CrowdSurf: Empowering Informed Choices in the Web

Original
CrowdSurf: Empowering Informed Choices in the Web / Metwalley, Hassan; Traverso, Stefano; Mellia, Marco; Miskovic, S.; Baldi, Mario. - STAMPA. - (2016), pp. 5-12. (Intervento presentato al convegno ACM SIGCOMM 2016 tenutosi a Florianópolis nel Agosto).

Availability:
This version is available at: 11583/2656559 since: 2017-05-16T12:11:22Z

Publisher:
ACM

Published
DOI:

Terms of use:
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)
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?
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

1. Network devices
   - Firewalls and proxies
     - Fail in case of encrypted traffic (HTTPS)
   - Lack scalability
   - Managed by third parties

2. Client devices
   - Browser plugins
     - Limited scope
     - No control on device traffic
     - Not transparent

---

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

Lara O'Reilly  📧  🗣️  1  Feb. 3, 2015, 6:57 AM  💬  60,452  🤝  22
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
CrowdSurf

Cloud

- 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

Client

- 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

Anonymization

Open Controller

Corporate Controller

TLS

TCP
CrowdSurf in a picture

Opinions + Traffic samples

Open Controller

Suggestions

Ruled Interaction

Web Services

Rules

Traffic samples

Corporate Controller
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


<table>
<thead>
<tr>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>sid</td>
<td>a b c</td>
<td>d e f</td>
</tr>
<tr>
<td>tmp</td>
<td>m m m m</td>
<td>n n n n</td>
</tr>
<tr>
<td>uid</td>
<td>x y z</td>
<td>x y z</td>
</tr>
</tbody>
</table>

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 times 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
CrowdSurf

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
  - Monitor the HTTP traffic before encryption takes place

- Takes actions
  - Block/manipulate/report transactions to undesired services

- Under user’s control
  - 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

### Third-party Trackers Found

<table>
<thead>
<tr>
<th>Third-party Trackers</th>
<th>34 new third-party trackers found</th>
</tr>
</thead>
<tbody>
<tr>
<td>atemda.com</td>
<td></td>
</tr>
<tr>
<td>x.bidswitch.net</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.77tracking.com">www.77tracking.com</a></td>
<td></td>
</tr>
<tr>
<td>rack.movad.net</td>
<td></td>
</tr>
<tr>
<td>ovo01.webtrekk.net</td>
<td></td>
</tr>
<tr>
<td>dis.criteo.com</td>
<td></td>
</tr>
<tr>
<td>p.rfihub.com</td>
<td></td>
</tr>
<tr>
<td>ib.adnxs.com</td>
<td></td>
</tr>
<tr>
<td>user_id</td>
<td></td>
</tr>
<tr>
<td>rand</td>
<td></td>
</tr>
<tr>
<td>us</td>
<td></td>
</tr>
<tr>
<td>cs2</td>
<td></td>
</tr>
<tr>
<td>uid</td>
<td></td>
</tr>
<tr>
<td>bk-uuid</td>
<td></td>
</tr>
<tr>
<td>xid</td>
<td></td>
</tr>
</tbody>
</table>

### Embedded Trackers

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

![Graph showing number of visits vs service rank with time to collect (T_c) on the y-axis and number of visits on the x-axis.]

googleanalytics
Monitoring the Web

[1] Popa, L. et al., "HTTP As the Narrow Waist of the Future Internet," ACM HotNets, 2010
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