

PROJET INDIVIDUEL GYMINF

Trace(r)s on the web

Qui nous observe sur le web et comment

Julia Rebstein

supervisé par Linus Gasser, C4DT

Juillet 2023

Sommaire

1. Introduction	4
1.1. Révision des études existantes	5
1.2. Buts du travail de recherche	7
2. Côté technique	8
2.1. Cookie : Historique	8
2.2. Cookie : Fonctionnement et restrictions	8
2.3. HTTP : requête et réponse (TutorialsPoint Contributor, 2023)	10
2.3.1. Requête HTTP	10
2.3.2. Réponse HTTP	12
2.4. Utilité des cookies	13
2.5. La classification des cookies	14
2.5.1. First-Party Cookies	14
2.5.2. Third-Party Cookies	14
2.5.3. Second-Party Cookies	15
2.6. Fingerprinting, cookie syncing et autres techniques de traçage	15
2.6.1. Fingerprinting	15
2.6.2. Pixel	16
2.6.3. Image	16
2.6.4. Cookie syncing	16
2.6.5. Les acteurs publicitaires du web	17
3. Présentation de l'étude	20
3.1. Participants et récolte des données	20
3.2. Terminologie utilisée dans les extensions Lightbeam et Thunderbeam	22
3.3. Lightbeam - Thunderbeam	23
3.3.1. Situation 1	26
3.3.2. Situation 2	28
3.3.3. Situation 3	29
3.3.4. Situation 4	30
3.3.5. Situation 5	31
3.3.6. Situation 6	32
4. Résultats et Analyse de l'étude	34
4.1. Nombre de sites visités vs. nombre de sites tiers	34
4.2. Sites visités par plusieurs participants	40
4.3. Compagnies publicitaires et sites tiers majoritaires	42
4.3.1. Liens entre les sites primaires via les compagnies publicitaires	44
4.3.2. Sites tiers contactés par un grand nombre de sites primaires	46
5. Études sur sites spécifiques	48
5.1. Introduction à OpenWPM	48
5.2. Terminologie utilisée dans OpenWPM	49
5.3. Études de cas	49
5.3.1. Surf sur le site lematin.ch à travers le temps	50
5.3.2. Surf sur différents sites	51

5.4. Comparaison avec Lightbeam	54
6. Proposition de discussion avec élèves	57
6.1. La sphère privée c'est quoi ? (15 min)	57
6.1.1. Pourquoi se protéger ?	58
6.2. Les cookies (5 min)	58
6.3. Lightbeam (25 min)	58
6.4. Dépouillement des résultats (35 min)	59
6.5. Comment se protéger ? (10 min)	60
7. Conclusion	61

1. Introduction

Depuis quelques années déjà, les cookies ont envahi nos pages internet et sont partout autour de nous. Il en va de même pour la publicité. Impossible de surfer "tranquillement" sans avoir constamment des sollicitations et des suggestions pour tel ou tel objet, très probablement en lien avec l'un de nos centres d'intérêt.

Qu'est-ce qu'un cookie ? Y en a-t-il des bons et des mauvais, des gentils et des méchants ? Sont-ils vraiment nécessaires ? Quel est leur lien avec la publicité que l'on reçoit ? Faut-il les accepter ou les bannir ? Autant de questions auxquelles il est parfois difficile, voire impossible de répondre pour un néophyte. Ce travail a pour vocation d'éclairer quiconque le voudra afin de lui permettre de décider en toute connaissance de cause ce qu'il veut faire lorsqu'on lui pose la fameuse question " Acceptez-vous les cookies ? " .

Afin que ce travail ne soit pas une simple répétition de ce qui a été publié sur le sujet, j'ai suivi le trafic internet de 10 personnes *volontaires* pendant une semaine. J'ai pu le faire grâce à une extension proposée sur Firefox (Lightbeam) et aussi sur Google Chrome (Thunderbeam). Deux navigateurs largement utilisés par la population.

L'intérêt de cette étude face à toutes celles qui ont déjà été publiées est qu'elle fait intervenir des personnes réelles, qui surfent sur des sites internet, non pas choisis au hasard ou selon une popularité quelconque, mais pour leur propre intérêt. Les utilisateurs choisissent de s'identifier ou non, de procéder à des achats ou non, d'accepter les cookies ou non etc... En un mot, ils naviguent sur internet comme tout un chacun. À des fins statistiques, les études publiées font, au contraire, appel à des robots qui surfent sur des milliers de sites, et ce, 24 heures sur 24. Les sites visités ont été choisis en fonction de leur position dans des listes de sites populaires, comme par exemple sur le site Similarweb (Similarweb LTD, 2023). Les résultats qui en ressortent sont dès lors statistiquement fiables mais ne relèvent pas toujours d'une réalité qui nous est propre, en ce sens que nous ne restons pas sur internet sans interruption pendant des semaines, voire des mois, sans dormir ni manger, travailler ou toutes autres activités "humaines". De plus, les personnes choisies habitent en Suisse et se promènent donc également sur des sites suisses et pas exclusivement sur des sites à énormes audiences, représentés dans les études à large échelle.

Le nom des participants ne sera jamais dévoilé, ainsi, ni les étudiants ni ceux lisant ce rapport, ne pourront faire des liens entre une personne et les résultats présentés.

Ce rapport commence par une revue de quelques études réalisées sur les traces laissées sur internet et la manière dont nous sommes ciblés. Le chapitre 2 donne des explications détaillées sur les différents types de cookies existants, leur utilité ainsi que leur fonctionnement, incluant les protocoles d'échange HTTP (HyperText Transfer Protocol). Les différentes manières d'utiliser les cookies à des fins de traçage sont également exposées dans ce chapitre. Le chapitre 3 commence par la description détaillée du déroulement de l'étude puis du fonctionnement des extensions utilisées, à savoir Lightbeam et Thunderbeam. Le détail de chaque situation est alors illustré par des exemples spécifiques. Le chapitre 4 présente les résultats de l'étude et leurs analyses. Un autre type de recherche, décrit dans le chapitre 5, est effectué avec un autre programme, OpenWPM, qui va plus loin que les deux extensions en termes d'enregistrement de données. Ce deuxième paquet de

résultats permet une comparaison entre Lightbeam/Thunderbeam et OpenWPM. Finalement, dans le chapitre 6, je propose une trame de cours à donner à des élèves du gymnase. Le rapport se termine par quelques pistes de réflexion et une conclusion.

1.1. Révision des études existantes

Au moment de démarrer le projet, j'ai pris connaissance de quelques études réalisées sur le même sujet et j'en livre ici un résumé. Je tiens toutefois à préciser que je n'ai pas fait une révision extensive de toutes les études existantes mais j'en ai sélectionné quelques-unes, en fonction de certains mots-clés, tels que "tracking", "retargeted ads", "cookies". Elles m'ont aidée à comprendre le fonctionnement des cookies et des diverses techniques utilisées pour le traçage.

Bashir et al. (2016) présentent une manière de tracer les échanges d'information entre acteurs de la publicité sur le web. Ils utilisent pour cela la publicité ciblée qu'un utilisateur *lambda* voit lorsqu'il fait une recherche sur un site précis pour un objet spécifique. Contrairement à d'autres études qui cherchent à comprendre les mécanismes derrière les échanges d'information pour pouvoir proposer de la publicité ciblée, Bashir et al. se concentrent sur la publicité elle-même. Ils utilisent une version améliorée du navigateur Chromium qui leur permet d'enregistrer l'origine des requêtes d'images présentes sur les sites, même si celles-ci sont générées dynamiquement. Ils arrivent ainsi à détecter tous les liens existants entre les acteurs du web. Ils ont ensuite créé des personnages types (*personas*) imitant des personnes réelles, allant sur des sites de e-commerce. Ces *personas* ont été créés de telle sorte à couvrir un grand nombre de sites d'e-commerce différents, basés essentiellement aux États-Unis d'Amérique. Les chercheurs ont ainsi pu collecter des images publicitaires et s'assurer qu'elles étaient ciblées en fonction des intérêts des *personas*. Dans un deuxième temps, ils ont fait appel à des travailleurs réels pour classer ces images en tant que publicité ciblée ou publicité aléatoire. Lorsque la publicité est issue d'une redirection, donc d'un ciblage, les chercheurs ont pu observer les liens entre acteurs du web à l'aide des chaînes de requêtes et tout cela sans connaître le mécanisme utilisé pour le reciblage. Ils s'affranchissent ainsi de certains problèmes empêchant le suivi des cookies, tels que le hachage des identifiants.

Les résultats de l'étude montrent quels acteurs jouent quel rôle dans le reciblage publicitaire et aussi qui est plus présent que d'autre. Google ressort de manière très présente.

L'étude de Imane Fouad et al. (2020) se concentre sur le traçage à l'aide des pixels invisibles (image de taille 1 x 1). Ils précisent que les sites traqueurs sont souvent détectés, lors de recherches sur le sujet, ou bloqués, par des outils de navigateur, grâce à l'utilisation de filtres contenant une liste des sites traqueurs. Le problème de ces listes est qu'elles doivent être maintenues à jour perpétuellement et il est courant qu'elles ratent les nouveaux sites traqueurs. Ces derniers utilisent très fréquemment des pixels, dissimulés dans les pages web, pour envoyer des informations à leurs propres serveurs et obtenir des informations sur l'utilisateur. Imane Fouad et al pensent que la détection des pixels est un bon moyen de tracer les traqueurs.

Cette recherche a été menée en France en 2019 sur 10'000 sites web. Les auteurs utilisent l'outil OpenWPM pour enregistrer les requêtes et réponses HTTP lors du chargement des pages web afin de détecter les échanges de communication entre le navigateur et le serveur des pages chargées. Ils peuvent ainsi élargir le champ de leur recherche et observer

plusieurs manières de tracer les utilisateurs. Les résultats montrent que 91,92% des sites observés comportent au moins un moyen de traquer l'utilisateur basé sur l'utilisation de cookie. Ils dénombrent cinq types de traçage et les pixels invisibles arrivent en 2e position, juste derrière l'utilisation de javascript.

Leur étude a également permis de mettre en évidence des compagnies qui s'échangent des informations au sujet des utilisateurs, ainsi que la pratique utilisée.

Les résultats montrent que, en fonction de la liste de filtre utilisée, env. 15% des requêtes ne sont pas bloquées alors qu'elles devraient l'être.

Maaz Bin Musa et Rishab Nithyanand (2022) utilisent les publicités comportementales afin d'identifier les relations de partage de données entre les traqueurs et les annonceurs. Ils utilisent pour cela une méthode qu'ils nomment ATOM (pour Ad-network TOMography), en ce sens qu'elle étudie les images diffusées sur le web. Ils postulent qu'en bloquant systématiquement les traqueurs connus et en observant les changements que cela induit sur la publicité présentée, ils pourront montrer les liens entre annonceurs et traqueurs. Ils parviennent ainsi à détecter plus d'échanges entre partenaires du web que les études précédentes qui ne se basaient que sur un type de mécanisme observable côté client, comme par exemple le "cookie syncing" ou synchronisation de cookies, dont j'explique le principe de fonctionnement dans la section 2.6.4.

ATOM se veut un outil d'aide et de soutien à la réglementation en lien avec la protection de la vie privée. En effet, les annonceurs disposent de beaucoup plus de moyens financiers pour développer de nouvelles technologies que n'en ont les régulateurs pour détecter ces violations. ATOM contribue à fournir un cadre général permettant de recueillir des preuves statistiquement solides de violations potentielles des réglementations en matière de divulgation du partage de données. Il peut servir de base pour inciter à des enquêtes approfondies.

Les auteurs admettent toutefois que leur outil n'est pas totalement abouti et devra encore faire l'objet d'amélioration, notamment en incluant les vidéos ou les GIF.

Papadopoulos et al. (2019) s'intéressent au partage des cookies entre les différents acteurs du web, aussi appelé synchronisation des cookies. Ils se basent sur des données récoltées auprès de 850 utilisateurs de téléphones mobiles durant une année. Leur étude met en lumière qu'en moyenne, l'identité d'un utilisateur est partagée avec 3,5 domaines et que la synchronisation des cookies augmente le nombre de domaines traquant les utilisateurs par un facteur 6,75. Ils catégorisent les traqueurs en cinq domaines d'intérêt, et sans surprise les domaines qui ressortent le plus sont les sociétés de publicité, suivi par les sociétés actives dans l'analyse statistique puis par les média sociaux. Ils remarquent également que trois compagnies obtiennent des informations sur 30% et plus d'utilisateurs grâce au partage de cookies, tandis que 14 entreprises en obtiennent sur 20% des utilisateurs du panel. Ceci montre que, malgré le grand nombre de compagnies publicitaires existantes, le partage des cookies a lieu entre quelques compagnies majeures, dont Magnite (propriétaire de Rubiconproject), 360yield et openx.

Dans une autre étude, la première du genre selon ses auteurs, Imane Fouad et al (2021) mettent en évidence la réapparition de cookies effacés rendue possible grâce à une nouvelle manière de traquer les utilisateurs, le *fingerprinting* (technique décrite dans la section 2.6.1 de ce document) combiné avec les techniques habituelles de traçage. Le fingerprinting est un type de traçage dit sans état (*stateless*) en ce sens que ce ne sont pas

des données concernant l'utilisateur qui sont enregistrées, comme dans le cas des cookies, mais des données concernant sa machine, son navigateur ou autre. Cette étude a été menée en mars 2021 à l'aide de l'outil OpenWPM que les auteurs ont fait tourner sur deux machines distinctes pour enregistrer différentes données telles que les URL et en-tête des requêtes et réponses HTTP. Ils ont alors configuré OpenWPM pour surfer sur 30'000 sites de manière séquentielle et distincte sur les deux machines afin de différencier les cookies issus d'une réapparition due au *fingerprinting* de ceux issus d'un ciblage d'utilisateurs. Les auteurs ont alors pu identifier les acteurs responsables de la réapparition de cookies effacés et de la collaboration qui s'opère entre eux. Ils ont également démontré que la récupération de cookie est largement utilisée dans des sites web populaires et concluent que cette technique viole le Règlement Général sur la Protection des Données (RGPD).

1.2. Buts du travail de recherche

Dans mon travail je cherche à atteindre trois objectifs : obtenir des données collectées par de véritables utilisateurs, vulgariser les résultats pour en permettre l'accès à un large public et créer un cours pour des élèves du gymnase.

À l'exception de l'étude réalisée par Papadopoulos et al, toutes les études présentées ici se basent, pour des raisons évidentes de validation statistique, sur les résultats obtenus par des robots imitant des humains surfant sur le web. De plus, ces études s'effectuent sur des milliers de sites web durant des heures innombrables, sans interruption, ce qui n'est pas représentatif du comportement d'une personne réelle. Mon travail, au contraire, se base sur les données récoltées de personnes réelles, effectuant leur propre navigation.

Ces études abordent rapidement des détails techniques très pointus. Dès lors, elles sont la plupart du temps inaccessibles aux personnes intéressées mais non spécialisées. Mon deuxième objectif est de suffisamment vulgariser les techniques employées lors du traçage afin de le rendre compréhensible pour le plus grand nombre.

Finalement, je désire créer une séquence de cours pour des étudiants de première et deuxième année du gymnase, voie maturité, afin de leur montrer, avec des données et des faits concrets, ce que cela implique, en termes de publicité et de traces laissées sur internet, d'accepter ou non des cookies.

En menant cette étude sur des personnes réelles et non pas réalisées par des robots, je pense que cela rendra les conclusions et les recommandations plus tangibles pour les étudiants.

2. Côté technique

Dans ce deuxième chapitre, le lecteur fait connaissance avec les cookies. Tout d'abord, il découvre un bref historique sur les cookies, d'où ils viennent, pourquoi et par qui ils ont été inventés. Il en apprend ensuite davantage sur leur fonctionnement et les règles qu'ils doivent respecter. Une section dédiée au protocole HTTP est présentée, expliquant comment les requêtes et les réponses s'échangent entre le navigateur et les sites web

Par la suite, l'importance des cookies est mise en évidence, en les classant en tant que *First-Party Cookie*, *Third-Party Cookie*, voire *Second-Party Cookie*. Le lecteur comprend alors leur différence et leur utilité respective. Finalement, la dernière partie du chapitre aborde des sujets plus techniques. Des explications sur quelques procédures élaborées par les acteurs du web pour tracer ses utilisateurs, telles que Fingerprinting, Pixel/Image Tracking ou Cookie Syncing y sont exposées.

2.1. Cookie : Historique

Les cookies ont fait leur apparition en 1994 lorsqu'un employé de la compagnie Netscape Communication a développé, à la demande d'une compagnie cliente (MCI), un site de vente en ligne de produits de télécommunication longue distance (Wikipedia - MCI, 2023). Cette compagnie ne voulait pas stocker le contenu du panier du client sur leur propre serveur. Ils ont alors demandé au développeur de Netscape Communication de trouver un moyen pour stocker le contenu du panier sur l'ordinateur de l'utilisateur jusqu'à ce que ce dernier procède à l'achat (Andreessen, 2023). Ce fut la naissance des cookies.

