CrowdSurf: Empowering Informed Choices in the Web

Hassan Metwalley
Stefano Traverso
Marco Mellia
Stanislav Miskovic
Mario Baldi

NARUS

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATIONS

12th Italian Networking Workshop
Cavalese, Italy - January 14, 2015
Motivations

- Online advertising
- E-commerce / recommendation
- Analytics in general

Each service can know everything about you.

CrowdSurf: Empowering Informed Choices in the Web

Data brokers selling lists of rape victims, AIDS patients

The data brokers: selling your personal information

Steve Kroft investigates the multibillion-dollar industry that collects, analyzes and sells the personal information of millions of Americans with virtually no oversight.
Third-Party Trackers

- **Third-party web tracking** refers to the practice by which a service records user web activities often for profit.

- **Many techniques**
  - Cookies
  - HTML5 LocalStorage
  - Finger printing (browser/OS/IP)
Quantify tracking activity
Are you tracked?

- Top third-party tracking services are contacted by **more than 95% of users**
  - **77%** of PC users contact the first tracker in less than 1 second
  - **71%** of services embed at least one tracking service

- **Yes, probably you are tracked!!**
Privacy and tracking: the role of HTTPS

- **HTTPS** is becoming more used [1]
  - This clearly improves people privacy
  - But it makes it harder to verify and regulate tracking services from a neutral third party observing traffic…

- How many third party tracking services are using HTTPS?

Privacy and tracking: the role of HTTPS

Percentage of HTTPS flows (%)

Third party tracking services

May 2014
May 2013
May 2012
Countermeasures?

- Some countermeasures are available as browser extensions
  - Disable cookie sending, disable javascripts, blacklisting,…
- ….but they worsen user’s browsing experience
- ….are ineffective for mobile users and few users use them
How to deal with this scenario?
How to know what information is being collected?
How to protect not experts?

Our proposal: give back to the user the control on web browsing!
Current Scenario

- Regulators are reacting, but they don't have enough data
- Still, now tracking services can collect information **without any authorization**
- It is mandatory to:
  - inform users about privacy leakage problem
  - help people to understand what information is collected
  - create a system in which each user can choose what share
- **Making the web economy sustainable**
CrowdSurf

- Crowd-Sourced system (called **CrowdSurf**)
  - Users can collaborate by providing implicit (e.g., traffic samples) and explicit (e.g., their opinion) information
  - They obtain information about web services, i.e., **advices**

- Cloud runs data mining algorithms to produce advices containing indications about **trustfulness** of web services

- Challenges
  - Unified system
  - Semi-supervised approaches
  - Overcome limitation of current systems (fragmented and not automated)
CrowdSurf Layer

- Mandatory Layer in the Internet stack
  - To handle HTTP(S) traffic **before encryption**
- It processes, filters and check all web traffic
  - Using the advices provided by the community
  - Under complete control of user
Testbed implementations and benchmarking

CrowdSurf: Empowering Informed Choices in the Web
CrowdSurf Prototype

- CrowdSurf layer implemented as a **Firefox extension**
  - Watch all HTTP(S) traffic, immediately before encryption
  - Support for both PCs and Mobile Devices

- **Cloud server** collects traffic samples, elaborates these and distributes advices

- Personalized policies
  - Three cases
    - Corporate
    - Kid
    - Paranoid

<table>
<thead>
<tr>
<th></th>
<th>Block</th>
<th>Redirect</th>
<th>log&amp;report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>Corp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td></td>
<td>Corp</td>
<td></td>
</tr>
<tr>
<td>Dropbox</td>
<td></td>
<td>Corp</td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td></td>
<td>Corp</td>
<td>Corp (--&gt; Bing)</td>
</tr>
<tr>
<td>YouTube</td>
<td>Corp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebay+Amazon</td>
<td>Corp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult Sites</td>
<td>Corp</td>
<td>Kid</td>
<td></td>
</tr>
<tr>
<td>Trackers</td>
<td>Par</td>
<td></td>
<td>Kid</td>
</tr>
<tr>
<td>Ads+NoJS</td>
<td>Par</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CrowdSurf Extension - Benchmark

- Performance impairment test
- 20 websites
  - 10 Top Global Sites
  - 8 News Sites
  - 6 Tracker-Free Sites
- Performance index:
  - Average rendering time
Clients have enough power to easily handle the extra load generated by possible CrowdSurf implementation!!
Smart algorithms to automatically flag possible problems
Algorithm to **automate advice generation**

- For a set of events in which hostname is different from referrer field

1. Extract all keys from the URLs in the set that includes a query string:
   
   `http://www.acmeAds.com/query?key1=X&......&keyN=Y`

2. For each hostname and for each key, investigate one-to-one mapping between the client and the values taken by each of the keys.

