Web tracking technologies
and privacy protection on the Web

Nataliia Bielova

Inria Rennes,

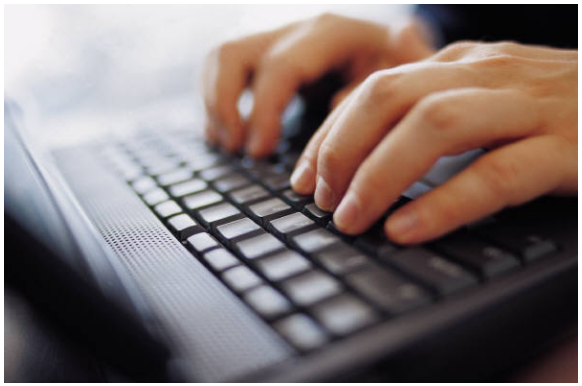
25 October 2013

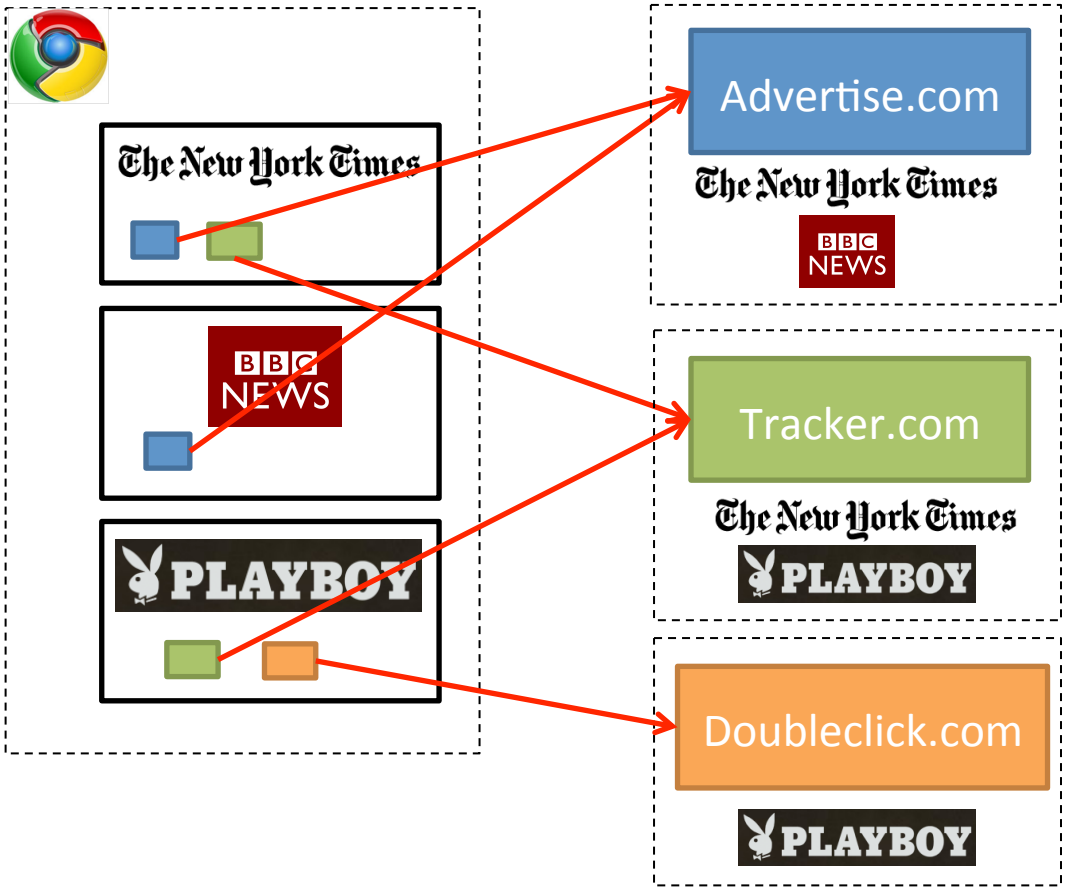
Back in 1993…

"On the Internet, nobody knows you're a dog."

Today…

Web Tracking



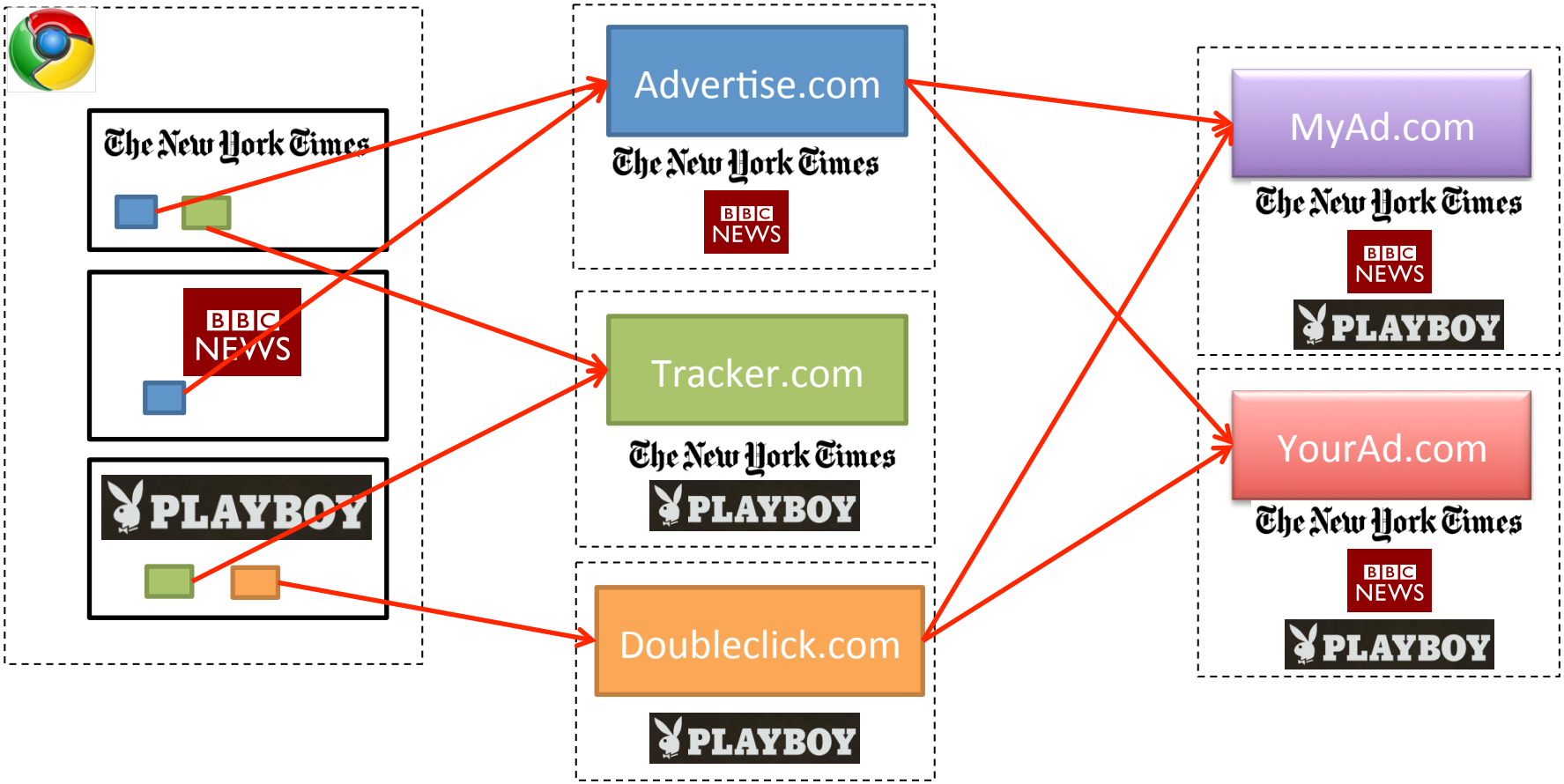


Bigger browsing profiles
= increased value for trackers
= reduced privacy for users

(Hypothetical tracking relationships only.)