La première utilisation réelle des cookies a eu lieu en 1995 sur la version bêta 0.9 de Mosaic Netscape pour vérifier si un utilisateur était déjà venu sur le site Netscape. À ce moment, Netscape était le navigateur prédominant sur le marché. Par la suite, d'autres navigateurs, comme Internet Explorer ou Mozilla Firefox, ont vu le jour et ont intégré cette technologie innovante dans leur système de navigation.

Le grand public n'était pas au courant de l'existence des cookies, ni du fait qu'ils étaient stockés dans leur machine. Dès 1996, suite à une publication du Financial Times (Jackson, 1996) émettant des inquiétudes au sujet de la protection de la sphère privée, les cookies ont fait l'objet de beaucoup d'attention et leur utilisation a été discutée lors de deux consultations du US Federal Trade Commission.

Des spécifications claires sur ce que pouvaient contenir ou non un cookie furent débattues et ont été publiées en 1997 sous la forme de la RFC2109 (Request for Comments). Elle a été, depuis lors, revue plusieurs fois pour aboutir à l'actuelle RFC6265 (cette dernière est toutefois sur le point de devenir obsolète si la nouvelle version est acceptée (Bingler et al., 2023)).

2.2. Cookie : Fonctionnement et restrictions

Les échanges d'information qui ont lieu lors de la navigation sur internet sont régis par le protocole HTTP (Andreessen, 2023). Ce protocole utilise un fonctionnement "sans état" (stateless), ce qui signifie que dès qu'un utilisateur change de page, même en restant sur le même site, le serveur du site aura oublié cet utilisateur. Dès lors, ce dernier devrait à nouveau s'identifier à chaque nouvelle requête, ce qui rendrait la navigation terriblement

pénible et inefficace ! Pour pallier ce problème, lorsqu'un utilisateur veut charger une page web, son navigateur envoie une requête au serveur, qui, s'il répond positivement, inclut le chargement de la page ainsi qu'un cookie, contenant un identifiant unique (UUID), spécifique à cet utilisateur. L'identifiant unique n'est autre qu'un bout de texte généré aléatoirement. Celui-ci est alors enregistré dans le navigateur du client. À chaque nouvelle transaction entre le navigateur et le serveur, c'est-à-dire à chaque changement de page demandé par le client, le navigateur envoie une requête au serveur accompagnée par le cookie, permettant ainsi l'identification du client. Le protocole HTTP peut être schématisé comme suit :

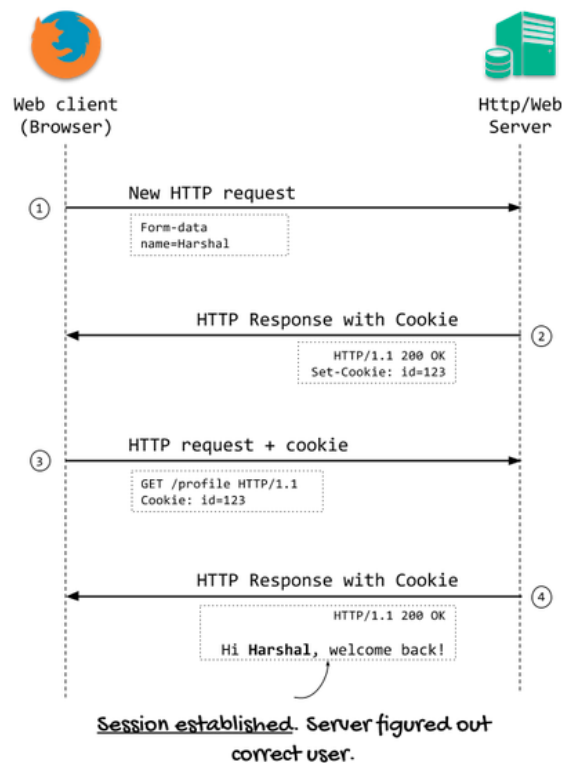


Figure 1 : Protocole HTTP (Patil, 2018)

Un cookie peut contenir toutes sortes d'informations au sujet de l'utilisateur. Cela peut aller du nombre de pages que son moteur de recherche doit afficher au mode d'affichage en passant par sa langue préférée. Voire beaucoup plus. Et c'est là que les dérives peuvent arriver.

Comme mentionné plus haut, des règles ont été établies pour protéger les utilisateurs. On les trouve dans la RFC6265 qui stipule ceci (Bingler et al., 2023) :

- Les cookies ont une taille limitée à 4KB max ;
- Un serveur ne peut pas enregistrer plus de 300 cookies dans un navigateur ;
- Un serveur ne peut pas lire les cookies d'un autre site web.

Ce dernier point vient de la politique de "Same Origin Policy" (W3C contributeurs, 2010) et il est central dans les restrictions vis-à-vis des cookies. Il mérite donc quelques clarifications.

Une origine est établie en comparant des parties bien spécifiques des adresses URL. Ces parties sont :

- Le protocole (http ou https) ;
- L'hôte (www.example.com) ;
- Le numéro de port (80).

Prenons les adresse URL suivantes comme exemple :

1. http://www.example.com/dir/page.html ;
2. http://www.example.com/dir2/page.html ;
3. https://www.example.com/dir/page.html ;
4. http://example.com/dir/page.html.

Les URL's 1 et 2 sont de même origine tandis que l'URL 3 n'a pas la même origine que la 1 car le protocole n'est pas le même (https vs. http) et l'URL 4 n'a pas la même origine que la 1 car le nom de domaine est différent (example vs. www.example).

Pour être à même de déterminer l'origine du cookie, examinons maintenant un peu plus en détail la requête HTTP à laquelle il est associé.

2.3. HTTP : requête et réponse (TutorialsPoint Contributor, 2023)

2.3.1. Requête HTTP

Une requête HTTP contient une ligne de requête, un en-tête et un corps de message (optionnel) qui eux-mêmes contiennent différentes informations :

Une ligne de requête : Méthode / URI / Version du protocole

Un en-tête de requête : zéro ou plus

Une ligne vide de séparation

Un corps de message : optionnel

La *méthode* de la ligne de requête définit ce que cette requête doit exécuter. Il en existe une grande variété, mais en ce qui nous concerne, nous pouvons nous arrêter sur les méthodes POST, GET et HEAD.

- La méthode POST : Elle permet d'envoyer des informations spécifiques au serveur, par exemple des informations clients ou un chargement de page utilisant des formulaires HTML. Dans ce cas, le corps du message est généralement renseigné ;
- La méthode GET : Elle permet de récupérer des données du serveur en utilisant une URI spécifique ; si le corps du message est renseigné, il sera aussi transmis ;
- La méthode HEAD : Elle agit de la même manière que la méthode GET mais en ne transférant que la ligne de requête et l'en-tête (sans le corps du message).

L'*URI (Uniform Resource Identifier)* est une chaîne de caractères qui identifie une ressource, soit grâce à son emplacement sur le web, soit grâce à son nom. Si l'emplacement est spécifié, c'est une *URL (Uniform Resource Locator)*. Si la ressource est identifiée par son nom alors c'est une *URN (Uniform Resource Name)*, comme par exemple le code ISBN d'un livre. (*URI - Glossaire MDN : Définitions Des Termes Du Web | MDN, 2022*)

Finalement on trouve la *version* du protocole HTTP utilisé.

En ce qui concerne l'en-tête de la requête, il fournit des informations réparties dans trois domaines :

- Des informations générales ;
- Des informations au sujet du client ;
- Des informations définissant le corps du message ou, s'il n'y a pas de corps, sur la ressource pointée par la demande.

Parmi les informations sur le client, on peut trouver, entre autres, les cookies.

Illustrons ce qui vient d'être discuté par deux exemples :

Le premier exemple montre la requête du client vers le serveur de watson.ch pour accéder au site. On voit dans la ligne de l'en-tête qu'il s'agit du chargement global de la page (*main_frame*). On trouve dans l'en-tête diverses informations inhérentes à l'utilisateur et à sa machine. Cette requête est la première que le navigateur envoie au site. Il n'y a pas encore de cookie qui y est associé.

```
GET / main_frame / HTTP 2.0

Host : www.watson.ch
User Agent : Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:105) Gecko/20100101 Firefox/105.0
Accept : text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,*/*;q=0.8
Accept language : en-US, en q=0,5
Accept encoding : gzip, deflate, br
Connection : Keep-alive
...

Une ligne vide de séparation

Corps de message : ∅
```

Le deuxième exemple illustre toujours une requête vers le site watson.ch mais cette fois on voit qu'un cookie y est associé. Ainsi, à chaque nouvelle requête du client vers le serveur de watson, le client enverra cet identifiant afin que le serveur puisse le reconnaître.

```
GET / style-sheet / HTTP 2.0

Host : www.watson.ch
User Agent : Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:105) Gecko/20100101 Firefox/105.0
Accept : ext/css,*/* q=0.1
Accept language : en-US, en q=0,5
Accept encoding : gzip, deflate, br
Connection : Keep-alive
Cookie : id=123

Une ligne vide de séparation

Corps de message : ø
```

À toute requête est associée une réponse. Sa structure est décrite ci-après.

2.3.2. Réponse HTTP

Suite à une requête HTTP, le serveur va envoyer une réponse HTTP qui contiendra cette fois une ligne de statut, un en-tête de réponse et un corps de message également optionnel.

```
Une ligne de statut : Version HTTP / code de statut / raison associée

Un en-tête de réponse : zéro ou plus

Une ligne vide de séparation

Un corps de message : optionnel
```

Les *codes de statut* sont des valeurs numériques composées de trois chiffres. Le premier des trois définit l'état du statut. Il en existe cinq :

- 1xx : Informatif, la requête a été reçue et le processus est en cours ;
- 2xx : Réussite, la requête a abouti ;
- 3xx : Redirection, la requête demande plus d'action pour aboutir ;
- 4xx : Erreur Client, la requête contient une erreur côté client ou ne peut pas aboutir ;
- 5xx : Erreur Serveur, le serveur n'a pas pu remplir la requête.

Parmi ces codes de statut, celui que nous avons déjà tous expérimenté est "404 : Page not found".

Les valeurs suivant le premier chiffre sont spécifiques à chacun de ces états. Le lecteur qui voudrait s'informer pourra les trouver à cette ressource : [Liste des codes HTTP — Wikipédia](#).

La *raison associée* est un texte expliquant le statut, par exemple “OK” en cas de succès ou “Page not found” en cas d’échec.

Les champs de l’en-tête de réponse permettent au serveur de transmettre des informations qui ne peuvent pas être placées dans la ligne de statut. Ces champs donnent des informations sur le serveur et les futurs accès à la ressource identifiée par l’URI.

Illustrons à nouveau ce qui vient d’être discuté par un exemple :

```
HTTP 3.0 / 200 / OK

Server : nginx/1.14.0 (Ubuntu)
Content type : text/css charset=UTF-8
Content encoding : g-zip
Last modified : Fri, 13 Feb 2009 23:31:30 GMT
x region : ch
x backend : 138.201.83.130
x host : www.watson.ch
set-cookie : id = 123 expires= Mon, 13-Oct-2025 09:58:04 GMT

Une ligne vide de séparation

Corps de message : ø
```

L’exemple ci-dessus montre la réponse à une requête au serveur de *watson.ch*. La ligne d’en-tête contient la version du protocole utilisée dans cette réponse (*HTTP 3.0*), le code de succès (*200*) et le *raison associée* (*OK*), signifiant que la page a pu être chargée. On peut également voir différents contenus dans l’en-tête dont par exemple l’URL du serveur (*www.watson.ch*), la région où il se situe (*ch*) et un cookie (*id = 123*).

Maintenant que nous connaissons la manière d’envoyer un cookie, regardons de plus près quelles sortes de cookie existent et à quelles fins ils sont utilisés.

2.4. Utilité des cookies

Il existe toutes sortes de cookies utilisés à différentes fins :

Les *cookies de session*, en général temporaires, sont ceux dont nous avons déjà parlé. Ils permettent d’identifier un utilisateur à travers un site et d’éviter de lui demander de s’identifier à chaque changement de page. Ce genre de cookie permet également d’enregistrer l’état d’un panier d’achat. Actuellement, les paniers ne sont plus enregistrés dans le cookie de l’utilisateur mais dans une base de données sur un serveur du site où l’achat a été effectué. Afin de garder la trace de quel panier appartient à quel utilisateur, le serveur envoie un cookie contenant un identifiant unique de session. Comme à chaque nouvelle requête de l’utilisateur son navigateur envoie le cookie au serveur, cela permet au serveur d’identifier le client et de lui attribuer le bon panier. Les cookies de session sont généralement effacés lorsqu’on quitte le navigateur. Ils peuvent avoir une durée de validité très courte (par exemple lorsqu’on s’identifie sur des sites bancaires) ou *ad aeternam* (par exemple lorsqu’on s’identifie sur les réseaux sociaux). Ceci sera défini dans l’en-tête de la réponse HTTP. Ces cookies sont *nécessaires* ou *essentiels*, sans eux, l’expérience de navigation sur le web ne serait pas agréable et prendrait un temps infini.

Les *cookies de personnalisation*. Ceux-ci permettent au site visité de proposer du contenu approprié à un utilisateur qui serait déjà venu par le passé et qui se serait identifié. L'enregistrement de la langue de préférence se fait par ce type de cookies. De plus, ils permettent d'établir des statistiques au niveau du bon fonctionnement (ou non) du site, du nombre de visiteurs, des rubriques visitées, du temps passé par page etc... Ces cookies sont agréables mais pas forcément essentiels.

Les *cookies de traçage*. Comme leur nom l'indique, ces cookies permettent de suivre l'utilisateur à travers le web. Ce sont eux les "mauvais" cookies. Regardons de plus près ce qu'ils font exactement.

Dès qu'un utilisateur fait appel à une page d'un site, l'échange habituel de requête-réponse HTTP entre le serveur et le navigateur démarre, ce qui inclut les cookies. Toutefois, au lieu de n'enregistrer que le strict nécessaire, le serveur enregistrera aussi l'URL de la page appelée, la date, l'heure de la requête et le cookie dans un fichier de journalisation. Tout ceci sera conservé dans une base de données, ce qui lui permettra de connaître les habitudes de l'utilisateur et de les recouper avec d'autres informations qu'il aurait déjà récoltées.

Pour rappel, un cookie ne peut être lu que par le site à qui il appartient, ou dit différemment à l'URL d'où le cookie est émis. Dès lors, comment le traçage à travers différents sites est-il possible ? Plusieurs techniques existent en fonction de ce qui est recherché. Elles sont décrites au paragraphe 2.6. Mais avant de s'y plonger, j'aimerais encore définir une terminologie très utilisée et pas toujours très claire : Les *First-Party Cookies*, les *Third-Party Cookies* et les *Second-Party Cookies*. (Zawadziński, 2018)

2.5. La classification des cookies

2.5.1. First-Party Cookies

Un *First-Party Cookie*, ou cookie de site primaire ou visité, est un cookie appartenant au site primaire. Il est issu de la réponse du serveur lorsque le navigateur lui envoie une requête. Par exemple, si l'utilisateur se rend sur le site www.24heures.ch, le navigateur envoie une requête au serveur de 24heures, qui, s'il répond positivement, envoie, avec le chargement de la page, un identifiant unique (un cookie) permettant de se rappeler de l'utilisateur et d'éviter de lui reposer la même question d'identification à chaque nouveau chargement de page ou à chaque nouvelle visite. Dans cette situation, l'URL associée au cookie est de même origine que l'URL du site visité. (Wlosik & Sweeney, 2021)

2.5.2. Third-Party Cookies

Un *Third-Party Cookie*, ou cookie de site tiers, est un cookie créé par un site différent de celui visité par l'utilisateur, d'où son nom de partie tierce. Ils sont utilisés pour tracer les utilisateurs à travers différents sites et procéder à du retargetage publicitaire. Par exemple, un utilisateur va comparer des crèmes solaires en vente sur le site suncream.ch. Plus tard, ce même utilisateur va sur le site bestbuy.ch et une publicité pour de la crème solaire apparaît.

La manière dont cela fonctionne est la suivante : les deux sites font appel à une compagnie publicitaire, par exemple adservice.com. Lors du chargement de leur page, un bout de code de [adservice](http://adservice.com) est également chargé. Dans cette situation, l'URL associée au cookie

(adservice) n'est pas de même origine que l'URL du site visité (suncream ou bestbuy). (Wlosik & Sweeney, 2021)

2.5.3. Second-Party Cookies

Le terme *Second-Party Cookie* est parfois utilisé pour décrire la situation suivante : lorsqu'un site primaire revend son *First-Party Cookie* à une compagnie avec laquelle il est en partenariat. On peut imaginer qu'une chaîne de restaurant soit en partenariat avec une compagnie de taxi et lui revende son propre cookie. Ainsi la compagnie de taxi a maintenant accès à des informations à propos des clients de la chaîne de restaurant, par exemple leur adresse. Si un utilisateur réserve une table au restaurant *via* internet, la compagnie de taxi peut lui proposer un service personnalisé pour l'amener au restaurant ou le ramener chez lui. (Wlosik & Sweeney, 2021)

Le paragraphe suivant se concentre sur les techniques de traçage, qui pourront donc être classées sous la dénomination de *Third-Party Cookie*.

