

What do Internet companies know about you?

Pierre Senellart



15 September 2016

CERRE Private Seminar – Data & Competition Policy

Dissecting the title

Internet companies

- “Internet companies” is the unfortunate established term
- Only talking about **Web-based services** (dot-com’s)
- Amazon, Google, Facebook. . . and smaller ones as well
- Internet is **not the same thing** as the Web!

Dissecting the title

Internet companies

- “Internet companies” is the unfortunate established term
- Only talking about **Web-based services** (dot-com’s)
- Amazon, Google, Facebook. . . and smaller ones as well
- Internet is **not the same thing** as the Web!

What do they know about you?

- What **can** they technically know?
- What do they typically **use** it for?
- How can a user **hide** this information?
- How can a **new company** have access to the **same data**?

Dissecting the title

Internet companies

- “Internet companies” is the unfortunate established term
- Only talking about **Web-based services** (dot-com’s)
- Amazon, Google, Facebook... and smaller ones as well
- Internet is **not the same thing** as the Web!

What do they know about you?

- What **can** they technically know?
- What do they typically **use** it for?
- How can a user **hide** this information?
- How can a **new company** have access to the **same data**?

Not discussing **legal**, **ethical**, or **economic** aspects!

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives.`

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives.`

My browser: Hum, who is cerre.eu?

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives`.

My browser: Hum, who is `cerre.eu`?

My ISP's DNS server: It's the machine `54.246.205.67`!

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives`.

My browser: Hum, who is `cerre.eu`?

My ISP's DNS server: It's the machine `54.246.205.67`!

My browser: Hello `54.246.205.67`, I would like to talk to your Web server.

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives`.

My browser: Hum, who is `cerre.eu`?

My ISP's DNS server: It's the machine `54.246.205.67`!

My browser: Hello `54.246.205.67`, I would like to talk to your Web server.

`54.246.205.67`: Sure thing!

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives`.

My browser: Hum, who is `cerre.eu`?

My ISP's DNS server: It's the machine `54.246.205.67`!

My browser: Hello `54.246.205.67`, I would like to talk to your Web server.

`54.246.205.67`: Sure thing!

My browser: Web server, can you give me the page `/mission-objectives`?

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives`.

My browser: Hum, who is `cerre.eu`?

My ISP's DNS server: It's the machine `54.246.205.67`!

My browser: Hello `54.246.205.67`, I would like to talk to your Web server.

`54.246.205.67:` Sure thing!

My browser: Web server, can you give me the page `/mission-objectives`?

My browser: Oh, and by the way, here is a bunch of other things about who I am
and what I like.

A primer on the Web 1/3

How does the Web work? Say I want to visit
`http://www.cerre.eu/mission-objectives`.

My browser: Hum, who is `cerre.eu`?

My ISP's DNS server: It's the machine `54.246.205.67`!

My browser: Hello `54.246.205.67`, I would like to talk to your Web server.

`54.246.205.67`: Sure thing!

My browser: Web server, can you give me the page `/mission-objectives`?

My browser: Oh, and by the way, here is a bunch of other things about who I am and what I like.

My browser: And here is the weird character string you asked me to remind you of the last time I visited you.

A primer on the Web 2/3

Web server on 54.246.205.67: Here is what you requested; if you really want to see the content in full, you should also load all these scripts and images on the same site, as well as this bunch of scripts from the following companies: Oracle, MaxCDN, the JQuery foundation, Google, and Twitter.

A primer on the Web 2/3

Web server on 54.246.205.67: Here is what you requested; if you really want to see the content in full, you should also load all these scripts and images on the same site, as well as this bunch of scripts from the following companies: Oracle, MaxCDN, the JQuery foundation, Google, and Twitter.

My browser: Hum ok, sure.

A primer on the Web 2/3

Web server on 54.246.205.67: Here is what you requested; if you really want to see the content in full, you should also load all these scripts and images on the same site, as well as this bunch of scripts from the following companies: Oracle, MaxCDN, the JQuery foundation, Google, and Twitter.

My browser: Hum ok, sure.

My browser: Hey Oracle, could you give me that script that `http://www.cerre.eu/mission-objectives` told me to ask you?

A primer on the Web 2/3

Web server on 54.246.205.67: Here is what you requested; if you really want to see the content in full, you should also load all these scripts and images on the same site, as well as this bunch of scripts from the following companies: Oracle, MaxCDN, the JQuery foundation, Google, and Twitter.

My browser: Hum ok, sure.

My browser: Hey Oracle, could you give me that script that `http://www.cerre.eu/mission-objectives` told me to ask you?