Today…

Web Tracking



(Hypothetical tracking relationships only.)

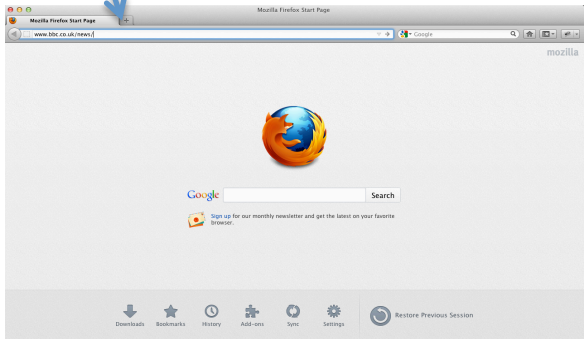
In this talk…

- What is Web Tracking?
  - How does it work?
  - Its basic components
- How to protect your privacy on the Internet?
  - What kind of defenses you can set in your browser?
  - Are they effective?
  - Which research solutions are proposed?
  - What about EU laws and regulations?

HTTP protocol is stateless

www.bbc.co.uk/news/

Web browser



URL path: bbc.co.uk/news...
Method: GET
Cookies: session-id=**2082787201l** & ...
...

HTTP request

HTTP response

Status: 200 OK
Content: HTML page
Set-cookies: session-id=**2082787201l** & ...

Web server

Cookie Database

bbc.co.uk/news:
session-id=2082787201l

- Launching new HTTP requests

- Mapping requests to apps
- Contacting DBs
- Constructing responses

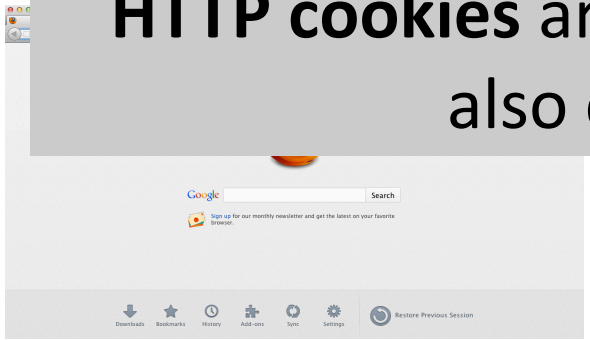
HTTP protocol is stateless

www.bbc.co.uk/news/

URL path: bbc.co.uk/news

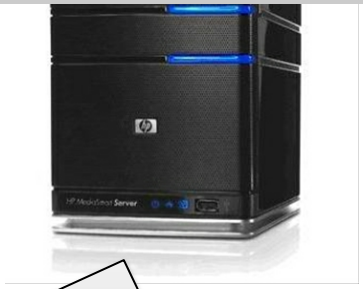
High-level point:

HTTP cookies are useful for web session handling, but also **can be used to track users.**



HTTP response

Status: 200 OK
Content: HTML page
Set-cookies: session-id=**2082787201l** & …



Cookie Database

bbc.co.uk/news:
session-id=2082787201l

- Launching new HTTP requests

- Mapping requests to apps
- Contacting DBs
- Constructing responses

Mechanisms Required By Trackers

- Ability to store/create user identity in the browser
  - HTTP cookies
  - other HTTP headers
  - other browser storages
  - device fingerprinting:
    - browser properties
    - OS properties
    - IP address…

Stateful tracking

Stateless tracking

- Ability to communicate user identity back to tracker
  - HTTP request headers
  - JavaScript