2.6. Fingerprinting, cookie syncing et autres techniques de traçage

Les techniques de traçage et leur fonctionnement sont décrits dans les sections 2.6.1 à 2.6.4. On y trouvera :

- Fingerprinting : ce ne sont plus les données personnelles de l'utilisateur qui sont enregistrées mais les données intrinsèques de l'ordinateur avec lequel il expérimente le web ;
- Pixel : image invisible de taille 1 X 1 placée sur le site ;
- Image : image visible comme l'icône Facebook ou de la publicité ;
- Cookie syncing : partage de données entre plateformes de publicité.

La section 2.6.5 illustre les différents flux entre les acteurs publicitaires du web.

2.6.1. Fingerprinting

De plus en plus d'utilisateurs effacent les cookies lorsqu'ils quittent leur navigateur ou alors le navigateur le fait pour eux. Ainsi, l'utilisateur redevient anonyme, son "état" a été effacé. Afin de pallier cet effacement, des techniques de traçage sans passer par des cookies ont été mises en place. C'est ce qu'on appelle le *fingerprinting*. Ces techniques ne sont pas dépendantes de l'utilisateur mais se basent sur l'hypothèse que chaque appareil est unique et que l'enregistrement de certaines caractéristiques en temps réel de la machine de l'utilisateur sont suffisantes pour l'identifier. Les caractéristiques visées sont, par exemple, le langage par défaut, l'encodage choisi, les extensions installées, le navigateur utilisé, sa version, la résolution de l'écran etc... Chaque information prise séparément n'est pas spécifique, mais lorsqu'elles sont toutes agrégées, elles permettent de cibler davantage un utilisateur. La difficulté de prouver le *fingerprinting* réside dans le fait que certaines informations pourraient être utiles au bon fonctionnement du site, comme par exemple la résolution de l'écran. Contrairement au traçage par cookie, le fingerprinting ne peut pas garantir à 100% que deux utilisateurs n'ont pas le même profil. Toutefois, en recoupant de nombreuses données, la probabilité de se "tromper" d'utilisateur reste suffisamment faible

pour que cette technique soit utilisée. Par ailleurs, c'est souvent un mélange des deux techniques, enregistrement du "cookie" habituel et *fingerprinting*, qui est utilisé. (Baudry & Laperdrix, 2015) (Musa & Nithyanand, 2022, 295-313).

2.6.2. Pixel

L'utilisation d'image transparente de taille 1 x 1, c'est-à-dire un pixel invisible, permet de tracer un utilisateur de la manière suivante : Il suffit pour cela que le site primaire ait inclus dans sa page un pixel d'une société de traçage. L'ajout du pixel se fait grâce à une ligne de code supplémentaire dans le code HTML de la page web. Ce code pointe vers le serveur de la société de traçage. Lorsque le navigateur charge la page web demandée, le navigateur suit le lien et ouvre l'image, invisible. Le serveur enregistre cette activité dans sa base de données, incluant toutes sortes de données intéressantes au sujet de l'utilisateur qui lui, ne se doute de rien.

D'autres types de balise, comme les balises de redirection ou balise de suivi de clic, peuvent exister, toutefois, de nombreux navigateurs se méfient des requêtes inhabituelles venant des sites et sont plus susceptibles de les bloquer, alors que très peu d'entre eux empêchent les requêtes de chargement d'une toute petite image. (Edwards, 2022)

2.6.3. Image

Une image plus grande et visible peut aussi être utilisée plutôt qu'un pixel. Le fonctionnement est le même, si ce n'est pour la visibilité de l'image, à savoir que le navigateur suit le lien inclus sur la page recherchée pour charger l'image en question. En même temps que l'image est chargée, le serveur enregistre des informations au sujet de l'utilisateur

Une autre situation pourrait être que YouTube a placé une vidéo sur le site primaire, lorsque l'utilisateur clique sur cette vidéo, son navigateur va envoyer une requête à YouTube qui saura dès lors de quel site primaire "vient" cet utilisateur. YouTube peut alors proposer du contenu susceptible d'intéresser l'utilisateur.

2.6.4. Cookie syncing

Le cookie syncing implique que plusieurs partenaires partagent des informations au sujet d'un utilisateur et ce, malgré la politique de même origine. Le principe se base sur les *Third-Party Cookies* quel que soit le processus grâce auquel ils ont été enregistrés.

La Figure 2 présente les différents types de cookie qui peuvent être enregistrés par le navigateur. Pour permettre une certaine clarté de l'image, seules les réponses HTTP des différents acteurs sont montrées :

- L'encadré A est un *First-Party Cookie* (Cookie : E_ID=ABC). Il a été fourni par le site visité, ici *example.org*, afin d'identifier le visiteur et de lui garantir une bonne expérience de navigation.
- L'encadré B montre une situation de *Third-Party Cookie*. Ici, la page *example.org* contient une image (logo.gif) de la société *foo.org*. Lorsque le navigateur charge la page *example.org*, il suit le lien vers *foo.org* pour charger l'image (chemin B1). Le serveur de *foo.org* enregistre cette action et assigne un identifiant à l'utilisateur (chemin B2) (Cookie : F_ID = 123).

- L'encadré C illustre le procédé de cookie syncing, qui a lieu en deux étapes :
 - Le site primaire, *example.org*, héberge également d'autres contenus, par exemple de la publicité sous forme d'image, appartenant à la société *bar.org* (chemin C1) ;
 - La société *bar.org* est en partenariat avec la société *sync.org* ;
 - Lorsque le navigateur charge l'image appartenant à *bar.org*, cette dernière envoie, en même temps que l'image, une requête de redirection vers son partenaire, *sync.org* (chemin C2) ;
 - Cette demande de redirection contient l'identifiant que *bar.org* a attribué à l'utilisateur (chemin C3 : *sync.org?bar user id=XYZ*) ;
 - La société *sync.org* sait maintenant que l'utilisateur portant l'identifiant unique de *bar.org* a visité le site *example.org* ;
 - Chaque société va enregistrer dans sa propre base de données les identifiants permettant de reconnaître l'utilisateur (chemin C4) ;
 - Les sociétés *bar.org* et *sync.org* étant en partenariat, échangent les données qu'elles ont au sujet de cet utilisateur, ce qui aboutira à un ciblage plus fin.

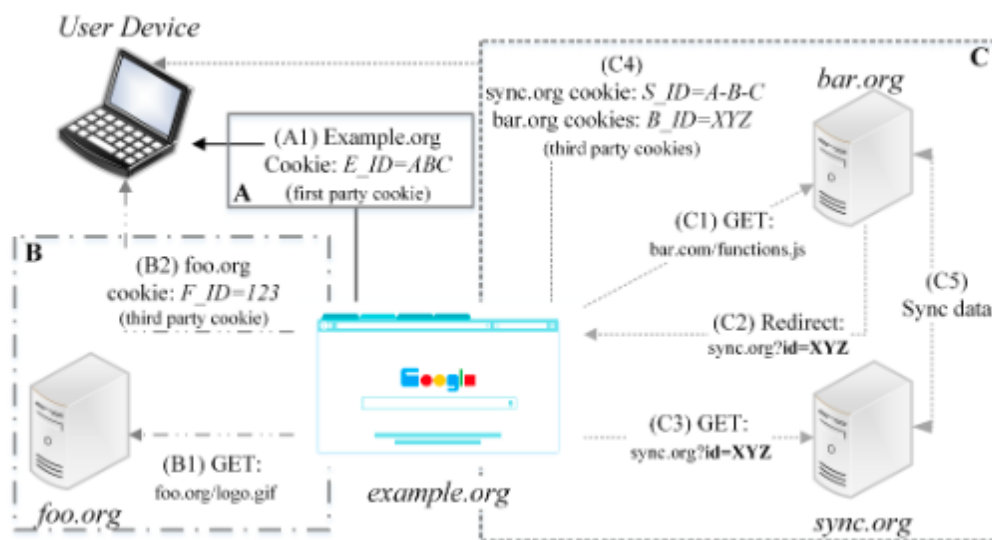


Figure 2 : Les différents type de cookie : A = first party cookie , B = third party cookie et C = cookie syncing (Urban et al., 2020)

2.6.5. Les acteurs publicitaires du web

Devant la quantité de données de plus en plus grande (Gaudiaut, 2022) (Schmitt, 2023) que les compagnies fictives comme *sync.org* ou *bar.org* peuvent récolter, et afin de les rentabiliser au maximum, ces dernières se sont organisées en fonction de leur compétence, à savoir les unes proposent de l'espace publicitaire au plus offrant (Sell Side Platform - SSP), les autres offrent du contenu (Demand Side Platform - DSP) et les troisièmes gèrent les données des utilisateurs (Data Management Platform - DMP). SSP et DSP sont mis en

relation par une plateforme d'échange publicitaire (ADX). Cette dernière gère l'ensemble des publicités proposées. (*Glossary of Marketing and Data Terms, 2023*)

La Figure 3 ci-dessous reflète les échanges qui ont lieu entre tous ces acteurs, rendus possibles grâce au cooking syncing.

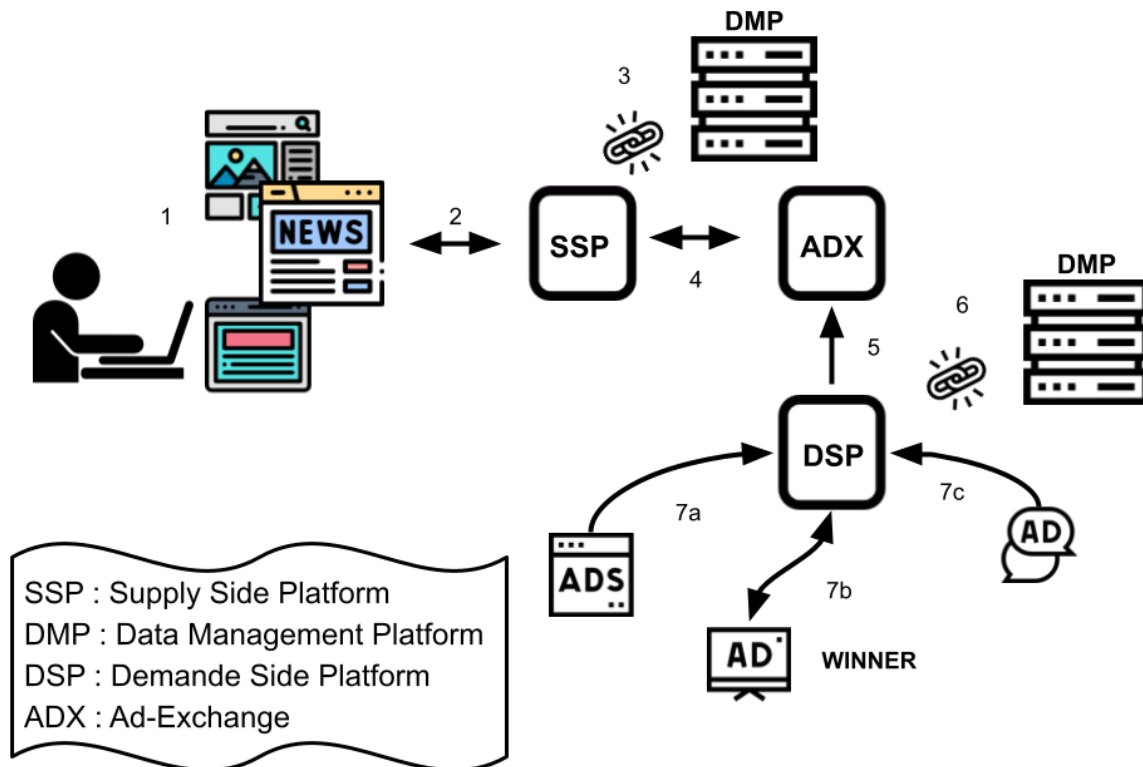


Figure 3 : Flux de données entre l'utilisateur et les acteurs du web

1. Les utilisateurs surfent sur le web, leur navigateur envoie des requêtes HTTP et accepte des cookies en échange du chargement de la page demandée ;
2. Le site est en lien avec une plateforme de vente en ligne (SSP) car il a de l'espace à vendre sur sa page pour y mettre du contenu publicitaire ;
3. La SSP peut monétiser son espace puisqu'elle connaît l'utilisateur grâce à son partenariat avec une DMP. Plus les informations sur l'utilisateur sont complètes, plus elle pourra vendre cher son espace ;
4. La SSP annonce aux échangeurs de publicité (ADX) qu'elle a une offre en cours ;
5. L'ADX va alors avertir la plateforme de fournisseur de contenu (Demand Side Platform - DSP) qu'elle peut fournir de la publicité ;
6. La DSP connaît aussi les préférences de l'utilisateur grâce à sa connexion avec une DMP~;
7. La DSP annonce à ses clients qu'elle a de la disponibilité pour eux. Un système d'enchères entre en jeu et l'espace publicitaire ira au plus offrant.

Tout ce processus a lieu en quelques fractions de secondes. Plus la SSP et la DSP sont en possession d'informations sur le client, plus elles pourront monétiser l'espace ou le contenu

dont elles disposent. Ainsi toutes deux sont en partenariat avec des plateformes de gestion de données (DMP), celles-là même qui récoltent les données des surfeurs du web. La boucle est ainsi bouclée.

Afin de se faire une idée des acteurs publicitaires du web existants, je liste ici quelques-unes des plateformes les plus couramment utilisées, dont nous verrons les noms apparaître au cours de ce rapport.

Supply Side Plateform : Google Ad Manager (anciennement DoubleClick) ; Rubicon Project (appartenant maintenant à Magnite) ; Index Exchange (casalemedia); OpenX ; Xandr (anciennement AppNexus, adnxs) ; Pubmatic. (Happy, 2023)

Demand Side Platform : DV360 ; Trade Desk ; Amazon DSP ; Yahoo ; Google Ads. (Schroeder & Tyler, 2023)

Ad-Exchange : Google AdX (anciennement DoubleClick Ad exchange) ; OpenX ; Xandr ; Magnite ; Index Exchange ; Pubmatic. (Trevisani & Martin, 2022)

Data Management Platform : OnAudience ; Oracle BlueKai ; Permutive ; Adobe Audience ; Lotame ; Ramp. (Thomas & Martin, 2023)

3. Présentation de l'étude

Ce chapitre présente les démarches entreprises auprès du comité d'éthique de l'EPFL pour être en mesure de réaliser l'étude, son déroulement à proprement parler ainsi que le fonctionnement de Lightbeam et de Thunderbeam, les deux extensions installées sur les ordinateurs des personnes ayant accepté de participer à l'étude. Afin de faciliter la compréhension des termes utilisés, une section est dédiée à la terminologie spécifique des extensions. En fin de chapitre, chaque situation pouvant être rencontrée lors d'un surf sur internet est expliquée en détail à l'aide d'exemples.

3.1. Participants et récolte des données

Comme déjà mentionné, l'objectif de l'étude étant de construire un cours destiné à des élèves de gymnase (secondaire II) sur le thème du traçage sur internet. Pour rendre l'étude plus concrète, j'ai choisi de la réaliser avec de vraies personnes naviguant sur internet dans leur vie quotidienne. S'agissant de données potentiellement sensibles, et sur les conseils de mon directeur de projet, avant de pouvoir procéder au recrutement, je me suis d'abord assurée que j'avais l'autorisation de mettre à exécution cette étude .

À cette occasion j'ai discuté avec Chiara Tanteri, Data Protection Officer, et Gaia Barazzetti, Research Ethic Compliance Officer, toutes deux à l'EPFL. Après leur avoir exposé le projet, elles ont confirmé la nécessité d'écrire une demande de consentement adressée au comité éthique de l'EPFL. Durant nos échanges, il est apparu très clairement que je ne pourrais pas demander aux élèves de participer à cette étude ni à mes collègues.

Les élèves, car ils sont pour la plupart mineurs, et quand bien même ils seraient adultes, le fait de connaître leurs habitudes de surf pourrait poser des problèmes éthiques justement. De plus, il est probable que les élèves ne possèdent pas un ordinateur personnel mais que ce soit un ordinateur familial. Dès lors, cela me mettrait en possession de données non centrées sur une personne mais sur toute une famille, or seul l'étudiant aurait été avisé de mon projet, ce qui est à nouveau éthiquement inacceptable.

Je ne pouvais pas non plus faire participer mes collègues, car il se pourrait que je me retrouve en possession d'informations sensibles sur eux, ce qui me mettrait dans une position délicate face à eux.

