$: ads

B. Krishnamurthy and C. E. Wills. Generating a privacy footprint on the Internet. IMC 2006
Traceability

- **TP-Cookies** enable
  - **Spatial** tracking: Track over different domains
  - **Temporal** tracking: Identify subsequent visits
- **Referrer** reveals visited website

- Example of data collected by advertisers:
  - 10h00: www.nytimes.com, cookie
  - 10h02: www.nytimes.com, cookie
  - 11h00: www.facebook.com/friends, cookie
Existing Solutions

• All or nothing
  – Block requests to advertisers
  – Block TP-cookies
  – Allow all

• Same origin policy
  – “Only the server that sets a cookie can access it”
  – Prevents loss of data confidentiality or integrity
  – But too permissive with respect to online tracking
Outline

1. Online Advertising
   – Privacy Implications
   – Existing Solutions

2. Proposed Solution
   – Privacy friendly Cookie management
   – User centric

3. Preliminary Evaluation
   – Firefox Extension
Proposed Solution

• Trade-off privacy and traceability
  – Limit spatial and temporal tracking
  – User-centric solution

• Define policies for use of cookies
  – User privacy preferences
  – User advertisement preferences
  – Visited web site
Key Idea

- Maintain a **collection of cookies** in parallel
  - Sent cookie depends on the visited web site and advertiser

<table>
<thead>
<tr>
<th>Domain</th>
<th>Cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td>ads.com</td>
<td>c1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain</th>
<th>Website</th>
<th>Cookie</th>
</tr>
</thead>
<tbody>
<tr>
<td>ads.com</td>
<td>nytimes.com</td>
<td>c1</td>
</tr>
<tr>
<td>ads.com</td>
<td>google.com</td>
<td>c2</td>
</tr>
</tbody>
</table>
Key Technique

• To obtain a new cookie
  – Do not send existing cookies in HTML header
  – Server assigns a new cookie

• Privacy-Friendly cookie management
  – Alternate among cookies in collection
Approach 1

Limit use of TP-cookies per domain
Use for a limited number of times

1. $u \rightarrow d_1$: ads.com, www.nytimes.com, $c_1$
2. $u \rightarrow d_1$: ads.com, www.nytimes.com/technology, $c_1$
3. $u \rightarrow d_1$: ads.com, www.google.com, $c_2$

because nytimes != google
Approach 2

Limit use of TP-cookies per web site category and within categories
Use for a limited number of times

• Categories define type of web site
  – nytimes.com => news
  – Readily available (e.g., Alexa)

• Spatial tracking threshold $L_s$
  – Limits spatial tracking across web sites within categories
Approach 2

$L_s = 2$

<table>
<thead>
<tr>
<th>Category</th>
<th>u -&gt; $d_1$:</th>
<th>u -&gt; $d_1$:</th>
<th>u -&gt; $d_1$:</th>
<th>u -&gt; $d_1$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Email</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Because $3 > L_s$
Because search $!=\text{news}$
Because email $!=\text{search and email}$ $!=\text{news}$
Approach 3

Limit use of TP-cookies based on URLs and user preferences
Use for a limited number of times

• URLs
  – Leak information through referrer
  – google.com/search?q=julien

• Preferences on web site categories
  – Privacy: What users do not want to share
  – Advertising: What users want to get
Setting up Preferences

Google Ad preference manager

Rely on online social communities
Approach 3

\[
\sum_{b_i \in H(B)} w_1(b_i) \cdot w_2(b_i) < L_s
\]

<table>
<thead>
<tr>
<th>URLs ((w_1))</th>
<th>User Privacy Pref. ((w_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\(u \rightarrow d_1:\) ads.com, www.google.com, \(c_1\)

\(u \rightarrow d_1:\) ads.com, www.google.com/search?q=computers, \(c_1\)

\(u \rightarrow d_1:\) ads.com, www.facebook.com, \(c_1\)

\(u \rightarrow d_1:\) ads.com, www.facebook.com/search?q=nevena, \(c_2\)

\(L_s = 1\)

Because 0.1 + 1 > \(L_s\)
Outline

1. Online Advertising
   – Privacy Implications
   – Existing Solutions

2. Proposed Solution
   – Privacy friendly Cookie management
   – User centric

3. Preliminary Evaluation
   – Firefox Extension
Implementation

• Firefox extension: PrivaCookie
  – Proof of concept code
  – Get it on http://icapeople.epfl.ch/freudiger

• TP cookie detection
  – Compare originating URL with current URL

• Local cookie table
  – Link cookies with hidden server that caused its assignment and visible server hosting ads
  – (Cookie, visible server, hidden server)
Study

• Chose 10 pages from each of the top 20 domains
• Firefox extension pagestats
  – Runs browser in batch mode with list of web sites
  – A total of 200 pages
Number of hidden servers for each of the top 20 domains
Number of visible servers for each hidden server
# Top 10 associated visible servers connected with the most popular advertisers

<table>
<thead>
<tr>
<th>Hidden Server</th>
<th>Visible Servers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yahoo</td>
</tr>
<tr>
<td>doubleclick</td>
<td>$c_1</td>
</tr>
<tr>
<td>quantaserve</td>
<td></td>
</tr>
<tr>
<td>atmdt</td>
<td>✓</td>
</tr>
<tr>
<td>advertising</td>
<td></td>
</tr>
<tr>
<td>yieldmanager</td>
<td>✓</td>
</tr>
</tbody>
</table>

Extension caused 81 additional cookies assignments
Tracking Countermeasures

• Track based on IP
  – Anonymizer/Tor

• Track with
  – Cache cookies
  – Browser history
  – Plugins (e.g., Flash cookies)
  – Proposed policies also apply to those cases

• Cooperative tracking?
Conclusion

• We propose a solution for trading-off privacy & traceability
  – Protects user privacy
  – Allows for targeted online advertising
  – No changes required from advertisers
  – Puts users in control

• Key idea: Maintains a collection of cookies in parallel

• Future Work:
  – Implement approach 2 & 3
  – Implement Javascript support
  – Consider other parameters in approach 3
URL Weight

• Parse URL for n-grams
  – “search”
  – “id”
  – “username”

• Can be done automatically before visiting URL