Stateful Tracking

COOKIES AND OTHER BROWSER STORAGES

Cookies: first- & third-party

Origin

nytimes.com

Origin

b.scorecardresearch.com

nytimes.com

```
<iframe
src=b.scorecardresearch.com>
```

Html page +

JavaScript code

```
</iframe>

<script src=facebook.net>
```

JavaScript code

```
</src>
```

First-party

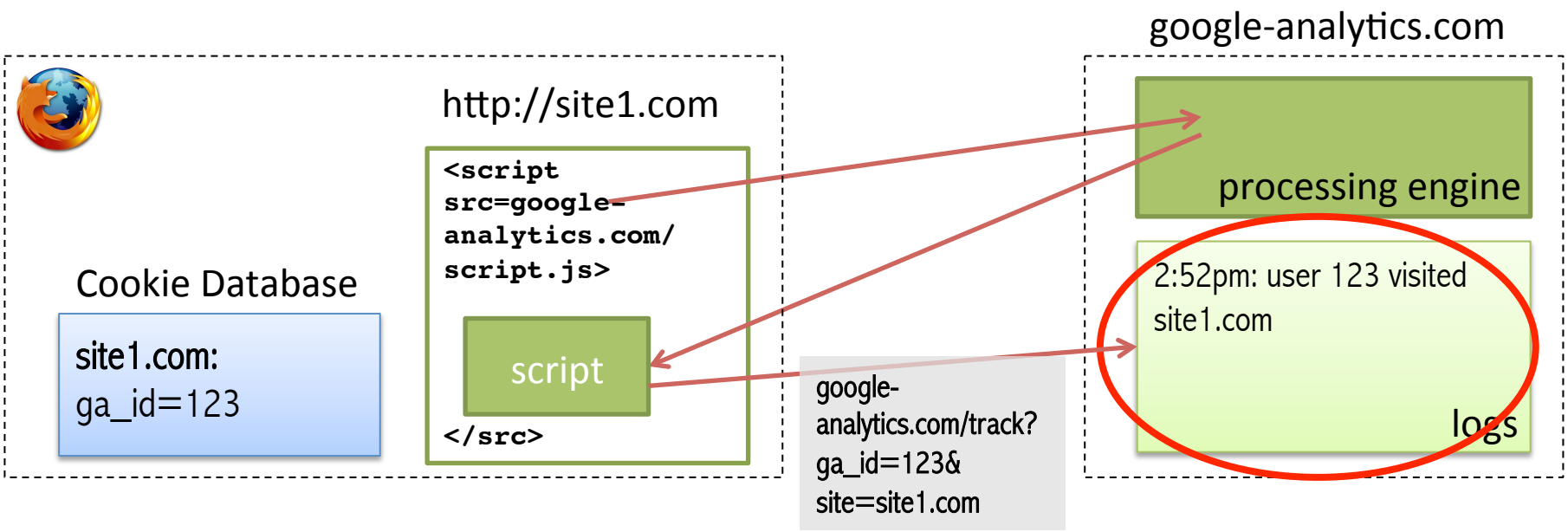
can read/write cookies of nytimes.com

Third-party

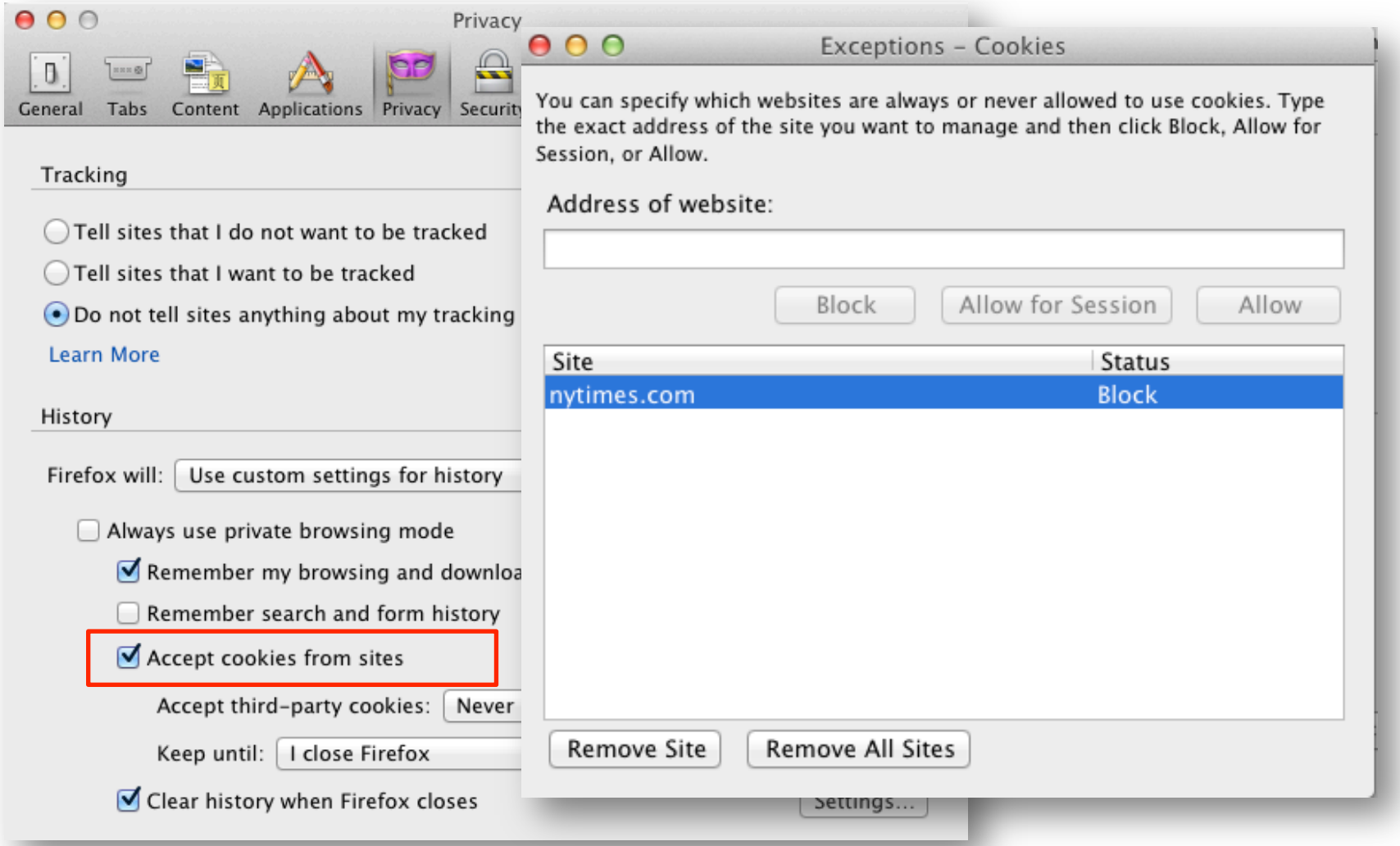
can read/write cookies of b.scorecardresearch .com

Within-Site Tracking

First-party cookies are used to track repeat visits to a site.

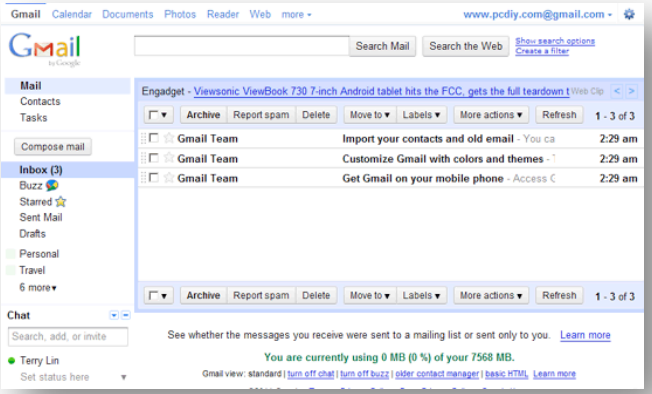


First-party cookie setting



First-party cookies benefits

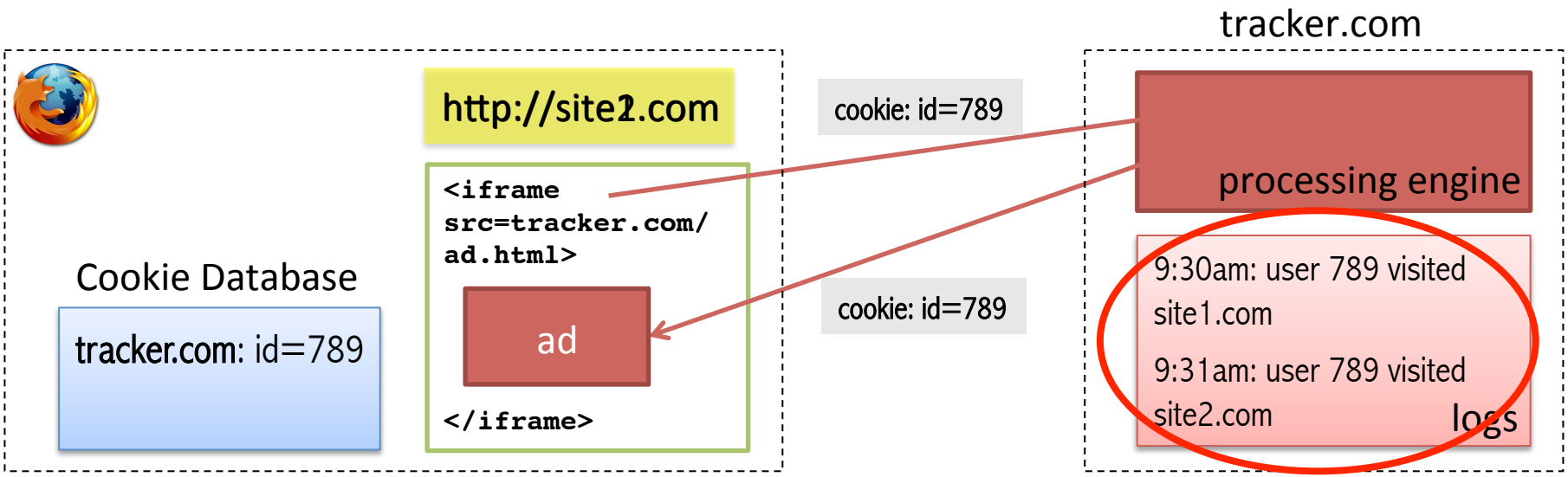
- Keep the session through different windows/tabs



- Website owners can evaluate
  - website statistics
  - popularity of certain pages
  - popularity of links
  - selected and copied phrases