Finalement, et ce pour des raisons techniques, il était essentiel que ce suivi puisse se faire sur ordinateur et non pas sur un appareil mobile de type smartphone ou tablette.

Lors de la rédaction à proprement parler de la demande de consentement, la personne de référence a été Mme Barazzetti uniquement. Elle m'a guidée tout au long de ma démarche.

Les premiers points à justifier étaient le nombre de personnes interrogées, leur âge et la manière de les recruter. J'ai également dû demander à un des responsables du programme GymInf de soutenir mon projet, en l'occurrence Monsieur Olivier Levêque.

La demande comprend une variété de documents dont un protocole de recherche, une fiche d'information destinée aux participants ainsi qu'un formulaire de consentement que ces derniers devaient signer s'ils acceptaient de participer à l'étude. Les questions posées se concentrent principalement sur la protection des données, en abordant des sujets tels que

l'accès aux données, les mesures de sécurité mises en place et la conservation des données après l'étude.

La demande au comité d'éthique a été soumise à une revue par deux relecteurs et un conseiller légal. Après plusieurs échanges, notamment concernant l'utilisation de Lightbeam et la garantie de non-conservation des données des utilisateurs sur des serveurs de Lightbeam, ainsi que le choix de mon groupe de participants, la demande a été validée le 22 novembre 2022. Elle se trouve en Annexe et contient, entre autres, les informations et documents suivants :

- L'âge des participants : entre 25 et 70 ans ;
- La taille du groupe de participants : 10 personnes ;
- Le moyen de les recruter : tout d'abord par l'envoi d'un questionnaire à "large" échelle, puis par une rencontre durant laquelle je leur expliquais leur rôle dans l'étude ;
- La liberté de se retirer de l'étude à tout moment ;
- La protection de l'anonymat : les données sont enregistrées sur un ordinateur dédié à cette recherche auquel seul moi ai accès ;
- Le moyen de transfert des données : il a lieu sur clé USB entre le participant et moi ;
- La conservation de données brutes à la fin de l'étude : détruites après la fin de l'étude ;
- Un protocole de recherche ;
- Un formulaire de consentement à signer par les participants ;
- Un formulaire d'information à lire et à distribuer aux participants ;
- La lettre de soutien de M Levêque.

Une fois l'accord du comité obtenu, pour garder la sphère privée j'ai créé un questionnaire sur Cryptpad (CryptPad, 2023) et l'ai envoyé par courriel à un grand nombre de mes connaissances. Il est accessible par ce lien [Formulaire Cryptpad](#). Ceci qui m'a permis de sélectionner des participants en fonction de plusieurs critères :

- Avoir plus de 25 ans ;
- Avoir un ordinateur personnel ;
- Avoir une connexion internet ;
- Utiliser Firefox ou Google Chrome ;
- Être d'accord pour participer à l'étude.

Une fois le panel de personnes trouvé, l'étude a pu commencer. Les participants devaient installer une extension fonctionnant sur Firefox - Lightbeam ou sur Chrome - Thunderbeam selon s'ils utilisent l'un ou l'autre de ces navigateurs. Chacun était libre d'accepter ou de refuser les cookies lors de leur expérience sur le web. Ils pouvaient également désactiver à tout moment l'extension, ainsi que se retirer de l'étude. Par défaut, les extensions n'enregistrent pas le trafic internet si le surf est effectué sur une page privée.

Les instructions qu'ils ont reçues étaient d'observer si de la publicité spécifique apparaissait s'ils avaient éventuellement effectué un achat, ainsi que d'enregistrer chaque soir les données collectées. Les extensions fonctionnant en mode continu, cette demande n'était pas réellement nécessaire mais uniquement une assurance en cas de problème.

Après trois jours de navigation je leur ai demandé de réaliser une recherche spécifique sur Tripadvisor et d'observer les publicités apparaissant les jours suivants, potentiellement en lien avec cette recherche.

Afin que les données ne soient vues que par mes yeux, j'ai utilisé un ordinateur dédié à cette étude pour toutes les étapes liées au traitement des données. J'ai récolté les données via des clés USB appartenant aux utilisateurs, que j'ai enregistrées de manière anonymisée sur l'ordinateur dédié puis, j'ai vidé les clés USB de leur contenu et les ai rendues à leur propriétaire. Je me suis également assurée qu'ils avaient désinstallé Lightbeam ou Thunderbeam, effacé le fichier contenant leurs données de surf et vidé la corbeille de leur ordinateur. S'il n'était pas possible de se voir physiquement, nous nous sommes rencontrés par visio et je me suis assurée des mêmes points que lorsque j'étais sur place. Les données ont alors transité via un site sécurisé, Cryptpad, qui ne conserve aucun enregistrement des données. J'ai, de plus, immédiatement après avoir enregistré les données sur l'ordinateur dédié, effacé les données de Cryptpad.

3.2. Terminologie utilisée dans les extensions Lightbeam et Thunderbeam

Avant d'aller plus loin dans l'explication du fonctionnement de ces extensions, je présente ici la terminologie trouvée dans le fichier enregistré lorsque l'utilisateur procède à une sauvegarde de ses données de surf :

Nom du champ	Fonction	Valeurs possibles
<i>First Party</i>	Indique si le cookie appartient au site primaire, visité par l'utilisateur	true (vrai) ou false (faux)
<i>Hostname</i>	Représente un site primaire ou tiers selon la valeur de <i>First Party</i>	adresse de site web
<i>First Party Hostnames (FPHN)</i>	Représente le site primaire visité par l'utilisateur et qui envoie des requêtes à d'autres sites	false ou adresse(s) de site(s) web
<i>Third Parties</i>	Indique si le cookie appartient à un site tiers	adresse(s) de site(s) web
<i>Favicon</i>	Code de l'icône	vide* ou code de l'icône

Tableau 1 : Nom des champs de Lightbeam, leurs fonctions et entrées possibles, trouvés lors de l'enregistrement des données

*En ce qui concerne le champ *Favicon*, parfois la requête utilisée par Lightbeam pour obtenir l'icône n'aboutit pas et dans ce cas, seul un rond blanc apparaît sur l'image et le champ "*Favicon*" n'est pas renseigné (Jonoxia, 2012).

La différence entre *Hostname* et *First Party Hostnames* est assez fine et dépend de la valeur trouvée sous *First Party* :

- Si la valeur *First Party* est *true*, le site visité se trouve sous *Hostname*.
- Si la valeur *First Party* est *false*, le site visité se trouve sous *First Party Hostnames* et le site tiers est sous *Hostname*.

On verra dans la section suivante les situations qui amènent le champ *First Party Hostnames* à être renseigné ou non.

Il est crucial de se rappeler qu'un *Hostname* n'est pas nécessairement un site visité par l'utilisateur

3.3. Lightbeam - Thunderbeam

Lightbeam est une extension du navigateur Firefox, soutenue par Mozilla Firefox jusqu'en 2019 ("L'extension Lightbeam Pour Firefox N'est Plus Prise En Charge | Assistance De Firefox," 2019). Ils ont, depuis lors, changé leur politique de protection face au traçage et ont intégré d'autres outils de protection dans leur navigateur. Toutefois Lightbeam est toujours développé et maintenu par son concepteur, Christoph Klassen (Klassen, 2023).

Lightbeam est écrit en javascript et permet une visualisation graphique du trafic internet généré lorsqu'une personne navigue sur internet.

Thunderbeam fonctionne de la même manière que Lightbeam mais sur Chrome au lieu de Firefox. (*Thunderbeam-Lightbeam for Chrome*, 2021)

Ces extension n'enregistrent pas la totalité des informations contenues dans la requête HTTP, mais uniquement l'adresse URL (adresse de type *www.24heures.ch*) des sites visités, que l'on appellera site primaire et celle des sites tiers vers qui le site primaire envoie des informations. Comme vu dans la section 2.3, l'adresse pointe vers un serveur et permet le chargement de la page. Elle peut parfois contenir d'autres informations, telles que des identifiants de login, un cookie, une adresse IP ou autre. On ne verra donc pas toujours les cookies envoyés ou reçus, mais c'est bien à travers ces cookies que le trafic ou l'échange d'informations entre différents partenaires peut avoir lieu.

L'extension permet également une représentation graphique de tout le trafic lié à internet comme le montre la Figure 4.

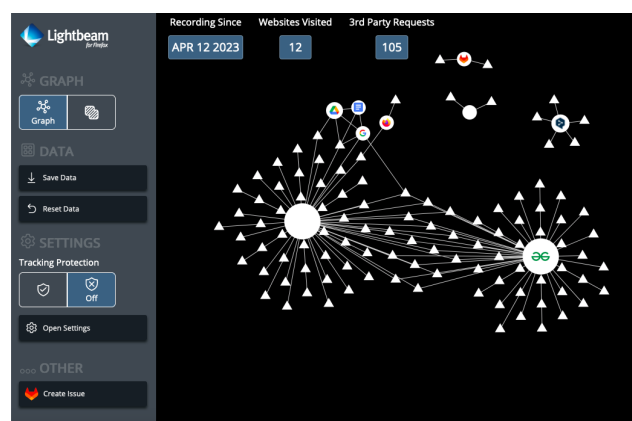


Figure 4 : Capture d'écran de Lightbeam

Sur la Figure 4, les ronds sont les sites primaires, c'est-à-dire les sites que l'utilisateur a voulu voir, ceux dont il a cherché l'adresse dans un moteur de recherche ou qu'il a tapé directement dans la bar URL. Certains de ces sites sont de simples ronds blancs tandis que d'autres montrent l'icône du site. Ces derniers sont alors plus facilement identifiables. Les ronds blancs apparaissent lorsque la requête de Lightbeam pour obtenir l'icône n'a pas abouti (Jonoxia, 2012). On trouve l'adresse du site sous le champ *Hostname* à la condition que le champ *First Party* soit renseigné à *true*.

Les triangles sont les sites tiers (*Third Parties*). Cette fois l'utilisateur n'a pas cherché à atteindre ces sites. Comme expliqué dans les sections 2.2 et 2.3, ce sont les sites primaires qui envoient des informations à ces sites tiers, avec les différents objectifs mentionnés dans la section 2.4, tels que permettre le maintien d'une session ouverte, le chargement d'une police de caractère, l'enregistrement d'un panier, de préférence de langage ou autres, mais aussi dans un but de traçage.

Il est également intéressant de constater que certains sites sont des îlots tandis que d'autres sont "reliés" les uns aux autres par des sites tiers (lorsqu'un triangle est connecté à deux ronds différents). Ce sont ces derniers qui seront mis sous la loupe lors de l'analyse des résultats.

Lightbeam rend compte de combien de sites ont été visités (dans la situation montrée, ici 12 sites ont été visités) et combien de sites tiers ont été contactés (ici 105 !). L'extension permet également de sauver les données (avec le bouton Save Data) ou d'effacer les données enregistrées (Reset Data). Toutes les données sont stockées dans le navigateur de l'utilisateur, aucune donnée n'est enregistrée sur les serveurs de Lightbeam.

Lorsque l'utilisateur clique sur le bouton *Save Data*, un fichier en format json est enregistré dans le disque dur de l'utilisateur. Comme mentionné dans le Tableau 1 de la section terminologie, cinq champs y sont renseignés dans le fichier json, à savoir "*Hostname*", "*Favicon*", "*First Party*", "*First Party Hostnames*" et "*Third Parties*".

Seule l'adresse URL est enregistrée. Ainsi, si elle contient les identifiants de l'utilisateur, ou l'adresse IP ou le cookie assigné à l'utilisateur, alors ces informations seront également accessibles. Ce n'est toutefois pas toujours le cas.

Pour que l'observation découlant de l'étude soit pertinente, les utilisateurs ont laissé le boutons de protection contre le traçage désactivé. En effet, ce qui nous intéresse ici n'est pas l'efficacité de l'extension mais bien le trafic internet "habituel".

Les différentes configurations que l'on peut rencontrer dans le fichier json sont décrites dans le Tableau 2, on trouve les explications détaillées de chaque cas sous les sections 3.3.1 à 3.3.6.

#	<i>First Party</i>	<i>Hostname</i>	<i>First Party Hostnames (FPHN)</i>	<i>Third Parties</i>
1	True	lematin.ch	false	adresse(s) de third parties
2	True	20min.ch	lematin.ch	adresse(s) de third parties
3	True	google.com	multiple noms de sites	adresse(s) de third parties
4	False	account.booking.com	www.booking.com	adresse(s) de third parties
5	False	fooby.ch	20min.ch	néant
6	False	htlb.casalemedia.com	multiples noms de sites	néant

Tableau 2 : Toutes les configurations possibles enregistrées dans le fichier json issu de Lightbeam ou Thunderbeam

3.3.1. Situation 1

```
"blog.webf.zone": {
  "hostname": "blog.webf.zone",
  "favicon": "",
  "firstPartyHostnames": false,
  "firstParty": true,
  "thirdParties": [
    "api2.branch.io",
    "app.link"
  ]
},
```

Figure 5 : Extrait du fichier *.json montrant un site primaire (*First Party : true*) et les sites auxquels il envoie des informations. Ici ce sont des cookies de fonctionnement.

L'extrait du fichier lightbeamData.json dans la Figure 5 montre un serveur qui envoie des informations au sujet de l'utilisateur uniquement à des fins de fonctionnement de son site.

- Le nom du site primaire est renseigné sous le champ *Hostname*. Ici il s'agit de , blog.webf.zone ;
- Le champ *Favicon* n'est pas renseigné car la requête Lightbeam pour obtenir l'icône n'a pas abouti (cf Figure 6) ;
- Le champ *First Party* est renseigné à *true* ; ce qui indique que l'utilisateur a voulu visiter ce site ;
- Le champ *First Party Hostnames* est renseigné à *false*. Ce site a été accédé en tapant l'adresse URL directement et il n'apparaît pas en tant que *Third Parties* dans un autre site ;
- Le champ *Third Parties* contient les adresses des sites nécessaires (2) pour l'affichage de la page.

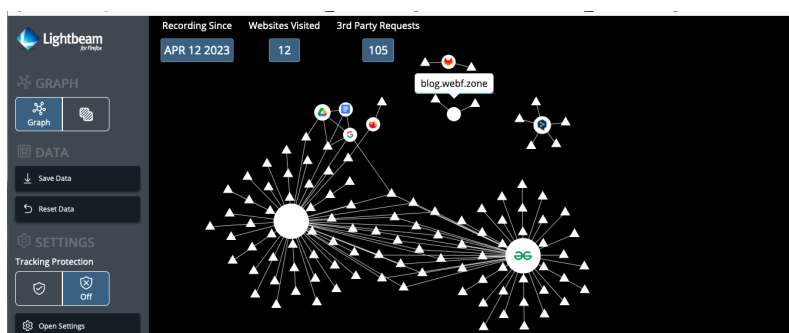



Figure 6 : Capture d'écran de Lightbeam montrant un site isolé des autres (blog.webf.zone), en lien avec deux sites tiers uniquement ainsi que d'autres sites, dont synonymo.fr (rond blanc à droite) et geekforgeeks (), en lien avec une multitude de sites tiers

```

"www.synonymo.fr": {
  "hostname": "www.synonymo.fr",
  "favicon": "",
  "firstPartyHostnames": false,
  "firstParty": true,
  "thirdParties": [
    "www.google.com",
    "cache.consentframework.com",
    "choices.consentframework.com",
    "www.google-analytics.com",
    "cdn.jsdelivr.net",
    "c.amazon-adsystem.com",
    "connect.facebook.net",
    "www.googletagservices.com",
    "js.sddan.com",
    "bidder.criteo.com",
    "htlb.casalemedia.com",
    "ib.adnxs.com",
    "aax.amazon-adsystem.com",
    "securepubads.g.doubleclick.net",
    "adservice.google.com",
    "adservice.google.ch",
    [...]
  ],
},

```

Figure 7 : Extrait du fichier *.json montrant un site primaire (*First Party : true*) et les sites tiers auxquels il envoie des informations.

La Figure 7 montre une situation similaire à la Figure 5. Les champs sont ici renseignés de la même manière si ce n'est le champ *Third Parties* qui contient une liste de sites tiers si longue qu'elle dépasse la capture d'écran !

- Le nom du site primaire est renseigné sous le champ *Hostname : synonymo.fr* ;
- Le champ *Favicon* n'est pas renseigné car la requête Lightbeam pour obtenir l'icône n'a pas abouti (cf Figure 6) ;
- Le champ *First Party Hostnames* est renseigné à *false*. Ce site a été accédé en tapant l'adresse URL directement et il n'apparaît pas en tant que *Third Parties* dans un autre site ;
- Le champ *First Party* est renseigné à *true*, ce qui implique que l'utilisateur a volontairement visité ce site ;
- Le champ *Third Parties* est renseigné par tous les sites auxquels synonymo.fr fait appel, que ce soit pour le chargement complet de sa page, son bon fonctionnement ou encore de la publicité.

