

INF344 – Données du Web

Pistage par les entreprises du Web

Antoine Amarilli ; adapté d'une présentation par Pierre Senellart

9 mai 2018

On veut visiter `http://www.cerre.eu/mission-objectives`.

On veut visiter `http://www.cerre.eu/mission-objectives`.

Mon navigateur : Qui est `cerre.eu` ?

Rappels sur le Web 1/3

On veut visiter `http://www.cerre.eu/mission-objectives`.

Mon navigateur : Qui est `cerre.eu` ?

Serveur DNS de mon FÀI : C'est la machine `54.246.205.67` !

Rappels sur le Web 1/3

On veut visiter `http://www.cerre.eu/mission-objectives`.

Mon navigateur : Qui est `cerre.eu` ?

Serveur DNS de mon FÂI : C'est la machine `54.246.205.67` !

Mon navigateur : Bonjour `54.246.205.67`, je parle à ton serveur Web.
Est-ce que tu pourrais me donner la page
`/mission-objectives` ?

Rappels sur le Web 1/3

On veut visiter `http://www.cerre.eu/mission-objectives`.

Mon navigateur : Qui est `cerre.eu` ?

Serveur DNS de mon FÂI : C'est la machine `54.246.205.67` !

Mon navigateur : Bonjour `54.246.205.67`, je parle à ton serveur Web.
Est-ce que tu pourrais me donner la page
`/mission-objectives` ?

Mon navigateur : Au fait, voici d'autres informations sur qui je suis

Rappels sur le Web 1/3

On veut visiter `http://www.cerre.eu/mission-objectives`.

Mon navigateur : Qui est `cerre.eu` ?

Serveur DNS de mon FÂI : C'est la machine `54.246.205.67` !

Mon navigateur : Bonjour `54.246.205.67`, je parle à ton serveur Web.
Est-ce que tu pourrais me donner la page
`/mission-objectives` ?

Mon navigateur : Au fait, voici d'autres informations sur qui je suis

Mon navigateur : Et voici les données (cookies) que tu m'as demandé
de t'envoyer à chaque fois que je viens

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Mon navigateur : Cette page appelle d'autres ressources :

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Mon navigateur : Cette page appelle d'autres ressources :

- D'autres scripts et images sur le même site

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Mon navigateur : Cette page appelle d'autres ressources :

- D'autres scripts et images sur le même site
- Des scripts de Oracle, MaxCDN, JQuery foundation, Google, Twitter

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Mon navigateur : Cette page appelle d'autres ressources :

- D'autres scripts et images sur le même site
- Des scripts de Oracle, MaxCDN, JQuery foundation, Google, Twitter

Mon navigateur : Je vais envoyer une requête à tous ces gens-là

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Mon navigateur : Cette page appelle d'autres ressources :

- D'autres scripts et images sur le même site
- Des scripts de Oracle, MaxCDN, JQuery foundation, Google, Twitter

Mon navigateur : Je vais envoyer une requête à tous ces gens-là

...

Mon navigateur : Maintenant que j'ai tout le contenu, il ne reste plus qu'à exécuter tous les scripts; ils vont sans doute me demander d'accéder à d'autres ressources

Mon navigateur : Maintenant que j'ai tout le contenu, il ne reste plus qu'à exécuter tous les scripts; ils vont sans doute me demander d'accéder à d'autres ressources

...

Mon navigateur : Maintenant que j'ai tout le contenu, il ne reste plus qu'à exécuter tous les scripts; ils vont sans doute me demander d'accéder à d'autres ressources

...

Mon navigateur : Cher utilisateur, la page est prête! Ça m'a pris 74 requêtes et m'a fait télécharger 1.4 MB de données

Mon navigateur : Maintenant que j'ai tout le contenu, il ne reste plus qu'à exécuter tous les scripts; ils vont sans doute me demander d'accéder à d'autres ressources

...