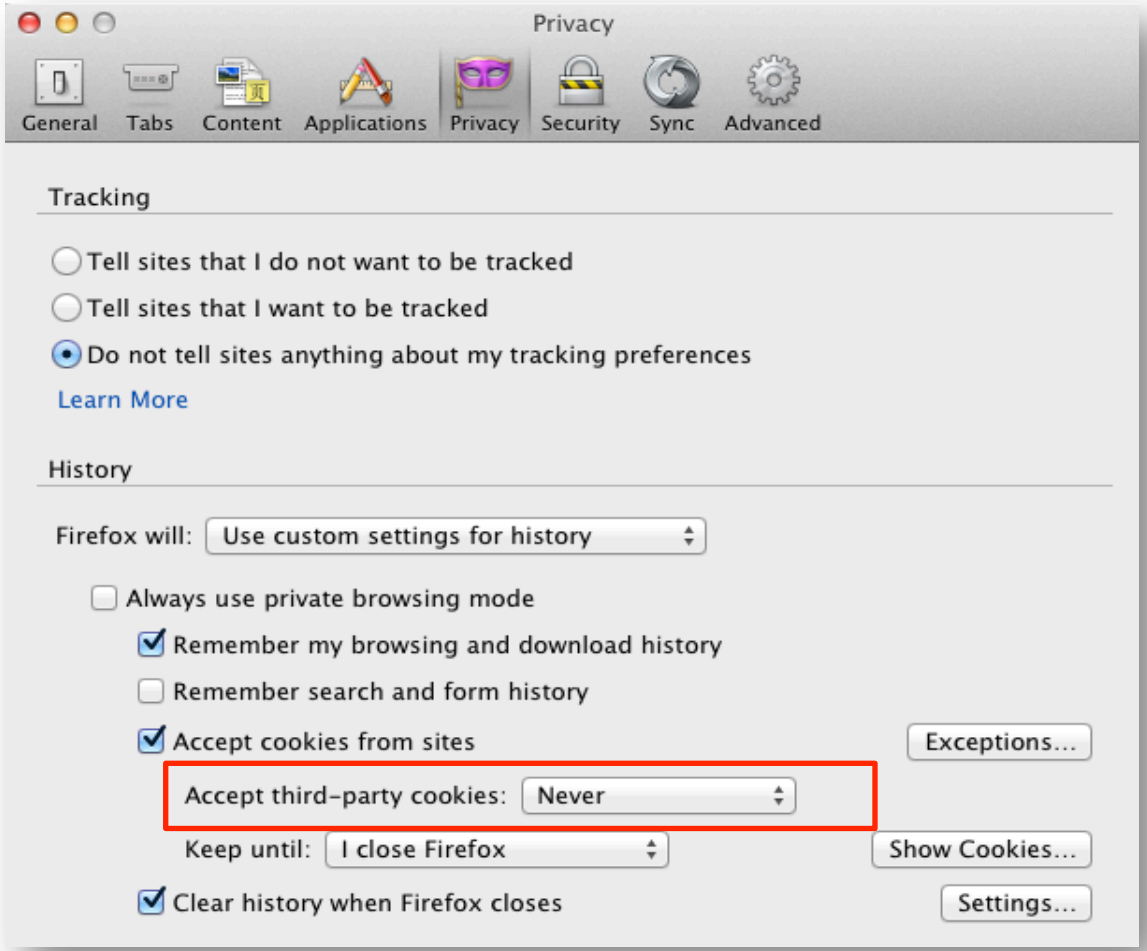
Cross-sites Tracking

Third-party cookies are used by trackers included in other sites to create profiles.



Courtesy of Franziska Roesner

Practical protection: Third-party cookies blocking



Practical protection: Third-party cookies blocking

- Does not influence your browsing experience
- Does not adjust advertisements for you

- **So why are third-party cookies still there?**
  - It's a business of advertisement companies

- **"How much are you worth?"**
  - New plugin shows what advertisers pay for you
  - http://yourvalue.inrialpes.fr by Inria Privatics team

Cookie respawning

- Cookies **can respawn** even if the user has deleted them

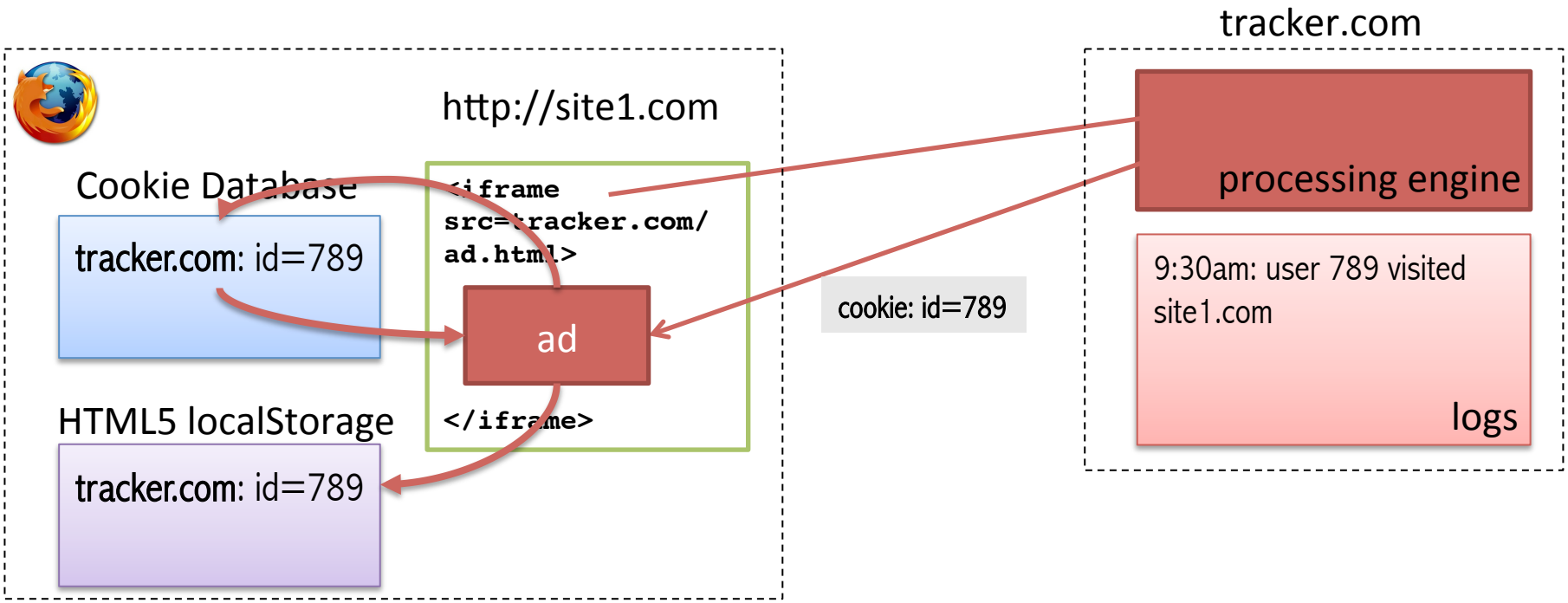
- Ashkan Soltani, August 2011 KISSmetrics and Hulu.com lawsuits
  - HTML5 localStorage
  - Flash LSOs
  - Other http headers

- Samy Kamkar: "evercookie" :
  - Even more storage mechanisms

Respawning - local storages

- KissMetrics lawsuit: HTML5 localStorage (across sessions)

User leaves the page

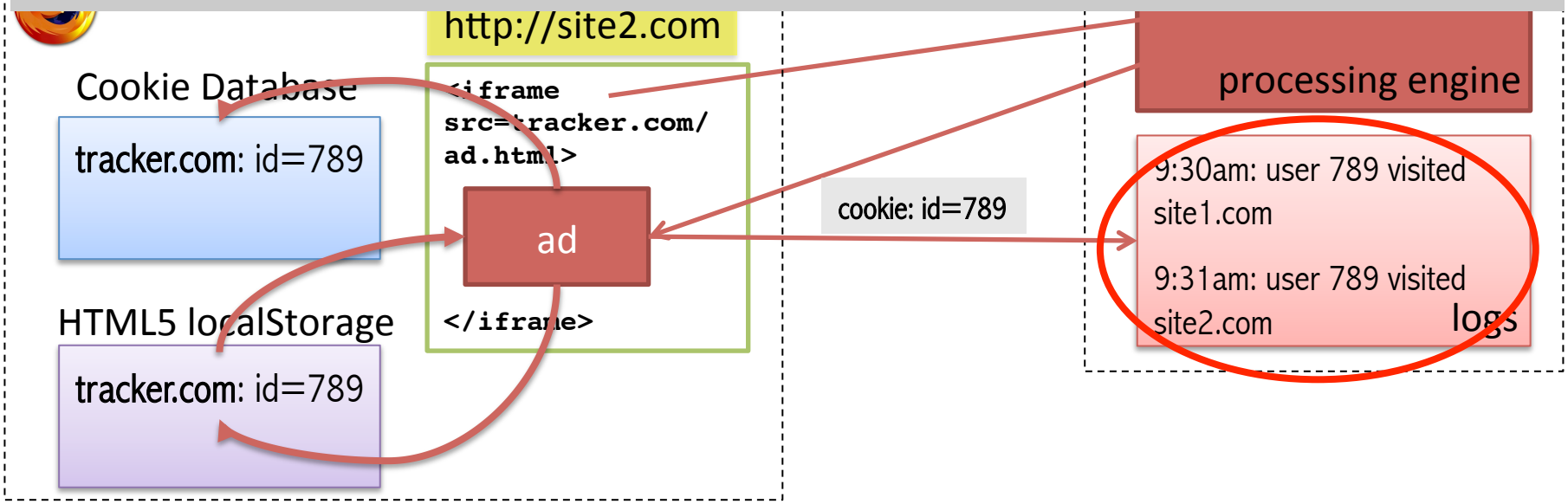


Respawning - local storages

- KissMetrics lawsuit: HTML5 localStorage (across sessions)

High-level point:

If these **local storages didn't store a copy of cookies**, this tracking would not be detected!