Certains sites tiers que l'on peut voir ici enregistrent des données à des fins statistiques (comme google-analytics, js.sddan.com) ou à des fins de consentement vis-à-vis des cookies (consentframework) et d'autres sont des sites connus pour proposer des contenus

publicitaires (googleadservices, critico, casalemedia, adnxs et doubleclick pour ne citer qu'eux) (Ghostery GmbH, 2023). On verra plus loin que ces sites font le lien avec d'autres sites primaires, ce qui permet aux publicitaires de mieux cibler les utilisateurs.

3.3.2. Situation 2

```
"www.20min.ch": {
  "hostname": "www.20min.ch",
  "favicon": "data:image/x-icon....",
  "firstPartyHostnames": [
    "www.lematin.ch"
  ],
  "firstParty": true,
  "thirdParties": [
    "beagle.prod.tda.link",
    "tdn.da-services.ch",
    "www.googletagmanager.com",
    "cdn.cookieclaw.org",
    "fooby.ch",
    "recipecontent.fooby.ch",
    "securepubads.g.doubleclick.net",
    "ad.doubleclick.net",
    "pubads.g.doubleclick.net",
    "weather.da-services.ch",
    "api.20min.ch",
    "ib.adnxs.com"
  ]
},
[...]
```

Figure 8 : Extrait du fichier *.json montrant un site primaire (*First Party : true*), une *FPHN* de nom différent du site primaire et une multitude de sites auxquels le site primaire envoie des informations.

```
"www.20min.ch": {
  "hostname": "www.20min.ch",
  "favicon": "",
  "firstPartyHostnames": [
    "www.lematin.ch"
  ],
  "firstParty": false,
  "thirdParties": []
},
```

Figure 9 : Extrait du fichier *.json montrant un site tiers (*First Party : false*) et le site primaire (*FPHN*) qui a envoyé une requête au site tiers.

La Figure 8 montre une situation particulière dont voici les détails :

- Le nom du site primaire est renseigné sous le champ *Hostname* : *www.20min.ch*;
- Le champ *Favicon* est renseigné par le code de l'icône du site (le code est tronqué pour une question de clarté) ;
- Le champ *First Party Hostnames* est renseigné par un nom de domaine différent du *Hostname*, *FPHN* = *www.lematin.ch* ;
- Le champ *First Party* est renseigné à *true*, ce qui implique que l'utilisateur a volontairement visité le site *www.20min.ch*;
- Le champ *Third Parties* est renseigné par tous les sites auxquels *www.20min.ch* fait appel, que ce soit pour le chargement complet de sa page, son bon fonctionnement ou encore de la publicité.

Cette situation reflète une procédure qui a lieu en deux temps. Au temps $t = 0$, l'utilisateur est allé sur le site lematin.ch. Au temps $t = 1$, il a visité le site du 20min.ch. Le site du 20min étant présent comme *Third Parties* du site lematin.ch (ce qui se voit sur la Figure 9), lorsque l'utilisateur a effectivement visité le site du 20min.ch, ce dernier a vu son statut passer de *First Party = false* (Figure 9) à *First Party = true* (Figure 8).

Ce que nous devons retenir de cette situation est le fait que, lorsque l'on a affaire à un site primaire, c'est-à-dire *First Party = true*, si *FPHN* est renseigné par un nom de site, plutôt que par *false*, cela signifie que ce site a déjà demandé l'enregistrement d'un cookie dans le navigateur de l'utilisateur. De plus, il se peut que plusieurs sites apparaissent sous le champ *FPHN*. C'est ce que nous allons décrypter dans la situation 3.

Mais avant cela, regardons un instant les *Third Parties* qui apparaissent dans la situation 2. À nouveau, la liste est si longue qu'elle est ici tronquée. On voit des sites de fournisseurs de services publicitaires tels que doubleClick, adnxs ou da-services, des sites de fonctionnement tels que googltagmanager, cookielaw ou api.20min et encore un site "fooby.ch", présent uniquement à des fins de recoupement d'information sur l'utilisateur.

3.3.3. Situation 3

```
"www.google.com": {
  "hostname": "www.google.com",
  "favicon": "data:image/x-icon;base64...",
  "firstPartyHostnames": [
    "www.anibis.ch",
    "www.autoscout24.ch",
    "www.bmw.ch",
    "www.lightinthebox.com",
    "drive.google.com",
    "docs.google.com",
    "fr.tripadvisor.ch",
    "www.cheapflights.com",
    "www.watson.ch",
    "www.lematin.ch",
    "www.20min.ch"
  ],
  "firstParty": true,
  "thirdParties": [
    "adservice.google.com",
    "fonts.gstatic.com",
    "www.gstatic.com",
    "sst.anibis.ch",
    "bat.bing.com",
    [...]
  ],
}
```

Figure 10 : Extrait du fichier *.json montrant un site primaire (*First Party : true*), une multitude de *FPHN* de noms différents et une multitudes de sites auxquels il envoie des informations.

La Figure 10 montre une situation où l'utilisateur est allé sur le site primaire, ici google.com.

- Le nom du site primaire est renseigné sous le champ *Hostname* : *www.google.com*;
- Le champ *Favicon* est renseigné par le code de l'icône du site (à nouveau tronqué pour une question de clarté) ;
- Le champ *First Party Hostnames* est renseigné par plusieurs noms de domaine différent du *Hostname* ;
- Le champ *First Party* est renseigné à *true*, ce qui implique que l'utilisateur a volontairement visité ce site ;
- Le champ *Third Parties* est renseigné par tous les sites auxquels google.com fait appel, que ce soit pour le chargement complet de sa page, son bon fonctionnement ou encore de la publicité.

Comme dans la situation 2, google.com était déjà présent comme *Third Parties* dans la multitude de sites que l'on voit sous *FPHN* et visité précédemment par l'utilisateur. Dans le cas de Google, on peut imaginer que les sites font appel à ses services pour des questions de statistique, mais on voit également qu'une des *Third Parties* visible est "adservice.google.com" qui n'est autre que le service de publicité de Google. Google pourra dès lors croiser toutes les données qui lui parviennent depuis ces différents sites et marchander ces précieuses informations pour une vente lucrative de publicité.

3.3.4. Situation 4

```
"accounts.google.com": {
  "hostname": "accounts.google.com",
  "favicon": "",
  "firstPartyHostnames": [
    "www.google.com",
    "www.linkedin.com",
    "drive.google.com",
    "docs.google.com"
  ],
  "firstParty": false,
  "thirdParties": []
},
```

Figure 11 : Extrait du fichier *.json montrant un site tiers (*First Party* : *false*) et les sites d'où ils tirent des informations (sous *FPHN*). La situation est ici particulière car 1) l'origine n'est pas la même entre *Hostname* et *FPHN* (d'où la mention "false" pour *First Party*) et 2) le domaine peut ou non être le même (google vs. linkedin).

La Figure 11 illustre une situation classique d'une personne qui s'identifie sur un compte, par exemple son compte Google. On peut, dès lors, être surpris de constater que cette situation est jugée comme un cookie de partie tierce. L'explication réside de la politique de même origine (*Same Origin Policy*), mentionnée dans la section 2.2.

En effet, lorsque l'on se connecte à un compte, le domaine peut être le même ou pas. Par exemple, si on se connecte à son GoogleDrive via son compte Google, le domaine est le même, tandis que si on se connecte à son compte LinkedIn *via* son compte Google, le

domaine n'est pas le même. Dans tous les cas, l'origine au sens de "Same Origin Policy" n'est pas la même, ainsi le fait de s'identifier sur un compte rend le cookie d'une origine différente, donc jugée de tiers partie.

- Le nom du site accédé est renseigné sous le champ *Hostname* : *accounts.google.com* ;
- Le champ *First Party* est renseigné à *false*, ce qui implique que l'utilisateur n'a pas volontairement visité ce site. Comme dit, ici la situation est particulière car le visiteur a voulu s'identifier ;
- Le champ *Favicon* n'est pas renseigné car Lightbeam n'enregistre pas les icônes des sites tiers (ce sont les triangles blancs sur la Figure 4) ;
- Le champ *First Party Hostnames* est renseigné avec les sites où l'utilisateur veut s'identifier ;
- Le champ *Third Parties* est vide puisque le site lui-même est en quelque sorte une partie tierce.

Cette situation ne sera pas comptée lorsqu'il s'agira de quantifier le nombre de sites tiers ayant accès à un site primaire.

3.3.5. Situation 5

```
"fooby.ch": {
  "hostname": "fooby.ch",
  "favicon": "",
  "firstPartyHostnames": [
    "www.20min.ch"
  ],
  "firstParty": false,
  "thirdParties": []
},
```

Figure 12 : Extrait du fichier *.json montrant un site tiers (*First Party* : *false* et *Hostname* : *fooby.ch*) et le site auquel il est lié (*First Party Hostnames* : "*www.20min.ch*")

La Figure 12 détaille un cas en lien avec la Figure 8 de la situation 2. Cette situation est la cinquième du Tableau 2.

- Le nom du site accédé est renseigné sous le champ *Hostname* : *fooby.ch* ;
- Le champ *First Party* est renseigné à *false*, ce qui implique que l'utilisateur n'a pas volontairement visité ce site ;
- Le champ *Favicon* n'est pas renseigné car Lightbeam n'enregistre pas les icônes des sites tiers (ce sont les triangles blancs sur la Figure 4) ;
- Le champ *First Party Hostnames* est renseigné avec le site à qui il a envoyé une requête, *www.20min.ch* ;
- Le champ *Third Parties* est vide puisque le site lui-même est en quelque sorte une partie tierce.

On voit ici un site tiers, fooby.ch, auquel www.20min.ch fait appel pour le chargement complet de sa page, par le biais d'un pixel caché ou d'un autre type de requête. Ainsi, chaque fois que l'utilisateur ira sur le site du 20min, le site fooby.ch sera averti de cette visite et pourra enregistrer des informations sur cet utilisateur, très probablement à son insu.

3.3.6. Situation 6

```
"htlb.casalemedia.com": {
  "hostname": "htlb.casalemedia.com",
  "favicon": "",
  "firstPartyHostnames": [
    "www.geeksforgeeks.org",
    "www.synonymo.fr"
  ],
  "firstParty": false,
  "thirdParties": []
},
"ib.adnxs.com": {
  "hostname": "ib.adnxs.com",
  "favicon": "",
  "firstPartyHostnames": [
    "www.geeksforgeeks.org",
    "www.synonymo.fr"
  ],
  "firstParty": false,
  "thirdParties": []
},
```

Figure 13 : Extrait du fichier *.json montrant deux sites tiers (*Hostname* : htlb.casalemedia.com et ib.adnxs.com), connus pour être des sites publicitaires, et les sites auxquels ils sont liés (*FPHN* : www.geeksforgeeks.org et www.synonymo.fr)

La Figure 13 est en lien avec la Figure 7 de la situation 1. Elle présente une situation qui peut paraître similaire à la Figure 12 si ce n'est que le champ *First Party Hostnames* contient cette fois deux sites. J'ai choisi de cibler deux *Third Parties* connues pour être des services publicitaires, à savoir htlb.casalemedia.com et ib.adnxs.com. Ils étaient tous deux présents comme *Third Parties* dans la Figure 7.

On les retrouve donc ici avec les spécificités suivantes :

- Le nom du site accédé est renseigné sous le champ *Hostname* : *htlb.casalemedia.com* ;
- Le champ *First Party* est renseigné à *false*, ce qui implique que l'utilisateur n'a pas volontairement visité ce site ;

- Le champ *Favicon* n'est pas renseigné car Lightbeam n'enregistre pas les icônes des sites tiers (ce sont les triangles blancs sur la Figure 4) ;
- Le champ *First Party Hostnames* est renseigné avec deux sites : *www.synonymo.fr* et *www.geeksforgeeks.fr* ;
- Le champ *Third Parties* est vide puisque le site lui-même est une partie tierce.

Les entrées sont les mêmes pour le deuxième site tiers mentionné ici, *ib.adnxs.com*.

La Figure 12 montrait une situation où le site tiers ne reçoit d'information que depuis un seul site primaire (*www.20min.ch*). La Figure 13 montre la situation où des sites tiers ont accès à des informations d'un utilisateur provenant de plusieurs sites primaires, ici *www.synonymo.fr* et *www.geeksforgeeks.fr*. C'est un hasard que l'utilisateur soit allé sur ces deux sites qui font appel aux mêmes agents publicitaires et c'est une aubaine pour ces mêmes agents. Ils peuvent ainsi récolter des informations au sujet de l'utilisateur depuis deux sources différentes et en plus ils peuvent s'échanger ces informations. Ils sont donc à même de proposer du contenu mieux ciblé pour cette personne.

La Figure 14 montre les très nombreux liens existants entre *synonymo.fr* (sans icône) et *geekforgeeks* (GG).

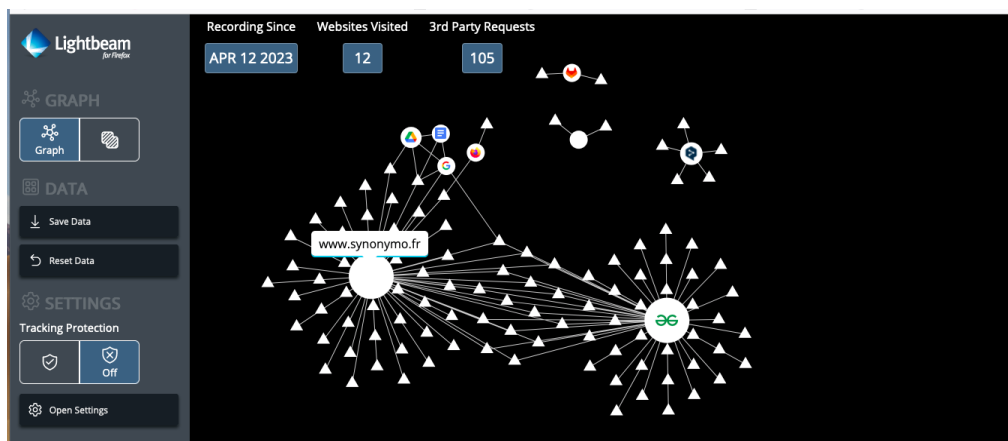


Figure 14 : Capture d'écran de Lightbeam montrant les liens entre *www.synonymo.fr* et *www.geekforgeeks.org*

Il faut garder en tête que tous les sites tiers contactés par le site primaire ne sont pas malveillants. Certains sont essentiels au bon fonctionnement du site primaire, certains sont là à des fins d'enregistrement de l'utilisateur à son compte (*account.google* par exemple mais aussi *account.booking* ou *secure.fnac*). L'extension Lightbeam ne permet pas un tri aisé des ces différents cas de figure. On peut toutefois raisonnablement penser que si l'on refuse systématiquement les cookies, que l'on purge son navigateur ou qu'on le ferme tout simplement, le trafic restant sera inévitable. C'est ce que je tente de mettre en lumière dans le chapitre suivant.

4. Résultats et Analyse de l'étude

Cette partie détaille les résultats de l'étude réalisée à l'aide des 10 participants. Pour rappel, j'ai demandé aux participants de surfer comme à leur habitude pendant une semaine tout en enregistrant leur expérience de surf à l'aide de Lightbeam ou Thunderbeam selon s'ils utilisent Firefox ou Google Chrome. Ils devaient quotidiennement m'informer des publicités qu'ils auraient reçues à la suite d'un éventuel achat. Au bout de trois jours, je leur ai demandé de visiter le site tripadvisor.ch et d'y effectuer une recherche de leur choix au sujet d'un restaurant, d'une ville, d'un voyage. Je leur ai demandé de rester attentif aux éventuelles publicités en lien avec leur recherche dans les jours qui ont suivis.

J'analyse dans ce chapitre les points suivants :

Le nombre de sites visités par rapport au nombre de sites tiers déposant des cookies dans les navigateurs des utilisateurs ;

Les sites visités par plusieurs participants et de leur comportement en terme de traçage vis-à-vis des utilisateurs ;

Quelques sites en particulier, connus pour pratiquer du traçage de surfeurs.