Mon navigateur : Cher utilisateur, la page est prête! Ça m'a pris 74 requêtes et m'a fait télécharger 1.4 MB de données

Mon navigateur : Au fait, tant que tu es sur cette page, je vais contacter Twitter toutes les 30 secondes, comme elle me demande de le faire

Tous les sites ne sont pas comme ça

- C'est **pire** sur les sites Web avec de la pub (sites d'actualité, blogs, etc.) ou qui ont des accords avec les régies publicitaires (sites de vente en ligne)

Tous les sites ne sont pas comme ça

- C'est **pire** sur les sites Web avec de la pub (sites d'actualité, blogs, etc.) ou qui ont des accords avec les régies publicitaires (sites de vente en ligne)
- C'est **mieux** sur certains sites "old-school" :

Tous les sites ne sont pas comme ça

- C'est **pire** sur les sites Web avec de la pub (sites d'actualité, blogs, etc.) ou qui ont des accords avec les régies publicitaires (sites de vente en ligne)
- C'est **mieux** sur certains sites "old-school" :
ec.europa.eu : pas de références externes

Tous les sites ne sont pas comme ça

- C'est **pire** sur les sites Web avec de la pub (sites d'actualité, blogs, etc.) ou qui ont des accords avec les régies publicitaires (sites de vente en ligne)
- C'est **mieux** sur certains sites "old-school" :
 - ec.europa.eu** : pas de références externes
 - europa.eu** : une seule référence externe (CloudFlare)

Tous les sites ne sont pas comme ça

- C'est **pire** sur les sites Web avec de la pub (sites d'actualité, blogs, etc.) ou qui ont des accords avec les régies publicitaires (sites de vente en ligne)
- C'est **mieux** sur certains sites "old-school" :
 - ec.europa.eu** : pas de références externes
 - europa.eu** : une seule référence externe (CloudFlare)
 - europarl.europa.eu** : références à Google, Yahoo! seulement

- Souvent, le serveur Web loggue **toutes** les requêtes traitées!
- NCSA Common log format :
 - **Adresse IP** du client
 - Identité **ident** (obsolète)
 - Identité **authentification HTTP** (rarement utile)
 - **Date** et **heure** de fin de traitement
 - Première ligne de la **requête HTTP**
 - **Code d'état** retourné au client
 - **Taille** (octets) de la réponse
- Ajouts fréquents :
 - **User-Agent** pour identifier le client
 - **Referer** pour connaître la page précédente

```
208.115.113.88 - - [22/Jan/2012:06:27:00 +0100]
```

```
"GET /robots.txt HTTP/1.1" 404 266 "-"
```

```
"Mozilla/5.0 (compatible; Ezooms/1.0; ezooms.bot@gmail.com)"
```

Utilisation des logs

- Pages les **plus visitées**
- Chemin sur le site, temps passé sur les pages, entrées, sorties (mais avec les onglets, difficile)
- Distinguer **humains et robots** (pas de garanties !)
- Emplacement **géographique** des visiteurs
- **Mots-clés** saisis dans les moteurs de recherche
- **Liens** vers son site
- Nombre d'abonnés à un flux RSS dans **User-Agent**
- **Spam** dans **Referer**
- Et plus?
 - Logiciels pour exploiter directement les logs du serveur
 - Tracking en PHP
 - Tracking tiers en JavaScript (Google Analytics)
 - Tracking possible du pointeur de la souris en JavaScript, etc.
 - Google Search Console

Types de données

- Données fournies par l'utilisateur
- Données réseau
- En-têtes HTTP
- Données de scripts
- Visites précédentes
- Visites précédentes sur des sites dépendants

Types de données

- Données fournies par l'utilisateur
- Données réseau
- En-têtes HTTP
- Données de scripts
- Visites précédentes
- Visites précédentes sur des sites dépendants

Ces données sont souvent **partagées** d'une entreprise à l'autre

Qu'est-ce qu'ils peuvent savoir ?

- Adresse email (**pseudo-identifiant**)
- Nom d'utilisateur (peut-être **réutilisé** ailleurs, donc croisements)
- Mot de passe (attention à la **réutilisation** !)
- Vente en ligne : numéro de carte, adresse postale, etc.
- Autres données : date de naissance, amis, emploi, centres d'intérêt, etc.

Que vont-ils faire avec?

Utile pour des **raisons techniques**, mais aussi pour du **profilage**

Que vont-ils faire avec?

Utile pour des **raisons techniques**, mais aussi pour du **profilage**

Comment cacher ces informations?

- Utiliser des comptes **jetable**s : `mailinator.com`, `bugmenot.com`
- Fournir le moins d'information possible, ou mentir

Qu'est-ce qu'ils peuvent savoir ?

