INF344 – Données du Web
Pistage par les entreprises du Web

Antoine Amarilli ; adapté d'une présentation par Pierre Senellart

2 mai 2017

Rappels sur le Web 1/3

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

Rappels sur le Web 1/3

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 FÀI : 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 FÀI : 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 FÀI : 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

Rappels sur le Web 2/3

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Rappels sur le Web 2/3

Serveur Web de 54.246.205.67 : Voici la page Web demandée

Mon navigateur : Cette page appelle d'autres ressources :

Rappels sur le Web 2/3

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

Rappels sur le Web 2/3

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

Rappels sur le Web 2/3

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

Rappels sur le Web 2/3

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

...

Rappels sur le Web 3/3

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

Rappels sur le Web 3/3

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

…

Rappels sur le Web 3/3

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

Rappels sur le Web 3/3

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

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

Données fournies par l'utilisateur 1/2

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.

Données fournies par l'utilisateur 2/2

Que vont-ils faire avec ?

Utile pour des raisons techniques, mais aussi pour du profilage

Données fournies par l'utilisateur 2/2

Que vont-ils faire avec ?

Utile pour des raisons techniques, mais aussi pour du profilage

Comment cacher ces informations ?

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

Données réseau 1/3

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

Données réseau 2/3

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

Données réseau 3/3

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.

En-têtes HTTP 1/2

Qu'est-ce qu'ils peuvent savoir ?

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

En-têtes HTTP 1/2

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

En-têtes HTTP 1/2

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

En-têtes HTTP 1/2

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 technologie, algorithmes de chiffrement, etc.

En-têtes HTTP 2/2

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

En-têtes HTTP 2/2

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

Données de scripts 1/2

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`

Données de scripts 2/2

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.

Données de scripts 2/2

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

Visites précédentes 1/2

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

Visites précédentes 2/2

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.

Visites précédentes 2/2

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

Visites précédentes sur des sites dépendants 1/2

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

Visites précédentes sur des sites dépendants 2/2

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

Visites précédentes sur des sites dépendants 2/2

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

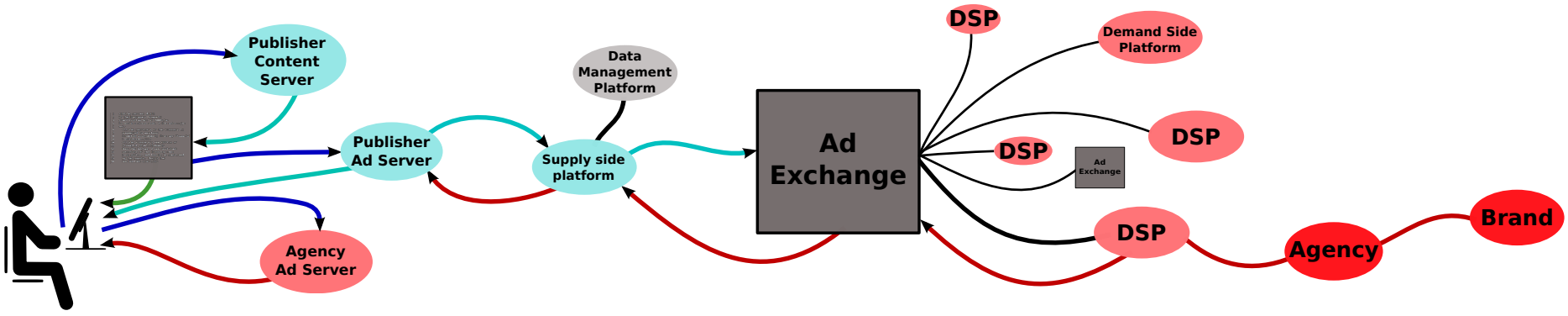
Comment naviguer sur le Web sans laisser de traces

- 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

Crédits

Online advertising schema CC-BY-SA John Nagle, see
`https://commons.wikimedia.org/wiki/File:Adservingfull.svg`

Beyond third-party Web sites : third parties of third parties I



- The publisher ad server and its supply-side platform (Google DFP, Rubicon) can identify the user with cookies…

Beyond third-party Web sites : third parties of third parties II

- Huge leak of information ! Mechanisms obscure, `https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_bashir.pdf`

Use case : Google I

What can Google know about you ?

- Every information you willingly or semi-willingly provided the company (credit card for Google Play, real name for Google Plus, full GPS history for Google Locations services on Android, etc.)
- Every past interaction you had with a Web site owned by Google (Search, Maps, Mail, Drive, etc.) unless you were not logged in and cookies were not shared (e.g., private mode)
- Every visit of a Web site that uses one of Google's hosted services (Google Analytics, Google Hosted Libraries, Google Fonts, Google AdSense…) unless third-party cookies are not shared

Use case : Google II

- Every visit of a Web site that includes advertisements served by a chain involving Google Doubleclick (the vast majority of Web sites with ads) unless third-party cookies are not shared

Not necessarily making full use of this, but the technical potential is there.