4.1. Nombre de sites visités vs. nombre de sites tiers

La Figure 15 utilise une échelle logarithmique et indique le nombre de sites volontairement visités par les testeurs (en vert) et le nombre de sites tiers (en rouge) contactés par les sites primaires. Les testeurs ont ici accepté les cookies (USER ID / O)

Nombre de sites visités vs. sites tiers avec des cookies acceptés

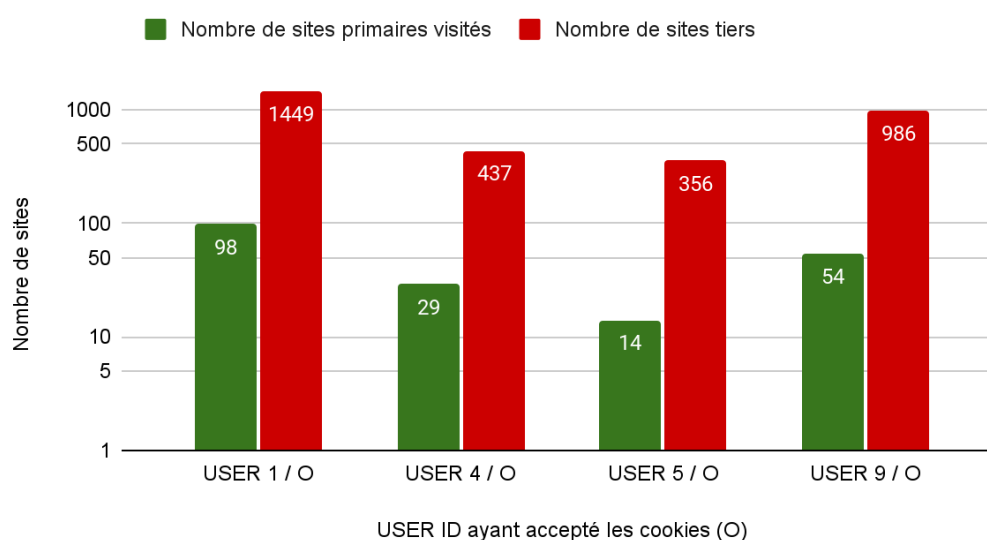


Figure 15 : Nombre de sites visités vs. nombre de sites tiers pour des utilisateurs ayant accepté les cookies.

Le rapport entre le nombre de sites visités a été calculé et montre que, pour un seul site visité entre 15 à 25 autres sites sont contactés pour le chargement complet de la page demandée ! Le plus grand rapport appartient à l'User 5. Ce dernier a peu visité de sites mais

ces sites ont contacté un grand nombre de sites tiers. Ceci tend à montrer que cet utilisateur est particulièrement envahi de demandes de sites tiers.

La Figure 16 montre que ce rapport diminue de moitié si l'utilisateur refuse les cookies (USER ID / N). Dans cette situation, entre 6 à 10 autres sites "seulement" sont chargés en même temps que la page demandée.

Nombre de sites visités vs. sites tiers avec des cookies refusés

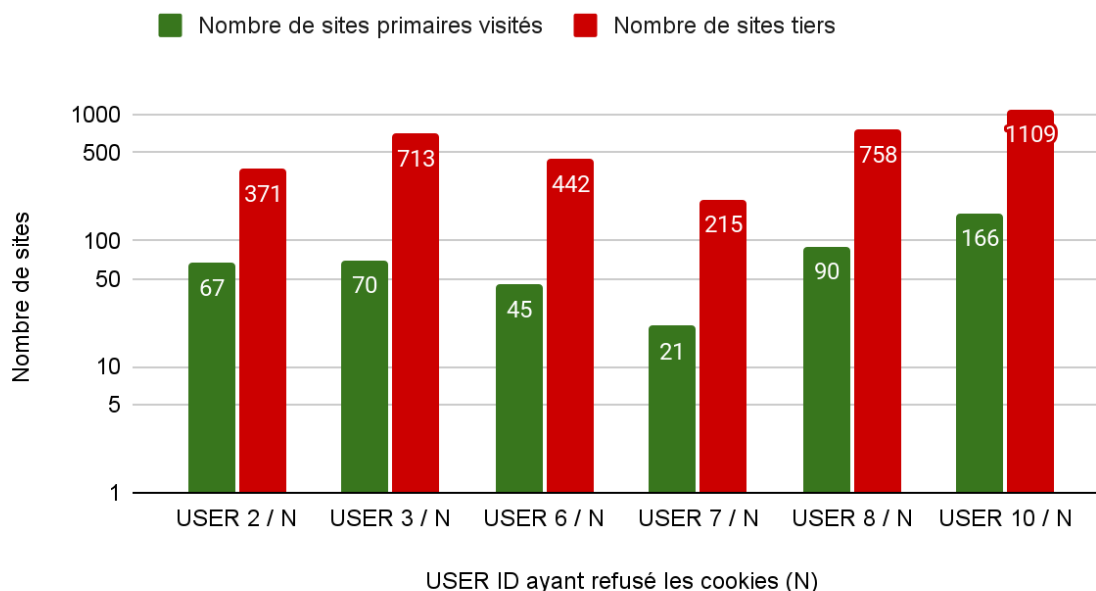


Figure 16 : Nombre de sites visités vs. nombre de sites tiers et rapport entre eux, pour des utilisateurs ayant refusé les cookies.

Ces valeurs sont données pour tous types de contacts établis en arrière-plan, qu'ils soient essentiels pour le fonctionnement de la page, pour l'obtention de données statistiques, pour des questions d'hébergement ou pour de la récolte de données à des fins de ciblage publicitaire.

Afin d'être en mesure de tenir compte des requêtes légitimes par rapport à de la publicité ou du traçage, j'ai éliminé certains sites du décompte final. Pour ce faire, je me suis basée sur le site [Whotracksme](#) (Ghostery GmbH, 2023) qui tient à jour un inventaire de nombreux sites et de leurs activités. On trouve ci-dessous une liste des catégories légitimes au bon fonctionnement d'un site et quelques exemples de sites entrant dans ces catégories.

Catégorie de sites	Fonction	Exemples
Hébergement	Serveur d'hébergement de site	Cloudfront, web-services
Essentiel	Inclut des gestions de balises permettant le fonctionnement du site, des politiques de confidentialité et toutes autres technologies nécessaires au bon fonctionnement du site.	Cookielaw, Onetrust, Googletagmanager, Googletagservices, Optanaon
Content Delivery Network	Ensemble de serveurs d'hébergement situés à des emplacements différents et mis en réseau via internet	Gstatic, Goggleapis, Cloudflare, Createjs, Jsdelivery
Statistique	Permet l'analyse du fonctionnement du site à l'aide de données collectées	Google Analytics,

En plus des sites ci-dessus, j'ai éliminé les adresses contenant les mots clé de type "consent", "login", "account", "bank", "pay" ou encore "documents", pour les raisons suivantes :

- consent : requête pour une demande d'accord ;
- login : requête pour une identification ;
- account : requête pour un accès à un compte ;
- bank : requête pour un accès ou une transaction bancaire ;
- pay : requête pour un paiement ;
- document : requête pour le chargement d'un document.

Dans un premier temps, j'avais également considéré le mot clé "secure" comme faisant référence à une source légitime d'identification sécurisée. Toutefois, en regardant plus en détail les adresses qui lui étaient associées, j'ai découvert que ce sont majoritairement des sites publicitaires qui l'utilisent, tels que "securepubads.g.doubleclick.net" ou "secure.adnxs.com". J'ai donc éliminé ce mot clé.

Les nouvelles valeurs sont représentées dans les Figures 17 et 18 pour les utilisateurs acceptant et refusant les cookies, respectivement.

Nombre de sites visités vs. sites tiers corrigé avec cookies acceptés

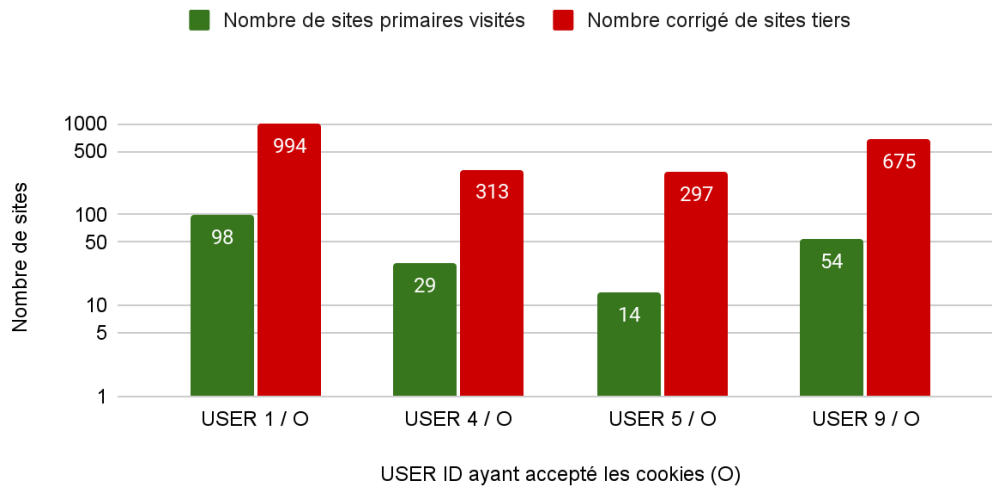


Figure 17 : Nombre de sites visités vs. nombre corrigé de sites tiers. Les utilisateurs ont ici accepté les cookies.

La Figure 17 utilise à nouveau une échelle logarithmique et montre le nombre de sites primaires (en vert), identique à ceux de la Figure 15, et le nombre corrigé de sites tiers (en rouge). Le rapport entre site tiers corrigé et site primaire a à nouveau été calculé et, en faisant abstraction du User 5, on observe que le rapport a diminué en passant de 15 - 18 à 10 - 13. Cela signifie que, pour la visite d'un site, entre 10 à 13 pages sont chargées en arrière-plan et que ces pages sont en lien avec du traçage ou de la publicité. Ainsi, on peut conclure qu'environ un tiers des sites contactés est essentiel au bon fonctionnement de la page tandis que deux tiers sont non essentiels.

Le navigateur du User 5 charge environ deux fois plus de pages que les autres utilisateurs, à savoir 21 pages par site visité contre 10, 11 ou 13 selon l'utilisateur. De plus, la prise en compte des catégories de sites n'a que peu modifié le rapport initial. En effet, il est passé de 25 à 21, ce qui implique que sur l'ensemble des sites tiers contactés (356) seuls 16,5% sont des sites essentiels (59) au fonctionnement des sites primaires visités.

J'explique cette différence par le genre de sites visités, relativement différents des autres utilisateurs. Le User 5 explore principalement des sites de vente en ligne de type anibis.ch. Ces sites de e-commerce ne demandant pas de frais pour déposer une annonce, dépendent majoritairement de la publicité qu'ils peuvent afficher. Ainsi, pour continuer à exister et à proposer leur services "gratuitement" ils ont un système de rémunération indirect via de la publicité ou, potentiellement, de la vente d'identifiants.

Nombre de sites visités vs. sites tiers corrigé avec cookies refusés

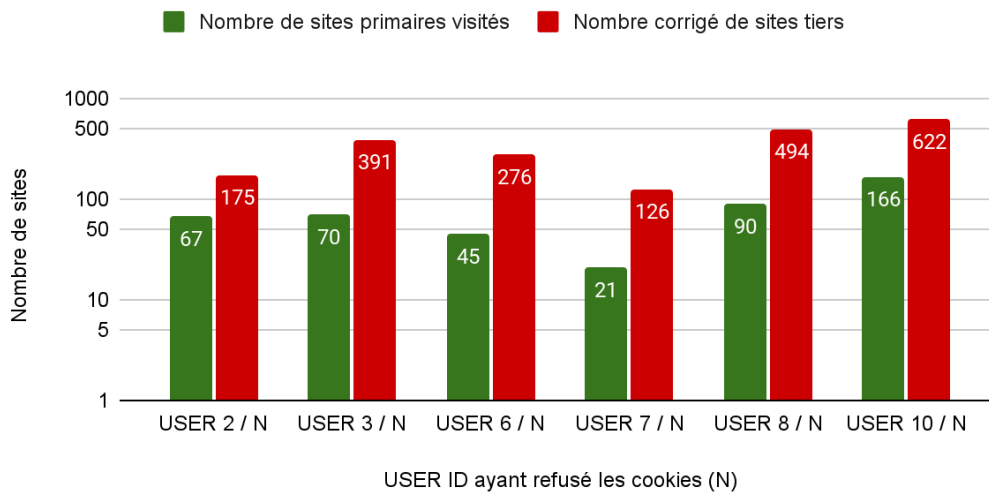


Figure 18 : Nombre de sites visités vs. nombre corrigé de sites tiers. Les utilisateurs ont ici refusé les cookies.

En ce qui concerne les utilisateurs ayant refusé les cookies, on observe que le nombre de sites contactés est presque deux fois moins grand lorsque l'on tient compte de leur catégorie.

Le nombre de sites issus de requêtes non voulues reste toutefois élevé puisqu'au mieux, il y a environ trois sites contactés pour un site demandé (User 2), au pire six contactés pour un site demandé (User 3, User 6 et User 7).

Afin de rendre le rapport entre page essentielle et non essentielle plus visuel, je l'ai représenté dans la Figure 19 ([en bleu clair](#)). Ce dernier montre que, mis à part pour l'User 2, le nombre de pages non essentielles est toujours supérieur au nombre de pages essentiels. Toutefois, lorsque l'utilisateur n'accepte pas les cookies, ce rapport est systématiquement inférieur à deux, tandis qu'il est systématiquement supérieur à deux si les cookies sont acceptés. Bien que la différence ne soit pas aussi grande qu'on l'aurait voulu, elle est bel et bien présente et peut en quelque sorte rassurer les utilisateurs qui refusent les cookies. Ils le font à bon escient et le trafic en arrière-plan en est amoindri. On remarquera encore la valeur pour l'User 5 (5,0), nettement supérieure à toutes les autres et montrant à nouveau la grande proportion de sites tiers non essentiels contactés par rapport aux sites essentiels.

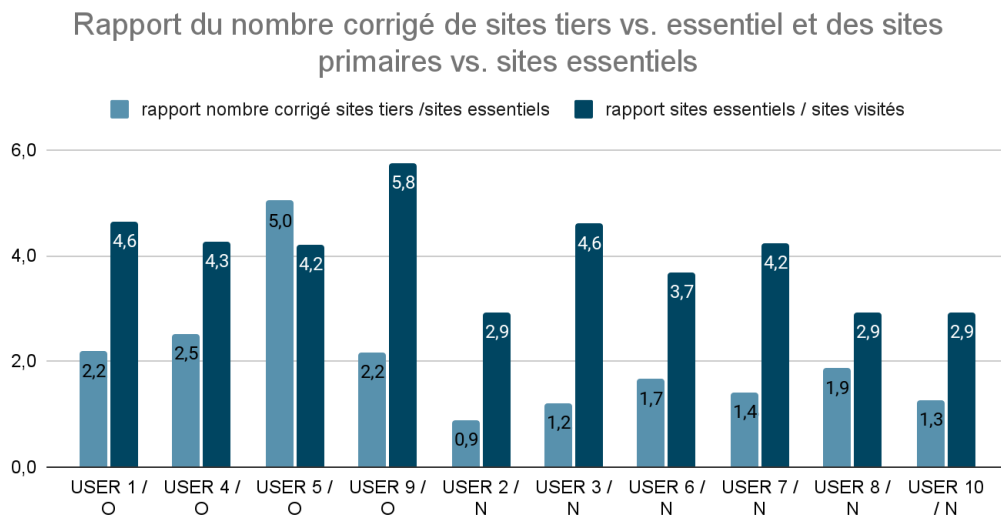


Figure 19 : Rapport entre nombre corrigé de sites tiers vs. sites essentiels et sites primaires vs. sites essentiels

Le deuxième rapport illustré sur la Figure 19 représente la proportion de sites essentiels contactés par site visité (en bleu foncé). On pourrait penser que cette valeur soit indépendante de l'acceptation ou non des cookies puisque les cookies essentiels sont, à priori, toujours acceptés. Les utilisateurs 4, 5, 6 et 7 montrent en effet des valeurs similaires proches de 4, signifiant que pour un site visité une moyenne de quatre sites sont contactés. Toutefois, les autres utilisateurs présentent des valeurs différentes. L'utilisateur 9 se voit connecté à presque six sites par site visité tandis que les utilisateurs 2, 8 et 10 ne se connectent qu'à trois sites par site visité. Les utilisateurs 1 et 3 sont contactés par 4,6 sites en moyenne.

Le nombre et le genre de sites visités ainsi que les actions qui y sont menées (achats, ventes, transactions bancaires, simples recherches) sont responsables du rapport entre sites essentiels et sites visités. Il est dès lors compliqué de tirer une conclusion claire et univoque. Ce qui peut être affirmé est qu'un utilisateur qui procède à de nombreux paiements en ligne aura un nombre de sites essentiels élevé. Les utilisateurs 1, 3 et 9 ont en effet procédé à de nombreux paiements en ligne. Toutefois, cette valeur peut être masquée selon le type et la diversité des sites sur lesquels il est allé surfer. Par exemple, l'utilisateur 10 a aussi procédé à de nombreux achats en ligne et son rapport entre sites essentiels et sites visités n'est que de 2,7. Cela provient du grand nombre de sites visités par cet utilisateurs (1109).

Idéalement on aimerait avoir uniquement des sites essentiels chargés en même temps que les sites primaires.