Not only Respawning, but Tracking

- Demonstrated by Vladimir Klimontovich, October 2012
- HTML5 localStorage instead of cookies

```
//Iframe code (http://pixel.sample-ad-exchange.com/iframe.html)
<html>
<head></head>
<body>
<script type="text/javascript">
    var userId = localStorage.getItem("user_id");
    if (userId == null) {
        //set user is if user is unknown
        userId = Math.random();
        localStorage.setItem("user_id", userId);
    }
    var img = document.createElement('img');
    img.src = "http://pixel.sample-ad-exchange.com/pixel.gif?user_id=“ + userId;
    var body = document.getElementsByTagName('body')[0];
    body.appendChild(img);
    </script>
</body>
</html>
```

Not only Respawning, but Tracking

- Demonstrated by Vladimir Klimontovich, October 2012
- HTML5 localStorage instead of cookies

```
//Iframe code (http://pixel.sample-ad-exchange.com/iframe.html)
<html>
<head></head>
<body>
```

One more point:

It **works in most browsers!** (except for Chrome)

```
        localStorage.setItem("user_id", userId);
    }
    var img = document.createElement('img');
    img.src = "http://pixel.sample-ad-exchange.com/pixel.gif?user_id=“ + userId;
    var body = document.getElementsByTagName('body')[0];
    body.appendChild(img);
    </script>
</body>
</html>
```

JavaScript access*

third-party cookies **blocked** in browser settings

	Third-party cookies	Third-party localStorage
	blocked	blocked
	blocked	allowed
	blocked	allowed
	blocked	allowed
	allowed	allowed

localStorage can be used instead of cookies!

also cookies can still be used!

*Checked on 24/10/2013

JavaScript access*

third-party cookies **blocked** in browser settings

	Third-party cookies	Third-party localStorage
	blocked	allowed
	blocked	allowed
	allowed	allowed

High-level point:

Cross-site tracking is **possible via JavaScript** even with third-party cookies blocking option!

instead of cookies!

also cookies can still be used!

*Checked on 24/10/2013

Respawning - other HTTP headers

- Was first described by Dean Gaudet in 2003:

> "other than cookies, there's typically only one other type of data a webserver can cause a browser to store on its local harddrive - **cacheable web content**."

=> Etag HTTP header

Respawning - Etag header

- KissMetrics lawsuit, August 2011

```
INITIAL REQUEST HEADER:
  GET /i.js HTTP/1.1
  Host: i.kissmetrics.com

INITIAL RESPONSE HEADER:
  Etag: "Z9iGGN1n1-zeVqbgzrlKkl39hiY"
  Expires: Sun, 12 Dec 2038 01:19:31 GMT
  Last-Modified: Wed, 27 Jul 2011 00:19:31 GMT
  Set-Cookie: _km_cid=Z9iGGN1n1-zeVqbgzrlKkl39hiY;
              expires=Sun, 12 Dec 2038 01:19:31  GMT;path=/;

SUBSEQUENT REQUEST HEADER (PRIVATE BROWSING MODE WITH ALL COOKIES BLOCKED):
  GET /i.js HTTP/1.1
  Host: i.kissmetrics.com
  If-None-Match: "Z9iGGN1n1-zeVqbgzrlKkl39hiY"
```

Respawning - Etag header

- KissMetrics lawsuit, August 2011

```
INITIAL REQUEST HEADER:
  Last-Modified: Wed, 27 Jul 2011 00:19:31 GMT
  Set-Cookie: _km_cid=Z9iGGN1n1-zeVqbgzrlKkl39hiY;
              expires=Sun, 12 Dec 2038 01:19:31  GMT;path=/;

SUBSEQUENT REQUEST HEADER (PRIVATE BROWSING MODE WITH ALL COOKIES BLOCKED):
  GET /i.js HTTP/1.1
  Host: i.kissmetrics.com
  If-None-Match: "Z9iGGN1n1-zeVqbgzrlKkl39hiY"
```

High-level point:

If **Etag header didn't store a copy of cookies**, this tracking would not be detected!

Practical solutions

- Browser setting: **block third-party cookies**
  - Protects from tracking (purely) via cookies
  - Does not protect from cookie respawning
  - Does not protect from tracking via other storages

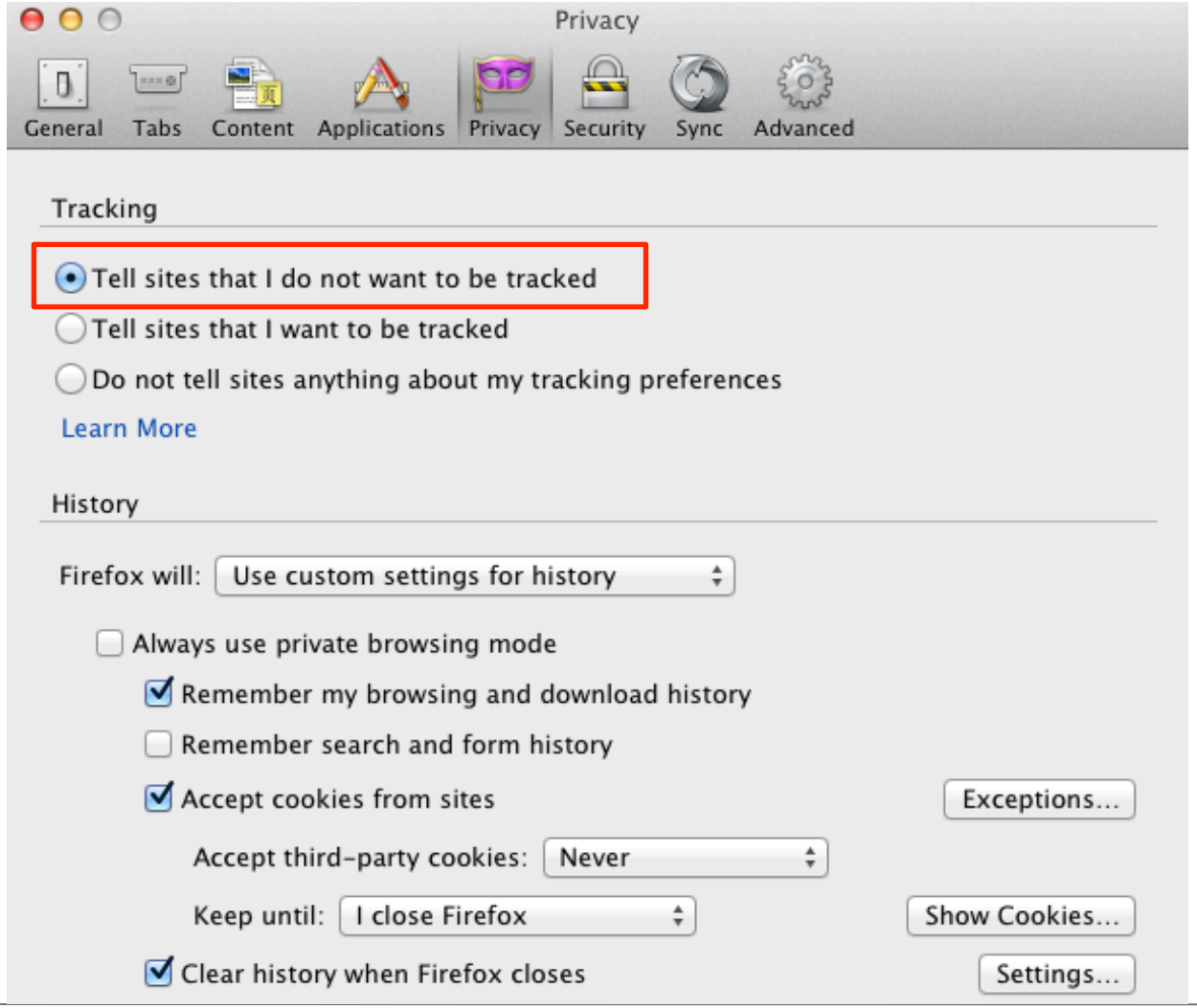
- Browser extension: block scripts/requests **only from known advertisement/tracking companies**
  - Does not protect from tracking by other companies
  - Does not protect form tracking by the main website