My browser: Oh, and by the way, here is a bunch of other things about who I am and what I like.

A primer on the Web 2/3

Web server on 54.246.205.67: Here is what you requested; if you really want to see the content in full, you should also load all these scripts and images on the same site, as well as this bunch of scripts from the following companies: Oracle, MaxCDN, the JQuery foundation, Google, and Twitter.

My browser: Hum ok, sure.

My browser: Hey Oracle, could you give me that script that `http://www.cerre.eu/mission-objectives` told me to ask you?

My browser: Oh, and by the way, here is a bunch of other things about who I am and what I like.

My browser: And here is the weird character string you asked me to remind you of the last time I visited you.

A primer on the Web 2/3

Web server on 54.246.205.67: Here is what you requested; if you really want to see the content in full, you should also load all these scripts and images on the same site, as well as this bunch of scripts from the following companies: Oracle, MaxCDN, the JQuery foundation, Google, and Twitter.

My browser: Hum ok, sure.

My browser: Hey Oracle, could you give me that script that `http://www.cerre.eu/mission-objectives` told me to ask you?

My browser: Oh, and by the way, here is a bunch of other things about who I am and what I like.

My browser: And here is the weird character string you asked me to remind you of the last time I visited you.

...

A primer on the Web 3/3

My browser: Finally! Got all the content. Now I need to execute all these scripts, I am pretty sure some of which will make me fetch content again from all over the internet

A primer on the Web 3/3

My browser: Finally! Got all the content. Now I need to execute all these scripts, I am pretty sure some of which will make me fetch content again from all over the internet

...

A primer on the Web 3/3

My browser: Finally! Got all the content. Now I need to execute all these scripts, I am pretty sure some of which will make me fetch content again from all over the internet

...

My browser: Hey, user, your page is ready! Had to do 74 requests and download 1.4 MB of data, so took me a few seconds. But the result is pretty cool, isn't it?

A primer on the Web 3/3

My browser: Finally! Got all the content. Now I need to execute all these scripts, I am pretty sure some of which will make me fetch content again from all over the internet

...

My browser: Hey, user, your page is ready! Had to do 74 requests and download 1.4 MB of data, so took me a few seconds. But the result is pretty cool, isn't it?

My browser: By the way, as long as you are on this page, I'll keep contacting Twitter every 30 seconds, they asked me to.

Different kinds of data

- Network-level data
- HTTP meta-information
- Browser scripting data
- Past interactions with the Web site
- Past interactions with a third-party Web site that uses a resource of mine
- Data provided by the user

Different kinds of data

- Network-level data
- HTTP meta-information
- Browser scripting data
- Past interactions with the Web site
- Past interactions with a third-party Web site that uses a resource of mine
- Data provided by the user

Don't forget: also technically easy for companies to **share** this information with each other (for a fee, with a reciprocity agreement, etc.)

Network-level data 1/3

What can they technically know?

- **IP address** (v4 or v6) of the computer sending the request
- From the IP address:
 - **Institution** the IP address belongs to (company, ISP, mobile phone operator)
 - Approximate **geolocation** of the IP address (somewhat precise at the country level, sometimes at the city level)
- **Network quality** information (latency and bandwidth of the communication)

Network-level data 2/3

What do they typically use it for?

- Proposing a different **default choice** of Web site (language, market) based on the geolocation
- Serving different content to **different markets** (e.g., copyrighted material with license only in specific countries)
- Optimizing **connection speeds** (serving a user from a server closer to her)
- Potentially, remembering **past interactions** (but very imprecise)

Network-level data 2/3

How can a user hide this information?

Hard. Route the traffic through a VPN, a proxy, Tor... but Web sites can use databases of IP addresses commonly used by these services. Always possible to route the traffic through another private computer, though.

How can a new company have access to the same data?

- IP addresses are readily available
- Databases mapping IP addresses to geolocations, companies, information about uses as VPNs, can easily be obtained, with various levels of quality (for free, for a fee, or semi-automatically built over time)
- Network quality information not as immediate, but can be obtained with a little work

HTTP meta-information 1/2

What can they technically know?

User-Agent identifies the **browser**, its version, its **operating system**, possibly some other information

Referer gives the URL of the Web page the browser is **coming from**

Accept-Language gives information on the user's **preferred languages** (typically, the language the OS is configured for)

Other headers (**Accept**, **Accept-Encoding**...) **indirectly** and partially **identify** the browser software; also possible through some analysis of protocol-level behavior (support of SPDY, of HTTP/2, of some cryptographic algorithms, of behavior w.r.t. pipelining, etc.)

HTTP meta-information 2/2

What do they typically use it for?