En conclusion,

- Refuser les cookies permet de diminuer le nombre de pages non essentielles contactées à notre insu. Le nombre de pages chargées en arrière-plan par site visité passe en moyenne de treize à cinq ;
- Le nombre de cookies, qu'ils soient essentiels ou non, dépend de ce que l'on fait lorsqu'on se rend sur la toile ;

- Le nombre de pages chargées dû à des requêtes non sollicitées est environ deux fois plus grand que le nombre de pages essentielles.

Dès à présent, lorsqu'un nombre de sites est évoqué, il s'agira de la valeur corrigée en tenant compte du nombre de sites essentiels ou acceptables.

4.2. Sites visités par plusieurs participants

Le Tableau 3 présente les sites visités par plusieurs testeurs de mon panel. Les conditions pour qu'un site apparaisse dans ce tableau est qu'au moins trois personnes aient surfé sur le site, qu'il ne soit pas en lien avec une banque ou un système de paiement et que la valeur de *First Party* du fichier Lightbeam / Thunderbeam soit *true*. Bien qu'on soit loin d'une valeur statistique, le fait de ne retenir que des sites visités par un tiers de mon panel permet de cibler les sites les plus populaires et au contraire d'éviter des sites "de niche". J'ai éliminé toute statistique en lien avec des paiements ou transactions bancaires pour des raisons évidentes de confidentialité et d'éthique.

Partant de l'hypothèse que plus un site est populaire plus il va pouvoir vendre son espace à des compagnies tiers, il m'a paru intéressant de connaître le nombre de sites tiers auxquels les sites primaires les plus visités par mon panel soumettent des requêtes en fonction de l'acceptation ou du refus des cookies. Seule la plus grande valeur enregistrée parmi les utilisateurs est présentée dans le tableau ci-dessous.

Hostname	# qui acceptent / # qui refusent	l'utilisateur accepte les cookie	l'utilisateur n'accepte pas les cookies
google.com/ch	4 / 6	55	75
fr.tripadvisor.ch	4 / 3	173	175
booking.ch	3 / 2	28	17
facebook.com	1 / 3	1	5
youtube.com	2 / 2	17	12
deepl.com	0 / 3	NA	0
lematin.ch	2 / 1	191	24
linkedin	1 / 2	8	12
local.ch	0 / 3	NA	33
qoqa.ch	2 / 1	21	25
wikipedia.ch	0 / 3	NA	0

Tableau 3 : Nombre de parties tierces contactées par le site primaire

Parmi les sites visités par les participants, celui qui ressort le plus souvent, hormis google.ch ou google.com (10 personnes sur 10), est fr.tripadvisor.ch (7 personnes sur 10), ce qui est normal puisque je leur avais demandé de procéder à une recherche spécifique sur ce site. Pour rappel, j'ai demandé aux participants de noter s'ils recevaient de la publicité en lien avec ce qu'ils ont cherché sur Tripadvisor ou avec un éventuel achat en ligne effectué les jours précédents.

Selon les réponses reçues, les utilisateurs qui n'acceptent pas les cookies ne voient pas de publicités en lien avec leur recherche sur Tripadvisor ou avec d'éventuels achats, tandis que les utilisateurs qui acceptent les cookies ont vu, par exemple, de la publicité pour des compagnies aériennes sur des sites comme lematin.ch, doodle.com ou encore 20min.ch après avoir effectué une recherche pour des vols d'avion sur le site de Tripadvisor. Je me serais donc attendue à ce que le nombre de sites contactés par Tripadvisor soit très différent entre une personne acceptant les cookies et une autre les refusant. Ce n'est pourtant pas ce que montre le tableau. En me penchant sur les résultats individuels de chacun, je me suis aperçue que les deux autres utilisateurs refusant les cookies ont chacun 21 et 45 sites contactés en arrière-plan. J'ai soupçonné l'utilisateur d'avoir exceptionnellement accepté les cookies au moment de faire sa recherche sur Tripadvisor, raison qui aurait amené cette valeur (175) si proche de la valeur maximale des utilisateurs acceptant les cookies.

Pour vérifier cette hypothèse, j'ai voulu reproduire une recherche en acceptant et en refusant les cookies. Or, je me suis aperçue que Tripadvisor ne demande pas de consentement pour les cookies. Ainsi lors d'une recherche sur ce site, mon navigateur a contacté 74 sites en arrière-plan ! Refusant systématiquement les cookies lorsqu'on m'en donne l'occasion, je n'ai pas observé de publicité en lien avec ma recherche.

J'en conclus que la grande disparité de valeurs vient de l'expérience de surf elle-même. Un utilisateur cliquant sur plusieurs liens ou images proposées par Tripadvisor se verra connecté à plus de sites tiers qu'une personne qui ne fait que se rendre sur le site.

On constate que la gestion des cookies est très différente d'un site à l'autre. En effet, pour des sites comme linkedin.ch, booking.ch, qoqa.ch ou encore youtube.com, le fait d'accepter ou non les cookies ne change pratiquement rien au nombre de sites qui sont contactés par le site primaire, tandis que les valeurs changent drastiquement en ce qui concerne lematin.ch. De prime abord on pourrait penser que cela ne signifie pas pour autant que ces sites ne respectent pas la demande de l'utilisateur puisque, dans le cas de Tripadvisor en tout cas, celui-ci ne reçoit de la publicité ciblée que lorsqu'il accepte les cookies. Malgré cela, la méfiance est de mise. En effet, sans pouvoir le prouver mais simplement par contrôle des types de sites tiers contactés même suite au refus des cookies, il se pourrait bien que les données de l'utilisateur soient envoyées à ces sites et qu'elles dorment dans leurs bases de données, en attendant qu'un jour peut-être cet utilisateur accepte les cookies.

D'autres valeurs du Tableau 3 sautent aux yeux. Celles obtenues pour Facebook par exemple. Elles sont si faibles qu'elles n'en paraissent pas croyables. À mon sens, cela montre que les utilisateurs n'ont pas surfé sur Facebook. Dès lors, comment Facebook peut apparaître sous la valeur *true* pour *First Party* ? Le fait que Facebook a pu déposer un cookie de site primaire, ferait penser qu'il peut s'auto-charger ou qu'il est déjà passé à une nouvelle manière de cibler les utilisateurs ? Si tel est le cas, cela fait froid dans le dos... Je tenterai de trouver quelque chose à ce sujet en scannant Facebook à l'aide de l'outil OpenWPM, expliqué dans le Chapitre 5.

Enfin, des sites comme wikipedia ou deepl.com ne demandent des contenus qu'à des adresses leur appartenant, raison pour laquelle j'ai rapporté la valeur de 0 dans le Tableau 3. Ces chargements représentent une certaine quantité de trafic internet mais pas de parties tierces.

4.3. Compagnies publicitaires et sites tiers majoritaires

Pour terminer, j'ai cherché à savoir quels sont les sites qui obtiennent le plus d'informations sur un utilisateur. Il y a trois situations intéressantes présentées dans le Tableau 2, dont voici un extrait ci-dessous, qui permettent de répondre à cette question : les situations 3, 5 et 6.

#	<i>First Party</i>	<i>Hostname</i>	<i>First Party Hostnames (FPHN)</i>	<i>Third Parties</i>
3	True	google.com	multiple noms de sites	adresse(s) de third parties
5	False	fooby.ch	20min.ch	néant
6	False	htlb.casalemedia.com	multiple noms de sites	néant

Extrait du Tableau 2 : Toutes les configurations possibles enregistrées dans le fichier json issu de Lightbeam ou Thunderbeam

Dans la situation 3, j'ai isolé les résultats de Lightbeam avec la paire clé-valeur *First Party = true* puis avec un grand nombre d'entrées sous le champ *First Party Hostnames (FPHN)*. Dans une telle situation, le site renseigné sous le champ *Hostname* est le site accédé par l'utilisateur et les sites sous *FPHN* sont ceux qui ont envoyé des données au site primaire. Parmi ceux-ci on trouve bien évidemment www.google.ch ou www.google.com ainsi que account.google.com. Bien que j'aie retiré du décompte des sites tiers le terme "account" il n'en reste pas moins que, lorsqu'on s'identifie à un site lambda à l'aide de son identifiant Google, cette dernière sait automatiquement où nous avons des comptes et peut ainsi regrouper nos centres d'intérêt. Google garde ces informations pour elle mais elle possède également des entreprises partenaires spécialisées dans la publicité, comme DoubleClick, 2mdn.net, invitemedia.com, [GoogleSyndication](http://GoogleSyndication.com) (Ghostery GmbH, 2023), ce qui lui permettra de valoriser nos données.

Passons aux situations 5 et 6 du Tableau 2. Elles sont plus dérangeantes que la situation qui vient d'être décrite, car ici, l'utilisateur n'est pas allé sur le site tiers mais seulement sur le site primaire, qui dans sa page, cache un pixel ou arbore de la publicité ou une icône. Au moment de charger la page, le site demande au navigateur d'envoyer une requête au site caché qui recevra un cookie permettant l'identification de l'utilisateur. Si le site tiers est présent sur plusieurs sites primaires (cas 6 du Tableau 2), il pourra récolter les données de l'utilisateur à travers ces différentes plateformes, s'il n'est présent que sur un seul site (cas 5 du Tableau 2), seul ce site aura accès à ses données. Toutefois, rien ne garantit que, lors d'une expérience de surf ultérieure, l'utilisateur ne va pas retomber une nouvelle fois sur ce même site tiers, auquel cas ses données pourront être regroupées.

Je parle de "sites tiers" mais dans la majeure partie des cas, on ne peut même pas accéder au site en question. En effet, l'adresse ne correspond pas à un site conventionnel mais à

une plateforme publicitaire de type Supply Side Platform (SSP). Pour rappel et comme détaillé dans la section 2.6.5, le site s'adresse à une SSP pour vendre son espace publicitaire car il doit gagner de l'argent pour vivre. En réalité, il ne s'adresse pas qu'à une SSP mais à plusieurs, dans l'idée de maximiser son profit. Du côté des SSP c'est pareil, plus une SSP aura de clients plus elle pourra garantir la vente à bon prix de l'espace publicitaire. Ces faits sont connus du développeur du site visité mais probablement pas du visiteur lui-même qui, même s'il peut voir des publicités défiler, un pouce en l'air ou une autre icône, ne sait peut-être pas que ses données sont envoyées à des compagnies tiers. Je rappelle que l'apparente gratuité du net ne peut être perpétuée qu'à ce prix.

Afin de me rendre compte de qui sont les acteurs qui nous tracent, j'ai, pour chaque utilisateur, dans un premier temps, isolé les trois compagnies publicitaires qui reçoivent le plus de cookies depuis les sites primaires, puis dans un deuxième temps, repéré le site tiers qui se cache derrière le plus de sites primaires.

Pour la première situation, j'ai isolé les noms de domaine dont la paire clé-valeur *First Party* = *false* puis j'ai scanné à travers toutes les entrées sous *Hostname* en utilisant des noms de compagnies publicitaires répertoriés dans la liste est la suivante, basée selon leur appartenance à la catégorie de publicitaires par le site Whotracksme (Ghostery GmbH, 2023) :

Doubleclick	Smartadserver	2mdn	Bing	Everest
Googlesyndication	Pixel	Demdex	Adgrx	Criteo
Braintreegateway	Rubiconproject	Casalemedia	Indexww	Teads
Adsafeprotected	Outbrain	Taboola	Bluekai	360yield
Amazon-Adsystem	Adnxs	Yahoo	Doubleverify	Pubmatic
Adservices	Qualtrics	Tinypass	Krxd	Adnz

Dans la seconde situation, j'ai à nouveau considéré uniquement les noms de domaine dont la paire clé-valeur est *First Party* = *false* puis j'ai gardé l'adresse commençant par "www" dénombrant le plus d'entrées sous *FPHN*. En gardant uniquement les adresses commençant par "www" je m'assure de garder des sites bien réels et pas des plateformes de publicité.

Les résultats sont présentés dans les Tableaux 4 et 5.

4.3.1. Liens entre les sites primaires *via* les compagnies publicitaires

USER / cookie	Compagnies publicitaires		
USER 1 / O	Googlesyndication	DoubleClick	Adnxs
USER 4 / O	DoubleClick	Googleadservice / Adnxs	Googlesyndication
USER 5 / O	Criteo	DoubleClick	Googlesyndication
USER 9 / O	DoubleClick	Criteo	Googlesyndication
USER 2 / N	Googleadservice	Adnxs	Criteo / Taboola
USER 3 / N	DoubleClick	Googlesyndication / Googleadservice	Criteo
USER 6 / N	DoubleClick	Googleadservice	Criteo
USER 7 / N	DoubleClick	Googlesyndication	Adnxs
USER 8 / N	Googlesyndication	DoubleClick	Criteo
USER 10 / N	Googlesyndication	DoubleClick	Googleadservice

Tableau 4 : Compagnies publicitaires les plus présentes chez les utilisateurs

Le Tableau 4 permet de constater tout d'abord que ce sont souvent les mêmes entreprises publicitaires qui reviennent. Cela montre qu'elles sont bien implantées et à elles trois, Googlesyndication, DoubleClick et Googleadservices, inondent le marché. Ce qui est intéressant à souligner est qu'elles appartiennent toutes les trois à différents secteurs de Google (GoldSparrow, 2021).

Revenons un instant aux observations de "mes" testeurs : l'un d'eux a rapporté avoir remarqué des publicités pour des vols d'avion sur doodle.com après avoir fait des recherches sur fr.tripadvisor.ch. Ces deux sites font appels à plusieurs compagnies, dont notamment DoubleClick (Wikipedia - DoubleClick, 2023), propriété de Google, Googleadservice, Criteo, une compagnie française (Wikipedia - Criteo, 2022), mais également Bing (propriété de Microsoft) (Microsoft, 2023), partenaire absent du Tableau 4 mais uniquement en raison de l'écrasante présence des trois autres. Deux autres compagnies sont également apparues sous l'appellation "pixel.adsafeprotected.com" et "pixel.rubiconproject.com" pour doodle.com et fr.tripadvisor.ch. Adsafeprotected appartient à Integral Ad Science (*INTEGRAL AD SCIENCE*, 2023) et RubiconProject appartient à Magnite (Wikipedia - Magnite, 2023), une des plus grandes plateformes de vente SSP (*The Largest Independent Sell-Side Ad Platform*, 2023).

La difficulté réside dans le fait d'identifier quel acteur fait quoi. En effet, bon nombre de ces entreprises se sont diversifiées et proposent leur service aussi bien en tant que SSP ou AdExchange, comme cela s'est vu dans les listes établies dans la section 2.6.5.

J'ai repris la Figure 3 pour illustrer les processus d'échange ayant lieu entre les deux sites primaires doodle.com et fr.tripadvisor.ch. J'ai utilisé quelques compagnies communes aux deux sites, trouvées grâce à Lightbeam. J'ai choisi ces sociétés car elles sont prédominantes dans ma recherche.

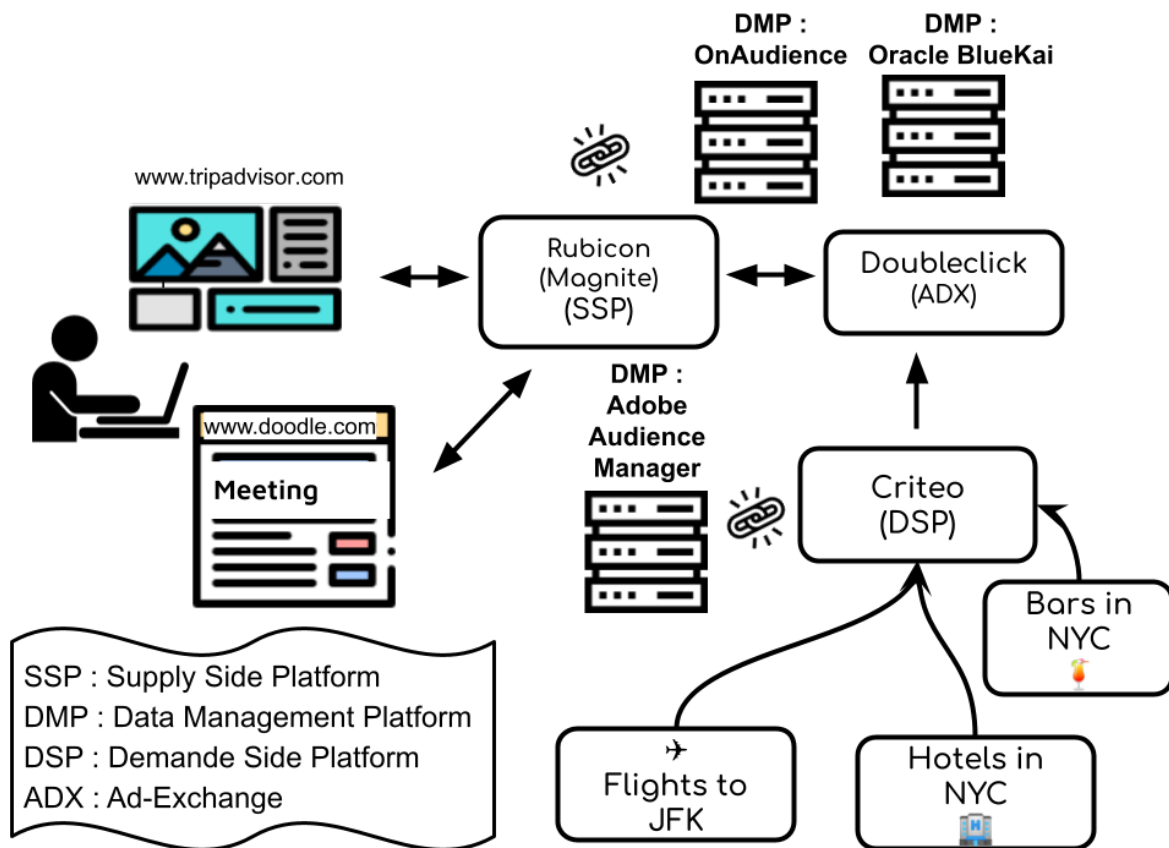


Figure 20 : Exemple de flux de données

Après qu'un utilisateur a parlé de publicité pour des vols apparaissant sur lematin.ch, j'ai pu épingler les mêmes parties tierces que celles vues sur Tripadvisor. Ces parties tierces sont donc à même de proposer du contenu ciblé également sur lematin.ch.

