Tracking Catalog: Uncovering and analyzing user tracking on the Internet

Tomasz Bujlow Valentín Carela-Español Pere Barlet-Ros

Computer Architecture Dept. (DAC)
UPC BarcelonaTech
`{tbujlow,vcarela,pbarlet}@ac.upc.edu`

DTL Workshop, Nov. 20, 2014

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

About us

- Broadband Communications Group (CBA)[1]
  - Research group at UPC BarcelonaTech
  - Several topics: New Internet architectures, optical networking, nano-networking, SDN, network measurements, . . .
- Network monitoring group within CBA[2]
  - 1 Full Professor, 1 Associate Professor, 2 Post-Doc, PhD Students
  - Network measurements, traffic classification, *machine learning*
  - Apply our expertise to the field of online privacy and DTL

[1] http://www.cba.upc.edu

[2] http://www.cba.upc.edu/monitoring

Motivation

- Different entities interested in tracking our online activity
  - Economical, political, security, even governmental interests
  - Examples: Verizon[3], NSA[4], political campaigns[5], . . .
- Users would like to know *when* and *how* they are tracked
  - Disable tracking when desired
  - Decide whether accessing a resource despite tracking
- Tracking is almost impossible to avoid
  - *Do not track* option is not respected
  - Erasing cookies is not always enough
  - Fingerprinting is hard to avoid (even in private browsing mode)

[3] How Verizon's Advertising Header Works, Web Policy (2014). `http://webpolicy.org/2014/10/24/how-verizons-advertising-header-works/`

[4] NSA uses Google cookies to pinpoint targets for hacking. `http://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking`

[5] How President Obama's campaign used big data to rally individual voters. `http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters`

Existing tools

- Tools available to users
  - Check browser/privacy settings (e.g. Panopticlick)
  - Block tracking traffic (e.g. Adblock Plus, Privacy Badger)
  - Visualize third-parties (e.g. Lightbeam)
  - Safer browsing (e.g. Private browsing mode, Tor, DuckDuckGo)
- Research projects
  - XRay: Transparency for the web (Columbia University)
  - $heriff for price discrimination (Telefonica, UPC BarcelonaTech)
  - TaintDroid (Intel Labs, Penn State, Duke University)
- No tools available to know *when* and *how* we are tracked

Objective

- Tracking Catalog: Tell how sites are tracking us
  - Identify tracking mechanisms used by popular sites
  - Including also third parties
  - Analyze existing (and future) tracking mechanisms

- Provide it as a service for the users (e.g. browser plugin)
  - Users will know *when* and *how* they will be tracked
  - Users will be able to decide whether they access the site or not
  - Increase transparency and trust in “good” services

Methodology

- Continuously visit and analyze most popular sites
  - E.g. Alexa top-10K and their third parties
  - Automatize the process (e.g. Selenium WebDriver, FourthParty)
  - Apply Machine Learning to detect patterns and build signatures
- Analyze tracking mechanisms
  - Collect most invoked Javascripts and analyze them
  - Discover new (unknown) tracking methods
- Provide it as an *open source* tool to the DTL community
  - Users and researchers can contribute (data and new functions)
  - Crowdsourcing and distributed infrastructures (e.g. PlanetLab)
  - Analyze all the collected data and publish a report

Tracking mechanisms

- HTTP cookies
- Cookie leaks and syncing
- Fingerprinting (e.g. Canvas)
- Web cache and ETags
- HTTP Redirect headers
- Headers in outgoing HTTP requests
- Explicit web-form authentication
- HTML5 Local Storage
- Flash cookies and LocalConnection object
- Browsing history
- Evercookies
- Many others . . .

An example: Canvas fingerprinting

- Some tracking mechanisms are difficult to uncover and block
  - Ustream - The leading HD streaming video platform (www.ustream.tv - Alexa rank: 1048) is using canvas fingerprint
  - Fingerprinting script: `http://d1g3gvqfdsvkse.cloudfront.net/assets/featurekicker.js`

```
getCanvasFingerprint: function() {
    var e = document.createElement("canvas"),
        t = e.getContext("2d"),
        n = "http://valve.github.io";
    return t.textBaseline = "top", t.font = "14px ’Arial’", t.textBaseline
= "alphabetic", t.fillStyle = "#f60", t.fillRect(125, 1, 62, 20),
t.fillStyle = "#069", t.fillText(n, 2, 15), t.fillStyle = "rgba(102, 204,
0, 0.7)", t.fillText(n, 4, 17), e.toDataURL()
}
```

Open questions

- Questions we expect to answer from our study
  - How prevalent is each tracking mechanism?
  - How tracking depends on different parameters?
  - How tracking is obfuscated?
  - Which tracking mechanisms have not been detected yet?
- Other questions we would like to address
  - What is the accuracy of each tracking method?
  - For what purpose is each tracking method used?
  - Are our social network connections used for targeted advertising?
  - Is our activity while not logged in attached to our personal profile?

Tracking Catalog: Uncovering and analyzing user tracking on the Internet

Tomasz Bujlow Valentín Carela-Español Pere Barlet-Ros

Computer Architecture Dept. (DAC)
UPC BarcelonaTech
{tbujlow,vcarela,pbarlet}@ac.upc.edu

DTL Workshop, Nov. 20, 2014

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH