CrowdSurf
Empowering Transparency in the Web

25 Aug 2016, ACM SIGCOMM, Florianopolis

Hassan Metwalley
Stefano Traverso
Marco Mellia
Stanislav Miskovic
Mario Baldi
Introduction
Do you know what you HTTP?
Example
Web tracking

Thousands of Web trackers collect our data

- Browsing histories
- Religious, sexual, and political preferences

- On average, the first tracker is met as soon as the browser starts
- Some trackers reach 96% of users
- 71% of websites host at least one tracker

The Open Question

How to **know** and **choose** which **services our data is exchanged** with and how?
Partial solutions

- **Firewalls and proxies**
  - Fail in case of encrypted traffic (HTTPS)
  - Lack scalability
  - Managed by third parties

- **Browser plugins**
  - Limited scope
  - No control on device traffic
  - Not transparent
A New System

Goal
Let **users** re-gain visibility and **control** on the **information** they exchange with **Web services**

Design Principles
- Holistic  
  working in any scenario
- Client-centric  
  available on any kind of device
- Practical, not revolutionary  
  use existing technology
- Crowd-sourced  
  knowledge built on a community of users
- Automatic  
  little engagement of the user
- Privacy-safe  
  never compromise users’ privacy
A controller collects information about the services users visit
- Explicit -> their opinion
- Implicit -> traffic samples

Users’ contributions processed by data-analyzers and the advising community

Results = suggestions about the reputation of services

Users download the suggestions they like

the CrowdSurf Layer translates them into rules

Rules = actions on users’ traffic
- Regexp + action
CrowdSurf Controllers

Open Controller
- Collaborative approach
- Users improve the wisdom of the system
  - Traffic samples and opinions
  - Build data analyzers and suggestions

Corporate Controller
- Builds directly rules for employees
- Employees can not customize rules
- All devices follow the same rules
The CrowdSurf Layer

HTTP

Regular Expression Matching

Action

- Block
- Redirect
- Allow
- Modify
- Log and Report

Suggestions to Rules

Open Controller

Corporate Controller

Anonymization

TCP

TLS
CrowdSurf in a picture

Opinions + Traffic samples

Suggestions

Open Controller

Ruled Interaction

Web Services

Rules

Traffic samples

Corporate Controller

Ruled Interaction
Proof of Concept
Prototype

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

Client
- Implemented as a Firefox plugin
- Supports block, redirect, log&report
Example of Data Analyzer:
Automatic Tracker Detector

Unsupervised methodology to identify third-party trackers [2]

- Observation:
  - 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) and their values
  4. Check the presence of key values uniquely associated to the users

Example of Data Analyzer: Automatic Tracker Detector


34 new third-party trackers found
Performance Implications of running CrowdSurf

Different user profiles

Paranoid Profile
- Blocks
  - adv/tracking
  - JS code
- Does not report traffic samples

Kid Profile
- Activates child protection rules
- Reports traffic to trackers

Corporate Profile
- Redirects
  - search.google.com to search.bing.com
- Blocks social networks, e-commerce sites, trackers
- Reports activity on DropBox
Impact on Web site loading time

Paranoid is 1.07 times faster than baseline
Kid is 1.08 times slower
Corporate is 1.18 time slower
Conclusion
Open Problems

- Lot of details to consider
- Design/develop/standardize a new network layer
- Protecting users’ privacy
  - Anonymizing HTTP/S traffic
- Usability
- Involve users to join
- Protection from malicious biases
Holistic, crowd-sourced system for the auditing of the information we expose in the Web

https://www.myermes.com
Thank you!
Need a new model that...

- Enables transparency and visibility
- Takes actions
- Under user’s control
- Monitor the HTTP traffic before encryption takes place
- Block/manipulate/report transactions to undesired services
- Automatic, but configurable
Example of Data Analyzer: Automatic Tracker Detector

Automatic Tracker Detector vs

Dataset
HTTP trace from ISP running Tstat
- 10 days of October 2014
- ~19k monitored users
- ~240k HTTP transactions per day

34 new third-party trackers found

<table>
<thead>
<tr>
<th>Third-party Trackers</th>
<th>Portal1</th>
<th>News1</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portal1</td>
<td>4</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>E-commerce1</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>E-commerce2</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>E-commerce3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Portal2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>E-commerce3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Porn</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Sportnews</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SearchEngine</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Third-party Trackers</th>
<th>News1</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>atemda.com</td>
<td>xid</td>
<td>26</td>
</tr>
<tr>
<td>x.bidswitch.net</td>
<td>user_id</td>
<td>13</td>
</tr>
<tr>
<td><a href="http://www.77tracking.com">www.77tracking.com</a></td>
<td>rand</td>
<td>12</td>
</tr>
<tr>
<td>rack.movad.net</td>
<td>us</td>
<td>9</td>
</tr>
<tr>
<td>ovo01.webtrekk.net</td>
<td>cs2</td>
<td>4</td>
</tr>
<tr>
<td>dis.criteo.com</td>
<td>uid</td>
<td>4</td>
</tr>
<tr>
<td>p.rfihub.com</td>
<td>bk-uuid</td>
<td>4</td>
</tr>
<tr>
<td>ib.adnxs.com</td>
<td>xid</td>
<td>1</td>
</tr>
</tbody>
</table>
Example
A growing business around our data

Loss of visibility and control

- HTTPS *protects* our privacy, but...
- ...prevents third parties to check *what’s going on under the hood* of encryption
- ...and *severely limits network functions*

“Child protection through the use of Internet Watch Foundation blacklists has become ineffective, *with just 5% of entries still being blocked* when HTTPS is deployed” [2]

Time to collect a dataset
Monitoring the Web

CrowdSurf Controllers

Open Controller
- Collaborative approach
- Users improve the wisdom of the system
  - Traffic samples and opinions
  - Build data analyzers and suggestions

Third party Controller
- Suggestions for commercial purposes
- Opens to a market of suggestions

Corporate Controller
- Builds directly rules for employees
- Employees can not customize rules
- All devices follow the same rules
CrowdSurf in a picture

- Open controller
- Third-party controller
- Corporate controller

Web Services:
- Google
- Twitter
- Yahoo
- Amazon
- Facebook
- YouTube
- Spotify

Data Analyzer

- Corporate
- Controller
- Private User Device
- Corporate Device

Suggestions
- Corporate Rules
- Web Browsing
- Traffic samples