Research solutions

- **Information flow control**
  - Analyses JavaScript and prevents cookie leakage
    - to remote servers & to other storages
  - Strong formal guarantee
    - sensitive data sources (cookies) do not interfere with untrusted data sinks (servers, storages)
  - Several implementations:
    - Enhanced web browser FlowFox [De Groef et al. CCS'12]
    - FireFox plugin ZaphodFacets [Austin&Flanagan POPL'12]

Do-not-track and EU ePrivacy directive

Do-Not-Track (DNT)



Do-Not-Track (DNT)

- Tracking preference expression
  - New HTTP request header **DNT:1**
  - Optional HTTP response header **Tk:1** (server is compliant)

- **How** the web servers should enforce DNT?
  - "do-not-track" → "do-not-target"
  - do not target the users based on collected data
  - but still allow data to be collected

- Did anything actually change?
  - IE 10 adds **DNT:1** by default, Yahoo! and Apache ignore it.

EU ePrivacy Directive 95/46

w.r.t. Stateful tracking

Actual Regulation

2002/58/EC:

- users should **be able to refuse** to have info stored in their browser

2009/136/EC:

- users should **give a consent** to have info stored in their browser

Interpretation

EU states:

- users can **change their cookie settings**

Some EU states:

- cookie setting is an **implicit consent**

Most of other EU states:

- no, we need other standard with **explicit consent**

Thanks to EU ePrivacy Directive

European Comission **is very interested** in sound DNT

Neelie Kroes (Vise-President of the EC on Digital Agenda)

- June 2011:
  - **It's not enough** that web businesses say **they honour DNT**
  - Citizens need to be sure **what exactly companies do**.
- January 2012:
  - EU ePrivacy directive is **not only about cookies**!
  - We need a sound **Do Not Track (DNT)** standard!
- October 2012:

> "I said it last June, and I said it in January. Loud and clear. But, for the avoidance of doubt, I will say it again today: the **DNT standard must be rich and meaningful enough to make a difference** when it comes to protecting people's privacy."

Stateless Tracking

DEVICE FINGERPRINTING AND EU EPRIVACY DIRECTIVE

ars technica

MAIN MENU MY STORIES: 25 FORUMS SUBSCRIBE JOBS

RISK ASSESSMENT / SECURITY & HACKTIVISM

Top sites (and maybe the NSA) track users with "device fingerprinting"

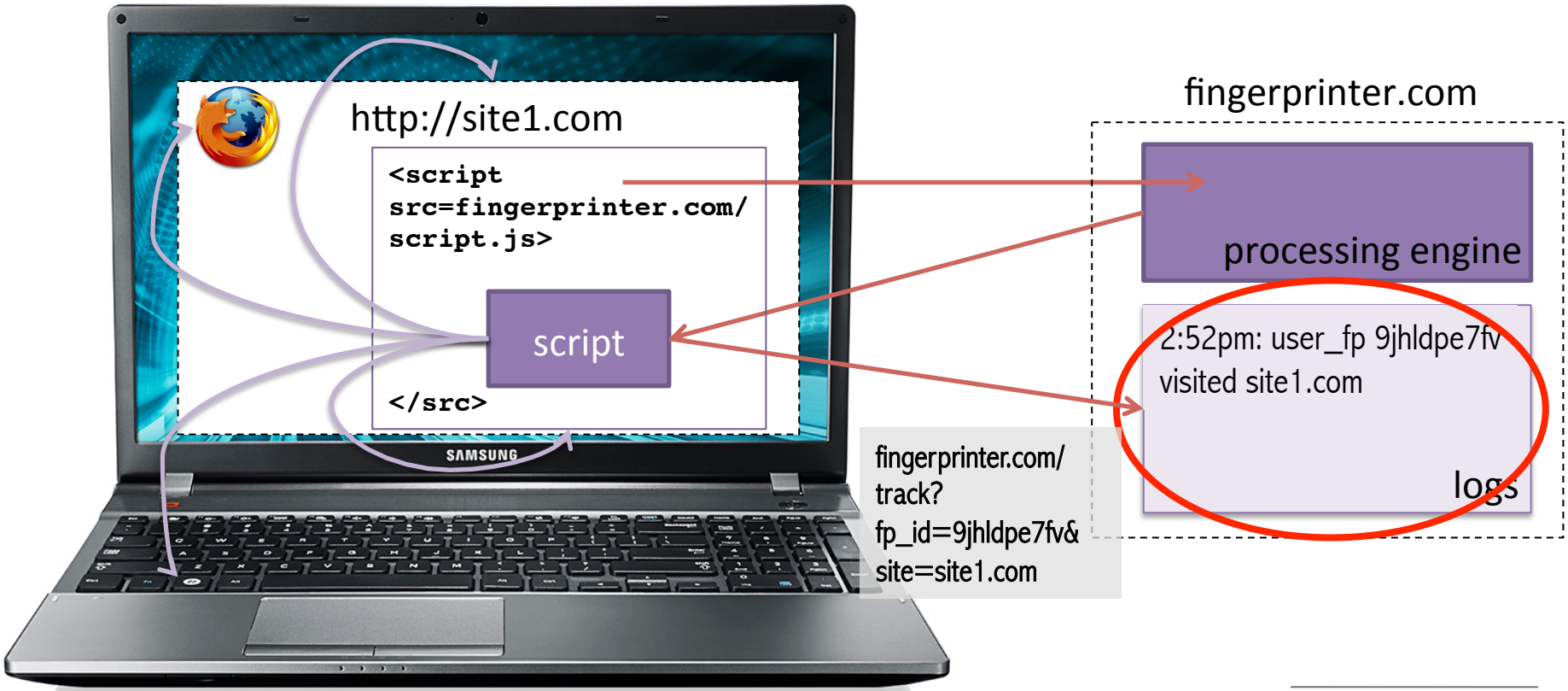
May make it easier to follow privacy-minded users on the darknet.

by Dan Goodin - Oct 11 2013, 7:31pm CEST

PRIVACY 98

Tracking by device fingerprinting

Browser and operating system properties are used to track repeated visits cross sites.



Tracking by device fingerprinting

Panopticlick

How Unique — and Trackable — Is Your Browser?

Your browser fingerprint **appears to be unique** among the 2,419,678 tested so far.

Currently, we estimate that your browser has a fingerprint that conveys **at least 21.21 bits of identifying information.**

Idea: **distinguish user's browsers** by accessing browser features and using their probability distributions

Panopticlick results [Eckersley, PETS'2010]

Browser property	Source
User Agent (browser name and version, OS version, etc)	HTTP
	JS
HTTP_ACCEPT header	HTTP
Browser plugin details	**JS**
Time zone	JS
Screen size and color depth	JS
System fonts	**Flash/JS**
Cookies enabled?	HTTP
	JS
Supercookies test	JS

83.6% are unique among **all observed**
94.2% are unique among **browsers with Flash and Java**

Plugins and fonts are the most identifying metrics!

[Acar etal. CCS'2013]

Prevalence of device fingerprinting

- First large-scale study
  - Flash-based: 97 sites out of 10 000
  - JavaScript-based: 404 sites out of 1 million
  - … and this is just a lower bound!

- Main idea:
  - scripts that access too many browser and device properties (e.g., more than 30 fonts) potentially implement fingerprinting.

EU ePrivacy Directive 95/46

w.r.t. Stateless tracking

> Art. 7: Member States shall provide that **personal data** **may be processed** **only if**:
> (a) the data subject has unambiguously given his consent;

- '**personal data**' = any information relating to an identified or identifiable natural person ('data subject')
- '**an identifiable person**' = one who can be identified, directly or indirectly
- '**processing of personal data**' = any operation or set of operations which is performed upon personal data

Scarlet vs Sabam case (Nov 2011): **IP addresses** are protected personal data because they allow those users to be precisely identified.

