Privacy leakage in users’ web browsing
Hassan Metwalley
2nd year 29th cycle

Supervisor: Marco Mellia, Politecnico di Torino
Collaborators: Stefano Traverso, Politecnico di Torino

Student’s ID 34473

Department of Electronics and Telecommunications
## Attended classes

<table>
<thead>
<tr>
<th>Classes</th>
<th>Hours</th>
<th>Points</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analisi statistica dei dati</td>
<td>15</td>
<td>20</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>Data mining: concetti e algoritmi</td>
<td>20</td>
<td>26.67</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>Cryptography</td>
<td>30</td>
<td>50</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>Information theory and codes</td>
<td>30</td>
<td>50</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>Il software libero</td>
<td>20</td>
<td>26.67</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>Parallel and distributed computing</td>
<td>25</td>
<td>33.33</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>Software and routing technologies in future networks</td>
<td>20</td>
<td>26.67</td>
<td>Hard Skills</td>
</tr>
<tr>
<td>(High qualification, Cisco Systems)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Etica Informatica</td>
<td>20</td>
<td>26.67</td>
<td>Soft Skills</td>
</tr>
<tr>
<td>Fondi competitivi per la ricerca</td>
<td>10</td>
<td>13.33</td>
<td>Soft Skills</td>
</tr>
<tr>
<td>Short Course on Entrepreneurship</td>
<td>7</td>
<td>9.31</td>
<td>Soft Skills</td>
</tr>
<tr>
<td>Internship NEC Europe Ltd., Heidelberg, Germany – From May 2015 to October 2015</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Soft Skills**

**Hard Skills + External Activities**

Total Hours = 37 (40 required)  
Total Points = 324.34 (200 required)
Research context and motivation

- In the last few years, there are numerous services that collect user’s information without that it knows nothing.
- **Third-party web tracking** refers to the practice by which a service register user web activities often for profit or advertising in general.
- Techniques
  - Cookies
  - HTML5 LocalStorage
  - Finger printing (browser/OS/IP)
- But each service possibly knows **everything** about you collecting data.

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.

Privacy leakage in users' web browsing
A growing business around our data - Example

ScorecardResearch, [...] a leading global market research effort that studies and reports on Internet trends and behavior. ScorecardResearch conducts research by collecting Internet web browsing data.
Available solutions

**Centralized in-network devices**
- Firewalls and middleboxes
  - Fail in case of encrypted traffic (HTTPS)
- HTTP/S Proxies
  - Lack scalability
  - Managed by third parties

**On-client**
- Browser plugins
  - Limited functionalities
  - Do not work on mobile devices
  - Can not access traffic generated by mobile apps
  - Not always open-source
  - Impossible to know how services are blacklisted

---

Google, Microsoft, and Amazon are paying Adblock Plus huge fees to get their ads unblocked

Lara O'Reilly  
Feb. 3, 2015, 6:57 AM  
60,452  
22
Adopted Methodologies

1. List of third party tracking services
   - Ghostery
   - Abine
   - Manual identification

2. Passive Monitoring of Internet Traffic
   - Tstat + DN-Hunter

3. Data Elaboration
   - Extraction of trackers’ activity
   - Big Data
Are you safe?

71% of Internet services embed at least one tracking service

On average, a user encounters the first tracker after as soon as she starts the browser.

Percentage of users (%)
Need something to...

...monitor the HTTP traffic generated by the client **before** encryption takes place

...report/manipulate/block transactions to undesired services

...automatically, but **under user’s control**

**Enable transparency and visibility**

**Take actions**

**Easy configurable**
**Novel contributions: CrowdSurf**

A cloud-based crowdsourced system

**Cloud**
1. A **controller** collects information about the services users visit
   - Explicit -> their opinion
   - Implicit -> traffic samples
2. Users’ contributions are analyzed by semi-automatic **data-analyzers** and the **advising community**
3. Results build **suggestions** about the reputation of services

**Client**
1. Users download the suggestions they like
2. the **CrowdSurf Layer** translates them into **rules**
3. Rules are used to take actions on the traffic leaving users’ device
4. Users are free to customize suggestions, depending on their needs
CrowdSurf Layer

- Embedded in all operating systems
- Sits between HTTP and transport layer
- Can handles HTTPS traffic before encryption
- Using regular expression, it can filter traffic depending on his trustfulness:
  - Block: layer blocks all traffic to a domain
  - Redirect: layer redirects traffic to another domain
  - Allow: layer doesn’t apply any change to traffic
  - Modify: layer modify part of the traffic (removes cookies, remove POST body, etc.)
  - Log and Report: layer send anonymized information to the cloud
CrowdSurf System

Open Controller
- **Collaborative approach** similar to Wikipedia and EFF
- Users called to improve the wisdom of the system
  - Provide traffic samples and opinions
  - Contribute to build data analyzers and suggestions

Third party Controller
- Build suggestions for **commercial purposes**
- Opens to a market of suggestions
- Users **might be** called to improve the wisdom of the system

Corporate Controller
- **Builds directly rules** for employees
- Employees are not allowed to customize rules
- All devices in the corporate domain follow the same rules
Adopted methodologies – CrowdSurf Prototype

**Controller**
- Java-based web service
- Communicates with CrowdSurf devices
- Hosts a data analyzer for identification of tracking sites
- Collects traffic samples
- Distributes suggestions

**CrowdSurf Layer**
- Implemented as a Firefox plugin
- Supports allow, block, redirect, log&report
- Works on both PCs and mobiles
- Available for download

Only +6-10% page loading time
Automatic Tracker Detector

Unsupervised methodology to identify third party trackers

- Intuition: trackers usually embed UIDs as URL parameters
- Procedure:
  1. Input: HTTP traffic samples provided by CS users
  2. Take all HTTP queries to third party services
     - http://acmetrack.com/query?key1=X&key2=Y
  3. Extract keys (key1, key2)
  4. Check the presence of key values uniquely associated to the users

Privacy leakage in users’ web browsing
Future work

- Architecture standardization
  - Protecting users' privacy
    - Anonymizing HTTP/S traffic is a challenging task
  - Usability
  - Protection from malicious biases
  - Web industry reaction

- More in real
  - More algorithms for trackers identification
  - Prototype on mobile devices
  - More understandable….."Can my grandparents manage the system?"
Publications

Accepted papers


Submitted papers


Submitted patent

H. Metwalley, S. Traverso, M. Mellia, “Method for detecting tracking web services”