   - If a key value assumes a different value for each client, and does not change over time:
     - We found a "client identifier"
     - Mark the service as a "possible tracker"
### Automatic Tracker Detection - Results

- **HTTP traces**
- **Third party sites** in
  - Repubblica.it
  - YouTube
  - Facebook
- **Keys** suggest the exchange of **client identifiers**
- With CrowdSurf should be possible to check also on HTTPS

<table>
<thead>
<tr>
<th>Website</th>
<th>Third party sites</th>
<th>Keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repubblica</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YouTube</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Automatic Tracker Detection - Results

- **HTTP traces**
- **Third party sites in**
  - Repubblica.it
  - YouTube
  - Facebook
- **Keys suggest the exchange of** **client identifiers**

- **With CrowdSurf should be possible to check also on HTTPS**

<table>
<thead>
<tr>
<th>Website</th>
<th>Third party sites</th>
<th>Keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repubblica</td>
<td>pix04.revsci.net</td>
<td></td>
</tr>
<tr>
<td></td>
<td>su.addthis.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>track.adform.net</td>
<td></td>
</tr>
<tr>
<td>YouTube</td>
<td>bh.ams.contextweb.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eu-jet-01.sociomantic.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ib.adnxs.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.wajam.com">www.wajam.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uip.semasio.net</td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>adadvisor.net</td>
<td></td>
</tr>
<tr>
<td></td>
<td>data.bncnt.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>go.flx1.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ira.spysomeone.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tags.bluekai.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ww1.collserve.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.skyscanner.com">www.skyscanner.com</a></td>
<td></td>
</tr>
</tbody>
</table>
## Automatic Tracker Detection - Results

- **HTTP traces**
- **Third party sites in**
  - Repubblica.it
  - YouTube
  - Facebook
- **Keys suggest the exchange of client identifiers**
- **With CrowdSurf should be possible to check also on HTTPS**

<table>
<thead>
<tr>
<th>Website</th>
<th>Third party sites</th>
<th>Keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repubblica</td>
<td>pix04.revsci.net</td>
<td></td>
</tr>
<tr>
<td></td>
<td>su.addthis.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>track.adform.net</td>
<td></td>
</tr>
<tr>
<td>YouTube</td>
<td>bh.ams.contextweb.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eu-jet-01.sociomantic.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ib.adnxs.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.wajam.com">www.wajam.com</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uip.semasio.net</td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>adadvisor.net</td>
<td></td>
</tr>
<tr>
<td></td>
<td>data.bncnt.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>go.flx1.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ira.spysomeone.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tags.bluekai.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ww1.collserve.com</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.skyscanner.com">www.skyscanner.com</a></td>
<td></td>
</tr>
</tbody>
</table>
# Automatic Tracker Detection - Results

- **HTTP traces**
- **Third party sites in**
  - Repubblica.it
  - YouTube
  - Facebook
- **Keys suggest the exchange of client identifiers**
- **With CrowdSurf should be possible to check also on HTTPS**

<table>
<thead>
<tr>
<th>Website</th>
<th>Third party sites</th>
<th>Keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repubblica</td>
<td>pix04.revsci.net</td>
<td>id</td>
</tr>
<tr>
<td></td>
<td>su.addthis.com</td>
<td>puid</td>
</tr>
<tr>
<td></td>
<td>track.adform.net</td>
<td>icid</td>
</tr>
<tr>
<td>YouTube</td>
<td>bh.ams.contextweb.com</td>
<td>vgd</td>
</tr>
<tr>
<td></td>
<td>eu-jet-01.sociomantic.com</td>
<td>fpc</td>
</tr>
<tr>
<td></td>
<td>ib.adnxs.com</td>
<td>uuid</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.wajam.com">www.wajam.com</a></td>
<td>sExtCookieId</td>
</tr>
<tr>
<td></td>
<td>uip.semasio.net</td>
<td>install_timestamp</td>
</tr>
<tr>
<td>Facebook</td>
<td>adadvisor.net</td>
<td>bk_uuid</td>
</tr>
<tr>
<td></td>
<td>data.bncnt.com</td>
<td>uid</td>
</tr>
<tr>
<td></td>
<td>go.flx1.com</td>
<td>anuid, euid</td>
</tr>
<tr>
<td></td>
<td>ira.spysomeone.com</td>
<td>s</td>
</tr>
<tr>
<td></td>
<td>tags.bluekai.com</td>
<td>google_gid</td>
</tr>
<tr>
<td></td>
<td>ww1.collserve.com</td>
<td>bk_uuid</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.skyscanner.com">www.skyscanner.com</a></td>
<td>ksh_id</td>
</tr>
</tbody>
</table>
Future work

- **CrowdSurf System** presents some practical challenges that must be faced
  - Preserve privacy of users when sending contribution to the cloud
    - Anonymization is mandatory
  - Research community is called to design **automatic algorithms**, and propose **scalable** implementations
  - Users must become aware of the risk of web tracking and thus embrace crowdsurf (or similar) solution
  - It shall pass through a long and difficult standardization process to get accepted as a compelling technology

- “But it can work”.....
Thanks!
Automatic Tracker Detection

Client1  →  Id1
Client2  →  Id2
Client3  →  Id3
Client4  →  Id4
ClientN  →  IdN

Hostname1 and Key1

Possible Third-party tracking service
Automatic Tracker Detection

Hostname2 and Key1

Client1  --  Id1
Client2  --  Id1
Client3  --  Id3
Client4  --  Id4
ClientN  --  IdN

Not Possible
Third-party tracking service