- **Adresse IP** du client
- **Institution** de l'IP (entreprise, FÀI, opérateur mobile)
- **Géolocalisation** approximative (pays, voire ville)
- **Qualité réseau** latence et bande passante

Que vont-ils faire avec ?

- Rediriger vers un **site local** (langue, marché) suivant la géolocalisation
- Géoblocage : pour respecter le copyright, pour la censure, pour le droit à l'oubli
- Améliorer la **qualité de service** en utilisant un serveur proche de l'utilisateur

Comment cacher ces informations ?

- Router le trafic via **Tor** torproject.org mais les adresses de sortie sont publiques (donc on sait que vous utilisez Tor, et on peut vous bloquer)
- Router le trafic via un **ordinateur personnel**, un **VPN**, etc.

Qu'est-ce qu'ils peuvent savoir?

User-Agent Identifie le **navigateur**, la version, le **système d'exploitation**, voire plus

Qu'est-ce qu'ils peuvent savoir?

User-Agent Identifie le **navigateur**, la version, le **système d'exploitation**, voire plus

Referer Indique de quel URL on **vient**

Qu'est-ce qu'ils peuvent savoir?

User-Agent Identifie le **navigateur**, la version, le **système d'exploitation**, voire plus

Referer Indique de quel URL on **vient**

Accept-Language Indique la **langue préférée** de l'utilisateur

Qu'est-ce qu'ils peuvent savoir ?

User-Agent Identifie le **navigateur**, la version, le **système d'exploitation**, voire plus

Referer Indique de quel URL on **vient**

Accept-Language Indique la **langue préférée** de l'utilisateur

Fingerprinting On peut deviner des choses sur le navigateur à partir des autres en-têtes, le support de différentes technologies, algorithmes de chiffrement, etc.

Que vont-ils faire avec ?

- Servir du **contenu différent** aux différents navigateurs (ordinateur/mobile, compatibilité, etc.)
- Choix de la **langue**
- Utile pour la **mesure d'audience**

Que vont-ils faire avec ?

- Servir du **contenu différent** aux différents navigateurs (ordinateur/mobile, compatibilité, etc.)
- Choix de la **langue**
- Utile pour la **mesure d'audience**

Comment cacher ces informations ?

- Pour User-Agent, Referer, Accept-Language : configurable
- Pour le comportement intrinsèque du navigateur : irréaliste

Qu'est-ce qu'ils peuvent savoir?

- **Fuseau horaire** de l'utilisateur
- Résolution, couleurs de l'**écran**
- **Géolocalisation GPS** sur mobile (consentement préalable)
- **Configuration** du navigateur (blocage des contenus tiers, etc.)
- **Événements** sur le site (ou autres fenêtres avec le même domaine) :
 - Déplacements de souris, clics
 - Frappes clavier, copier-coller
- **Fingerprinting** indirect (e.g., canvas, polices installées, etc.)
panopticlick.eff.org

Que vont-ils faire avec ?

- Améliorer l'**expérience utilisateur** avec des contenus dynamiques
- **Modifier** le site Web suivant la configuration (CSS, etc.)
- Mesures pour améliorer l'**expérience utilisateur** : Heatmaps, A/B testing, etc.

Que vont-ils faire avec ?

- Améliorer l'**expérience utilisateur** avec des contenus dynamiques
- **Modifier** le site Web suivant la configuration (CSS, etc.)
- Mesures pour améliorer l'**expérience utilisateur** : Heatmaps, A/B testing, etc.

Comment cacher ces informations ?

- **Blocage des scripts** (uMatrix, etc.), mais beaucoup de sites deviennent inutilisables

Qu'est-ce qu'ils peuvent savoir?

Les sites peuvent faire stocker au navigateur des informations qu'il devra renvoyer à chaque visite ultérieure, pour une durée possiblement très longue

Que vont-ils faire avec ?

- Se souvenir de **qui est l'utilisateur** et des précédentes interactions avec l'utilisateur
- Nécessaire pour de nombreuses fonctionnalités : maintenir une session ouverte, se souvenir du panier de l'utilisateur, etc.

Que vont-ils faire avec ?

- Se souvenir de **qui est l'utilisateur** et des précédentes interactions avec l'utilisateur
- Nécessaire pour de nombreuses fonctionnalités : maintenir une session ouverte, se souvenir du panier de l'utilisateur, etc.