- Serving **different content** to different browsers (in particular desktop vs mobile sites)
- Serving content in the **appropriate language**
- Collect **statistics** about origin of the visit

How can a user hide this information?

The browser can be customized to **hide or spoof** the explicit data. Near impossible for indirect clues identifying the browser.

How can a new company have access to the same data?

Explicit information **readily available**. Identifying a browser through indirect clues (very) hard, but feasible with effort.

Browser scripting data 1/2

What can they technically know?

- **Timezone** of the user
- Characteristics (resolution, color) of the user's **screen**
- (If the user agrees) Fine **geolocation** data
- Indirectly, computer **performance** data
- Information about the browser **configuration** (e.g., are images or ads displayed? is third-party content loaded?)
- Potentially, every single information about how the user is **interacting** with browser windows displaying the Web site (but not other Web sites! “same-origin policy”):
 - Every mouse move, every key press, every click
 - Indirectly, every copy/paste operation
- Indirectly, device **fingerprinting** (e.g., through canvas fingerprinting)

Browser scripting data 2/2

What do they typically use it for?

- **Customize** a Web site appearance based on a user's configuration
- Run **user experience studies** for fine analysis of a user's interaction with the Web site
- Improve the **user experience** with more reactive Web pages

How can a user hide this information?

Only reliable possibility is to **block all scripts**, but will make many Web sites unusable.

How can a new company have access to the same data?

Readily available. Advanced tracking or fingerprinting requires important development effort.

Past interactions with the Web site 1/2

What can they technically know?

The browser will happily provide the same piece of information (a **cookie**) every time it visits **the same Web site**. Can be stored (the Web site's choice):

- for a given navigation session;
- or until some date (possibly very far out in the future).

Past interactions with the Web site 2/2

What do they typically use it for?

Remember **who the user is** and **previous interactions** it had with her. Critical for many features of Web sites: keeping a user logged in, shopping baskets, etc.

How can a user hide this information?

Possible to selectively **remove cookies**, or to destroy all cookies after a navigation session (e.g., **private mode**). Possible to block all cookies, but will break many Web sites.

How can a new company have access to the same data?

Readily available. Obviously, only valuable if the user has had **many interactions**.

Past interactions with a third-party Web site 1/2

What can they technically know?

If a **third-party** Web site requests a **resource** (image, stylesheet, script, media) **hosted** by a company's Web site, this company can have access to **all** previously mentioned information while visiting the third-party Web site (even client-side scripting if the resource is a script), including the **Referer** URL.

In particular, **cookies** are provided, so that the company's Web site can identify the user requesting the resource.

Past interactions with a third-party Web site 2/2

What do they typically use it for?

User tracking. Ad networks in particular heavily rely on this to build a profile of what pages a user visits.

How can a user hide this information?

Block third-party scripts using plugins. **Block third-party cookies.** Will break some functionalities.

How can a new company have access to the same data?

Very hard! Requires convincing thousands (or more) of third-party Web sites to include a link to your site. Have to provide a service (ads, analytics, social networking, CDN, widget) that people want to include on their site.

Data provided by the user 1/2

What can they technically know?

- Any data that they user needs to provide to interact with the service:
 - Email (serves as a **pseudo-identifier**)
 - User-chosen identifier (may be **reused** on other Web sites!)
 - Password (beware of **password reuse!**)
 - For e-commerce: credit card numbers, address, etc.
- Any other data readily **provided by the user** (birthdate, friends, job, interests, etc.)

Data provided by the user 2/2

What do they typically use it for?

Some is needed for **technical** reasons. Some can be used for **profiling**.

How can a user hide this information?

Provide **throwaway** email accounts and logins. Don't **reuse** passwords from a site to the next. Don't provide optional information.

How can a new company have access to the same data?

Easy... as long as you manage to **attract users**.

A paranoid's toolbox for browsing the Web

- An **open-source** and heavily configurable Web browser that **doesn't phone home** (Firefox, Chromium, Pale Moon)
- Activate the “**Do Not Track**” option (but not clear meaning for this option!)
- Plugins to **spooof** the User-Agent and Referer information
- Plugins such as Adblock Plus or uBlock Origin to **block third-party ads** (based on lists and heuristics)
- Plugins such as Ghostery or DoNotTrackMe to **block tracking cookies and fingerprinting code**
- Plugins such as NoScript to selectively **block scripts**
- Possible to block **all third-party cookies** altogether (but some features won't work)
- Possible to block **all client-side scripts** (many sites won't work!)
- Use Private Mode to have information (esp., cookies) **not retained** from one navigation session to the next

Merci.