EU ePrivacy Directive 95/46

w.r.t. Stateless tracking

> Art. 7: Member States shall provide that **personal data** **may be processed** **only if**:
> (a) the data subject has unambiguously given his consent;

High-level point:

Web browser fingerprints are personal data

- '**an identifiable person**' = one who can be identified, directly or indirectly
- '**processing of personal data**' = any operation or set of operations which is performed upon personal data

Scarlet vs Sabam case (Nov 2011): **IP addresses** are protected personal data because they allow those users to be precisely identified.

[Acar etal. CCS'2013]

Practical solutions

- Tor Browser: **not easy to provide 100% unlinkability**
  - limited user base => even a partial fingerprint may uniquely distinguish a Tor user
  - bug found: OS fonts can be checked through CSS rule
- FireGloves browser extension: **not efficient**
  - spoofs browser's user-agent and platform
  - inconsistencies with reality found via JavaScript
  - fonts can still be effectively detected
    - via text's dimensions

Research solution:

Quantitative Information Flow analysis against Fingerprinting

WITH FREDERIC BESSON AND THOMAS JENSEN

How to distinguish fingerprinting scripts from useful scripts?

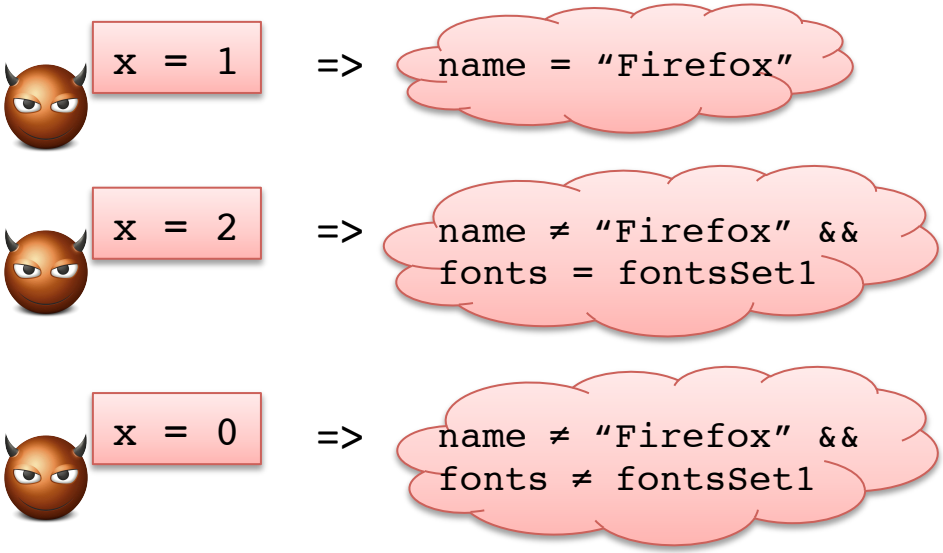
Script (possibly) provided by a tracker

```
var x = 0;
if (name == "FireFox") {
    x = 1;
}
else {
    if (fonts == fontsSet1) {
        x = 2;
    }
}
output x;
```

How much information does tracker learn about the user
if x=0?
if x=1?
if x=2?

Quantitative information flow

```
var x = 0;
if (name == "Firefox") {
    x = 1;
}
else {
    if (fonts == fontsSet1) {
        x = 2;
    }
}
output x;
```



Depending on user's browser, **different executions** of this script **leak different quantity** of information!

Challenge:

How to **automatically** evaluate **how much information** a tracker **learns through one execution** of the script?

[Besson, Bielova, Jensen CSF'2013]

Quantification of leakage

- **Self-information, or "surprisal"**
  - "amount of information about the identity" [Eckersley'10]
  - = beliefs for deterministic programs [Clarkson, Myers, Schneider'07]

$$I(A) = -\log_2 P(A)$$

```
var x = 0;
if (name == "Firefox"){
    x = 1;
}
output x;
```

Popularity of "FireFox" is 21%

$I(\texttt{name = "Firefox"}) = -\log_2 0.21 = 2.25$ bits

$I(\texttt{name} \neq \texttt{"FireFox"}) = -\log_2 0.79 = 0.34$ bits

- **Entropy-based definition = average leakage for all browsers!**

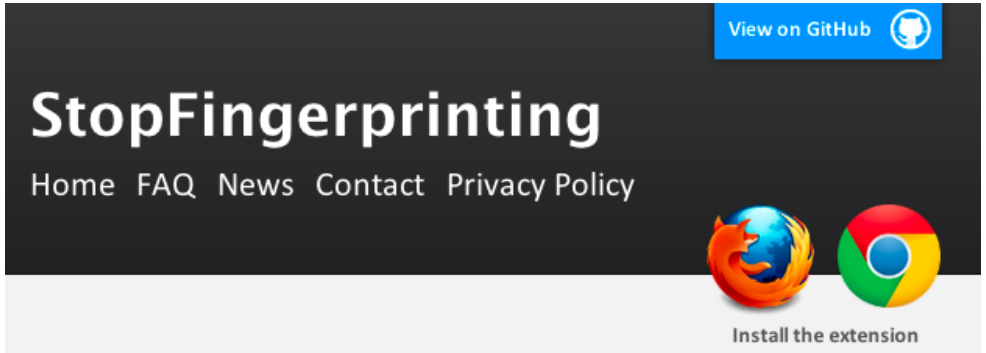
$H(\texttt{name}) - H(\texttt{name} \mid \texttt{x}) = 0.74$ bits

[Besson, Bielova, Jensen CSF'2013]

Our hybrid monitor for quantitative information flow

- Combination of dynamic and static analysis
  - Soundness and relative precision theorems
- Automatic quantification of information leakage
  - Symbolic representation of tracker's knowledge at runtime
- Strong formal guarantees:
  - Over-approximates the leakage of one execution

All the theorems are proven in Coq: http://www.irisa.fr/celtique/ext/QIF/



Analyzing the stability of web browser fingerprints

WITH PATRICIO PALLADINO

Panopticlick

How Unique – and Trackable – Is Your Browser?

With private browsing, cookies are allowed
- **unique among** 2,911,686 browsers
- **21.47** bits of identifying information.

Without private browsing, cookies are blocked (after deleting all cookies) – US
- **unique among** 2,911,727 browsers
- **21.47** bits of identifying information.

Without private browsing, cookies are allowed – US
- **unique among** 2,911,733 browsers
- **21.47** bits of identifying information.

Panopticlick

How Unique — and Trackable — Is Your Browser?

With private browsing, cookies are allowed

- **unique among** 2,911,686 browsers
- **21.47** bits of identifying information.

High-level point:

Panopticlick **does not recognize me** as the same user!

Panopticlick **counts** the same browsers **multiple times**!

- **unique among** 2,911,733 browsers
- **21.47** bits of identifying information.

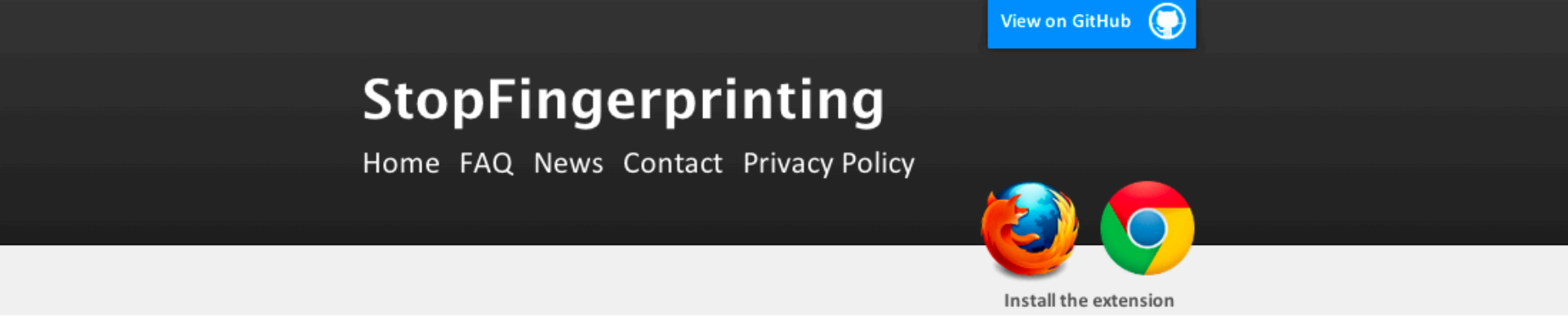
Panopticlick

How Unique — and Trackable — Is Your Browser?