Comment cacher ces informations ?

- **Supprimer** des cookies, manuellement, ou à la fin d'une session (mode incognito)
- **Refuser** les cookies, mais beaucoup de sites Web ne fonctionneront plus

Qu'est-ce qu'ils peuvent savoir ?

Si un site **dépend** de ressources hébergées par une entreprise (images, feuille de style, scripts, médias), cette entreprise peut faire son propre pistage

- Y compris en utilisant **Referer**
- Y compris avec ses propres **cookies**

Que vont-ils faire avec?

C'est comme ça que les **régies publicitaires** peuvent pister la navigation de l'utilisateur d'un site à l'autre

Que vont-ils faire avec?

C'est comme ça que les **régies publicitaires** peuvent pister la navigation de l'utilisateur d'un site à l'autre

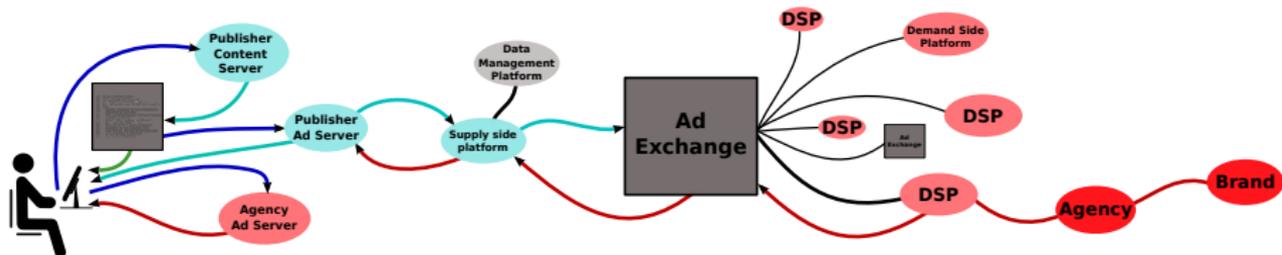
Comment cacher ces informations?

Bloquer les **scripts tierce-partie**, ou les **cookies tierce-partie**, mais certains sites Web ne fonctionneront plus

Comment naviguer sur le Web sans laisser de traces

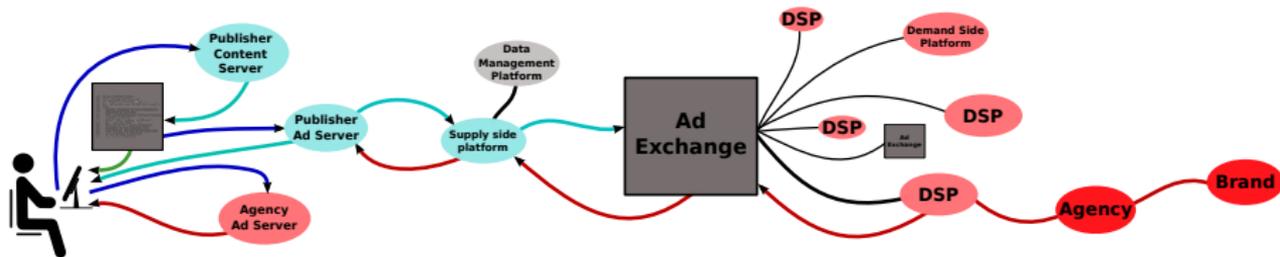
- Utiliser un navigateur **libre** et configurable qui respecte votre vie privée, par exemple Tor Browser, Firefox, Chromium, Pale Moon, etc.
- **Cacher** son IP avec Tor (directement dans Tor Browser)
- Activer l'option "**Do Not Track**" ?
- **Cacher** ou falsifier l'information User-Agent et Referer
- Bloquer les **publicités** avec uBlock Origin
- Bloquer sélectivement les **scripts** avec uMatrix (mais casse de nombreux sites Web)
- Bloquer les **cookies** ou les **cookies tierce partie** (mais même problème)
- Utiliser le **mode incognito** pour que les informations côté client (cookies, historique) ne soient **pas sauvegardées** d'une session à l'autre

