[Proceeding] CrowdSurf: Empowering Informed Choices in the Web

Original Citation:

Availability:
This version is available at: http://porto.polito.it/2656559/ since: November 2016

Publisher:
ACM

Published version:
DOI:10.1145/2831347.2831349

Terms of use:
This article is made available under terms and conditions applicable to Open Access Policy Article ("Public - All rights reserved"), as described at http://porto.polito.it/terms_and_conditions.html

Porto, the institutional repository of the Politecnico di Torino, is provided by the University Library and the IT-Services. The aim is to enable open access to all the world. Please share with us how this access benefits you. Your story matters.

(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?
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 [1]

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

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

On-client
- 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
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
Cloud
- **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

Open Controller

Corporative Controller

Anonymousization

CrowdSurf Layer

Rule Processor

TLS

TCP
CrowdSurf in a picture

- Web Services
- Opinions
- Traffic samples
- Suggestions
- Ruled Interaction
- Corporate Controller
- Rules
- Traffic samples
- Open 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


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

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

### Embedded Third-Party Trackers

<table>
<thead>
<tr>
<th>Third-Party Trackers</th>
<th>Portal1</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>portal1</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>E-commerce1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>E-commerce2</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>E-commerce3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Portal2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Porn</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Sportnews</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>SearchEngine</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**34 new third-party trackers found**

- atemda.com
- x.bidswitch.net
- www.77tracking.com
- rack.movad.net
- ovo01.webtrekk.net
- dis.criteo.com
- p.rfihub.com
- ib.adnxs.com
- bidderuid
- user_id
- rand
- us
- cs2
- uid
- bk-uuid
- xid

26 August 2016  CrowdSurf - Stefano Traverso
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