Browser property	Source
User Agent (browser name and version, OS version, etc)	HTTP
	JS
HTTP_ACCEPT header	HTTP
Browser plugin details	JS
Time zone	JS
Screen size and color depth	JS
System fonts	Flash/JS
Cookies enabled?	HTTP
	JS
Supercookies test	JS

Real trackers would not use all these properties!

Some properties are not stable!



- Experiment setting
  - Browser extension for FireFox and Chrome
  - Currently ~200 users
  - Hourly reports to Inria server

- Collected information (Panopticlick ++)
  - HTTP data: userAgent, IP, HTTP headers
  - JavaScript data: plugins, fonts, date/time,…
  - Flash data: IP, camera, keyboard, fonts, language, …

- Install the extension to help us collect more data!
  - https://stopfingerprinting.inria.fr

View on GitHub

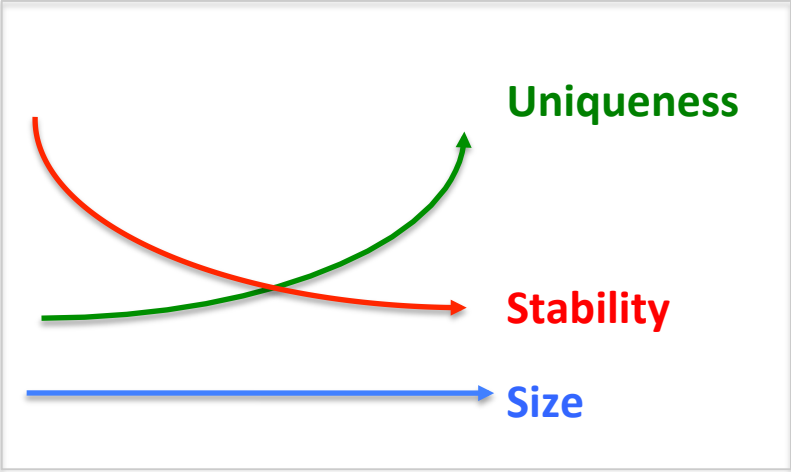
StopFingerprinting

Home FAQ News Contact Privacy Policy

Install the extension

What is the relation between fingerprints uniqueness, stability and size?

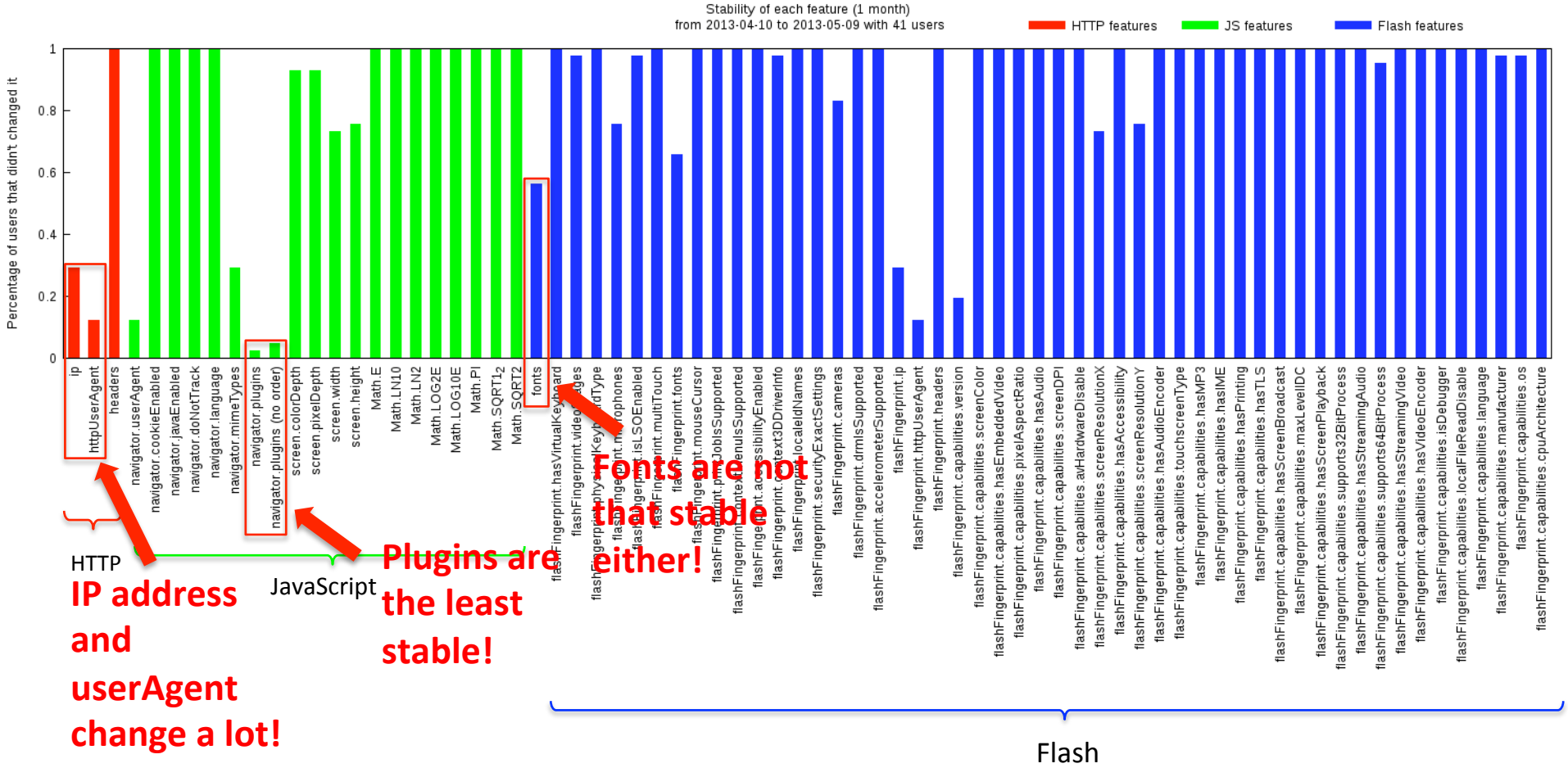
Our hypothesis:



- **Uniqueness**
  - How unique is a browser fingerprint in a long term?
- **Stability**
  - How stable are browser fingerprints?
- **Size**
  - Which subset of a fingerprint is actually useful for distinguishing the users?

Stability: by users

Percentage of users for whom a given browser feature was stable in a period of 1 month

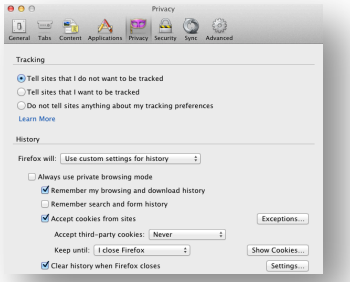


Conclusions

- **Web tracking**: stateful and stateless
  - cookies, storages, HTTP headers, device fingerprinting

- **Legal side**: EU ePrivacy directive and Do-Not-Track

- **Practical solutions**: none is 100% effective!
  - third-party cookies browser settings
  - browser extensions



- **Research solutions**
  - Information flow control against stateful tracking
  - Quantitative information flow against stateless tracking
  - Analysis of the stability of fingerprints:
    - http://stopfingerprinting.inria.fr