Les tierce-parties des tierce-parties



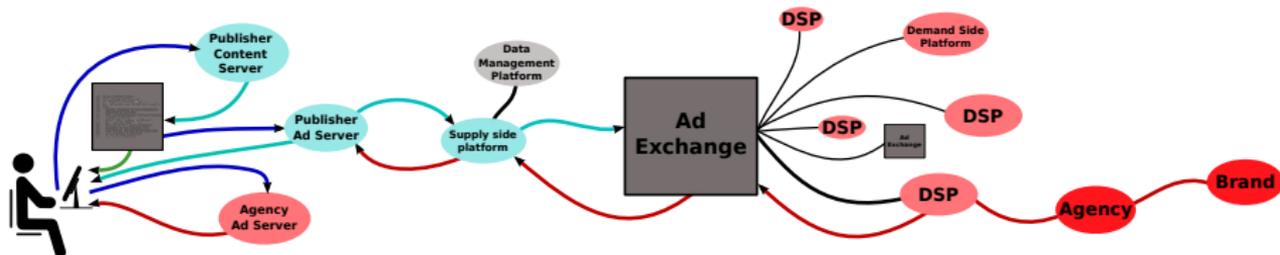
- L'annonceur du site Web et sa plateforme côté serveur (supply-side platform, SSP), par exemple Google DFP, Rubicon, peuvent **identifier** l'utilisateur avec des cookies...

Les tierce-parties des tierce-parties



- L'annonceur du site Web et sa plateforme côté serveur (supply-side platform, SSP), par exemple Google DFP, Rubicon, peuvent **identifier** l'utilisateur avec des cookies...
- ... mais les annonceurs sur la plateforme (demand-side platform, DSP), par exemple AppNexus, Criteo, ne peuvent pas, parce qu'ils n'interagissent pas **directement** avec l'utilisateur (jusqu'à ce que la publicité soit affichée)

Les tierce-parties des tierce-parties



- L'annonceur du site Web et sa plateforme côté serveur (supply-side platform, SSP), par exemple Google DFP, Rubicon, peuvent **identifier** l'utilisateur avec des cookies...
- ... mais les annonceurs sur la plateforme (demand-side platform, DSP), par exemple AppNexus, Criteo, ne peuvent pas, parce qu'ils n'interagissent pas **directement** avec l'utilisateur (jusqu'à ce que la publicité soit affichée)
- Pour résoudre ce "problème", la SSP partage ses informations de cookie avec la DSP, pour réidentifier l'utilisateur avec de **l'appariement de cookies**.

Exemple : Google i

Qu'est-ce que Google sait de vous ?

- Toute information que vous lui avez fourni, de votre plein gré ou non (carte de crédit pour Google Play, nom pour Google Plus, historique GPS pour Google Location Services, recherches pour Google Search, etc.)
- Toute **interaction passée** que vous avez eu avec un site de Google (Search, Maps, Mail, Drive, etc.) sauf si vous n'étiez pas connecté et n'utilisiez pas vos cookies habituels (e.g., mode incognito) et utilisiez une adresse IP différente
- Toute visite sur un site Web dépendant d'un service **hébergé** par Google (Google Analytics, Google Hosted Libraries, Google Fonts, Google AdSense...) sauf si vous refusez les cookies tierce-partie

Exemple : Google ii

- Toute visite sur un site Web qui inclut des **publicités** servies par une chaîne qui inclut Google Doubleclick (c'est-à-dire la majorité des sites avec publicités), sauf si vous refusez les cookies tierce-partie

Google n'utilise **pas forcément** toutes ces capacités, mais ils en ont la possibilité technique.

Protection légale des données personnelles

- 1974** Projet **SAFARI** pour lier les fichiers de l'administration
- 1978** Loi informatique et libertés, création de la **CNIL**
- 1998** Entrée en vigueur de la **directive européenne 95/46/CE**
 - Accord **Safe Harbor** avec les US
- 2004** Transposition en droit français de 95/46/CE
 - **Correspondants informatiques et libertés** (CIL)
- 2014** Arrêt de la CJUE pour le **droit à l'oubli**
- 2016** Accord **Privacy Shield** pour remplacer Safe Harbor
- 2017** Fichier **TES** (titres électroniques sécurisés)
- 2018** Entrée en vigueur du **RGPD**

Online advertising schema CC-BY-SA John Nagle, see

<https://commons.wikimedia.org/wiki/File:Adsgivingfull.svg>