Finalement, un utilisateur a rapporté avoir vu de la publicité pour un site de bricolage sur lematin.ch et 20min.ch après avoir recherché des offres pour des caillebotis sur google.ch. Pour rappel, un lien entre les deux sites d'information a déjà été relevé dans la Figure 8 de la section 3.3.2. On y voit aussi la présence de DoubleClick, appartenant à Google, preuve que ces trois sites communiquent. Doubleclick opère sous le nom de Google AdX est un

Un autre utilisateur parle encore de publicités ciblées apparaissant sur son compte Instagram (sur son smartphone) après avoir fait des recherches sur des vêtements depuis son ordinateur. Ici c'est instagram qui est présent sur le site de vêtements initial.

Sans surprise mais avec plus de connaissances qu'auparavant, je peux conclure que si un utilisateur accepte les cookies, il s'expose à de la publicité ciblée, ce qui peut ne pas le

déranger mais face à l'insistance de laquelle il pourrait finir par être influencé dans ses choix. Ce qui est toutefois rassurant c'est que le choix de l'utilisateur qui refuse la publicité est respecté. Cela ne signifie pas en revanche qu'il n'est pas traqué. Comme déjà mentionné, on pourrait imaginer que différentes compagnies l'auraient quand même enregistré dans leur base de données, à l'affût d'un éventuel changement d'avis de l'internaute, auquel cas elles ne manqueraient pas l'occasion de lui soumettre quelques images adaptées à ses préférences et ses centres d'intérêt. Le nombre de sites tiers accédés en arrière-plan ne changeant guère que l'on accepte ou refuse les cookies me fait pencher pour cette hypothèse.

4.3.2. Sites tiers contactés par un grand nombre de sites primaires

USER / cookie	Site tiers
USER 1 / O	www.facebook.com (28) / www.youtube.com (12)
USER 4 / O	www.facebook.com (10) / www.google.ch (9)
USER 5 / O	www.facebook.com (4)
USER 9 / O	www.google.ch (34) / www.facebook.com (26)
USER 2 / N	www.facebook.com (3) / www.youtube.com (3)
USER 3 / N	www.facebook.com (18) / www.google.ch (22)
USER 6 / N	www.youtube.com (4)
USER 7 / N	www.facebook.com (5)
USER 8 / N	www.youtube.com (9)
USER 10 / N	www.facebook.com (34) / www.google.ch (41)

Tableau 5 : Site tiers recevant des informations depuis de nombreux sites primaires (la valeur entre parenthèses correspond au nombre de sites primaires envoyant de requêtes au site tiers)

En ce qui concerne les sites tiers à qui le navigateur a envoyé une requête (Tableau 5), on peut voir que ce sont principalement, pour ne pas dire exclusivement des réseaux sociaux. Facebook est le grand vainqueur en étant présent chez 10 utilisateurs sur 10, puis vient YouTube, présent chez 4 personnes sur sur 10. Google est évidemment une fois de plus présent. J'ai été étonnée de ne pas voir les autres réseaux sociaux qu'on trouve en bas de page, en général au côté de Facebook et YouTube. Après une analyse plus pointue des données, je me suis aperçue que certains autres réseaux sont bel et bien présents mais pas directement sous leur adresse "www" mais sur leurs sites d'analyse ou de publicité (analytics.tiktok, analytics.twitter, ct.pinterest.com, ads.linkedin), ce qui m'avait fait les rater en première analyse. Les résultats sont présentés dans le Tableau 6.

USER ID / cookie	analytics.tiktok. com	analytics.twitter.com	ct.pinterest.com	ads.linkedin.com
USER 1 / O	9	9	7	12
USER 4 / O	5	0	3	1
USER 5 / O	1	1	2	1
USER 9 / O	4	3	7	6
USER 2 / N	1	0	1	2
USER 3 / N	0	0	0	5
USER 6 / N	5	1	4	4
USER 7 / N	0	0	0	3
USER 8 / N	0	3	1	6
USER 10 / N	1	5	7	7

Tableau 6 : Site tiers recevant des informations depuis de nombreux sites primaires. La valeur dans les colonnes correspond au nombre de sites primaires envoyant de requêtes au site tiers.

On constate que LinkedIn est présent chez les dix participants à l'étude, Pinterest chez 8 participants, puis viennent TikTok et Twitter avec une présence chez respectivement 7 et 6 participants. La présence ou non de ces réseaux sociaux comme traceurs dépend évidemment du genre de sites sur lesquels l'utilisateur va surfer. En comparant les résultats des Tableaux 5 et 6, on s'aperçoit que le nombre de sites primaires vers qui ces réseaux sociaux envoient des requêtes reste relativement faible en comparaison de Facebook et Youtube. Il n'en est pas moins intéressant de constater que LinkedIn, que l'on pourrait croire une plateforme respectueuse de la vie privée puisque en lien avec des relations de travail, est bien représentée ici. Peut-être est-ce un biais de mon panel de participants puisque toutes ces personnes ont plus de 25 ans et travaillent. Il serait intéressant de vérifier si cette tendance est la même dans un panel d'utilisateurs plus jeunes. Ce que je pourrai tester, dans le futur, avec les gymnasiens.

5. Études sur sites spécifiques

Dans ce chapitre, je décris l'utilisation d'un autre outil de suivi, OpenWPM, développé par l'Université de Princeton et je compare ces résultats avec Lightbeam.

5.1. Introduction à OpenWPM

OpenWPM est un outil informatique développé par l'université de Princeton (Englehardt & Narayanan, 2016). Il permet d'étudier les pratiques de collecte de données et de suivi subis par les utilisateurs du web.

Open WPM permet d'imiter un utilisateur surfant sur un site. Après configuration, l'outil va charger la page web, "cliquer" sur des liens ou des images présents sur cette page, permettre à l'utilisateur de s'identifier, d'accepter ou de refuser les cookies. En fin de surf, l'outil enregistre tout le trafic en lien avec ce site dans une base de données contenant plusieurs tables. On peut configurer OpenWPM pour "surfer" sur un seul site à la fois ou sur plusieurs. Chaque nouveau surf est ajouté à la base de données à moins qu'on ne change spécifiquement de fichier de sortie.

J'ai principalement utilisé cinq Tables sur les 14 proposées, à savoir les Tables "site_visit", "javascript_cookie", "http_request", "http_response" et "http_redirect".

Dans la Table "site_visit" on trouve l'identifiant du navigateur (browser_id) en lien avec le site visité (site_url).

La Table "javascript_cookie" est un extrait de l'en-tête qui accompagne une requête. Elle en permet une lecture plus aisée. On y trouve notamment le nom et la valeur des cookies déposés par les sites internet. Ces sites peuvent être ceux visités lors du surf de OpenWPM ou des sites tiers, ayant fait une requête au site visité en échange du chargement de sa page. Le nom du cookie peut parfois donner des renseignements sur l'entreprise propriétaire du cookie, par exemple le cookie _ga appartient à Google Analytics (La Rédaction JDN, 2019) et _fb à Facebook (CookieDatabase, 2023). La valeur du cookie pourra être utilisée pour chercher des entrées dans les autres Tables et voir quel(s) autre(s) site(s) utilise(nt) le même cookie et partage(nt) donc des informations sur un utilisateur.

Dans les Tables "http_request" et "http_response", on trouve les URL's des sites internet et les en-têtes associés à la requête / réponse.

Dans la Table "http_redirect", on trouve l'URL du site qui demande une redirection, l'URL du site vers qui la redirection pointe ainsi que l'en-tête associé.

Dans ces trois Tables on trouve également une identité de requête (request_id) qui permet de lier la réponse ou la redirection à la requête. Cette valeur est générée par OpenWPM en fonction de l'ordre dans lequel le navigateur reçoit les requêtes.

5.2. Terminologie utilisée dans OpenWPM

Le Tableau 7 résume les différents en-têtes de colonnes trouvés dans OpenWPM

Nom de l'en-tête	Fonction
<i>Browser ID</i>	Identifiant du site visité lors d'un surf donné, conservé à travers toutes les table et propre à OpenWPM
<i>Host</i>	Site à qui appartient le cookie
<i>Cookie name</i>	Nom du cookie
<i>Cookie value</i>	Valeur du cookie
<i>Request id</i>	Numéro d'identification de la requête
<i>Old request url</i>	URL de la requête d'origine
<i>Old request id</i>	Numéro d'identification de la requête d'origine, identique à Request id
<i>New request url</i>	Nouvelle URL (dans le cas d'une redirection)

Tableau 7 : Terminologie utilisée par OpenWPM

5.3. Études de cas

J'ai configuré OpenWPM pour surfer sur trois sites en particulier, à savoir www.lematin.ch, www.tripadvisor.ch et www.doodle.com afin de montrer les liens entre eux. J'ai choisi ces sites pour les raisons suivantes :

- Tripadvisor : Tous les participants y sont allés, puisque c'était une requête de ma part ;
- lematin.ch : D'une manière générale, les sites d'informations utilisent la publicité comme source de revenu. En particulier, lematin.ch échange des informations avec 191 sites tiers lorsque la personne accepte les cookies (voir Tableau 3, section 4.2), ce site m'a paru donc intéressant à étudier plus en détail ;
- doodle.com : Ce site est à l'origine d'observation de publicité ciblée par l'un des utilisateurs.

J'ai également étudié les surfs à travers le temps pour le site [lematin.ch](http://www.lematin.ch) afin de déterminer si les cookies sont conservés ou non.

En ce qui concerne [doodle.com](http://www.doodle.com) et [lematin.ch](http://www.lematin.ch), j'ai réalisé deux surfs. Lors du premier surf, j'ai accepté les cookies et lors du second, je les ai déclinés. Tripadvisor ne propose pas d'accepter ou non des cookies.

Par souci de clarté, pour toutes les Tables, je n’ai rapporté que les colonnes pertinentes ainsi que quelques lignes ciblées pour leur intérêt de comparaison avec Lightbeam.

5.3.1. Surf sur le site lematin.ch à travers le temps

	Browser ID	Host	Cookie name	Cookie value
4 juillet 2023	3657472016 (lematin.ch, Y)	lematin.ch	_ga_7P3NL6HKTL	GS1.1.1688477209.1.1.1688477240.0.0.0
8 juin 2023	1312132910 (lematin.ch, Y)	lematin.ch	_ga_7P3NL6HKTL	GS1.1.1686223942.1.1.1686223973.29.0.0
8 juin 2023	1468245175 (lematin.ch, N)	lematin.ch	_ga_7P3NL6HKTL	GS1.1.1686224030.1.1.1686224063.27.0.0
1 juin 2023	1152581835 (lematin.ch, Y)	lematin.ch	_ga_7P3NL6HKTL	GS1.1.1685623452.1.1.1685623493.19.0.0

Tableau 8 : Extrait de la Table “javascript_cookie” de OpenWPM pour trois surfs sur lematin.ch effectués sur plusieurs jours non consécutifs

On peut voir dans le Tableau 8 que le nom du cookie appartenant à Google Analytics est conservé à travers le temps. C’est le cas, que l’on accepte ou refuse les cookies. En filtrant la Table “http_request” pour le surf exécuté le 8 juin avec le nom du cookie, _ga_7P3NL6HKTL, cela permet de mettre en évidence 52 entrées. Toutes les entrées ont, en URL, l’adresse du matin.ch. Toutefois, si on pousse plus loin l’analyse, on peut voir que dans certaines de ces URL’s se cachent plusieurs valeurs de cookie, identifiables par la balise “&..id=...”. On les découvre dans le Tableau 9. J’ai à nouveau tronqué les adresses par souci de clarté. Ainsi, le nombre de cookies déposés dans le navigateur n’est pas équivalent au nombre de sites contactés.

Request id	URL
168	https://sst.lematin.ch/g/collect?v=2&tid=G-7P3NL6HKTL>m=45he3650&_p=137628467&_gaz=1&cid=1401536577.1686223942&ul=en-us&sr=1440x900&_fplc=0&sst.uc=&_s=1&dt=Le%20Matin%3A%20toute%20l%27actualit%C3%A9%20en%20Suisse%20et%20dans%20le%20monde&sid=1686223942 [...]

Tableau 9 : Extrait de la Table “http_request” de Open WPM

En filtrant la même Table à l’aide de l’un de ces cookies (cid=1401536577) , on retrouve 9 URL’s pointant vers la compagnie publicitaire DoubleClick et son partenaire google et à l’aide d’un autre cookie (sid=168223942), on tombe sur des URL’s pointant à nouveau vers Doubleclick et Google mais également Facebook. Ainsi donc, lorsque la page du matin.ch est chargée, des informations sur l’utilisateur sont envoyées à ces diverses entreprises ou réseau. Ce qui confirme les observations faites grâce à Lightbeam.

5.3.2. Surf sur différents sites

	Browser ID	Host	Cookie name	Cookie value
1.	1312132910 (lematin.ch, Y)	.lematin.ch	_fbp	fb.1.1686223942369.2045106199
2.	1312132910 (lematin.ch, Y)	.20min.ch	_ga_WSBS1PY35W	GS1.1.1686223943.1.0.1686223943.60.0.0
3.	1312132910 (lematin.ch, Y)	.casalemedia.com	CMID	ZIG8Ti2VXxfv5a2HYHsw.QAA
4.	1312132910 (lematin.ch, Y)	.pubmatic.com	KADUSERCOOKIE	06D27A52-C6CC-428C-BFF1-D07DECCA2DE6
5.	1468245175 (lematin.ch, N)	.20min.ch	_ga_WSBS1PY35W	GS1.1.1686224032.1.0.1686224032.60.0.0

Tableau 10 : Extrait de la Table “javascript_cookie” de OpenWPM pour lematin.ch

Le Tableau 10 présente des noms et valeurs de cookies lorsque le site lematin.ch est visité. Les lignes 1 à 4 montrent une navigation durant laquelle les cookies ont été acceptés (lematin.ch, Y), tandis que la ligne 5 montre une navigation avec des cookies refusés (lematin, N).

Les lignes 2 à 4 révèlent que lors du chargement de la page une demande est aussi envoyée à des sites différents de lematin.ch. J’ai choisi ces autres sites car ils représentent une entreprise publicitaire, un réseau social et un autre site d’information affilié à lematin.ch. À nouveau, en comparant la ligne 2 et la ligne 5, on peut voir que le cookie associé à Google Analytics est conservé, que les cookies aient été acceptés ou non.

La ligne 1 présente un cookie Facebook. On remarquera que le début de ce cookie est le même que celui présenté dans le Tableau 9 ([1686223942](#)), mais également dans le Tableau 8 (ligne 2). Lorsque la valeur complète du cookie est entrée dans la Table “http_request”, on trouve 28 entrées dont 3 ont pour URL facebook.com. Le type de ressource chargée est une image. Après vérification sur le site, il s’avère qu’il s’agit d’un pixel invisible. L’utilisateur n’a donc aucun moyen de savoir que ses données sont collectées par Facebook.

Lorsque la valeur du cookie Casalemedia ([ZIG8Ti2VXxfv5a2HYHsw.QAA](#)) est entrée dans la Table “http_request”, il en résulte 14 entrées dont aucune n’a pour URL lematin.ch. Les URL’s sont toutes en lien avec casalemedia, doubleclick, indexwww, amazon-adsystem et yahoo. Il en va de même lorsque cette même valeur est entrée dans la Table “http_response”. Ceci met en lumière le partage d’information qui a lieu entre ces entreprises, tout comme c’était le cas pour Tripadvisor et corrobore à nouveau ce qui avait été observé avec Lightbeam.

On ne retrouve pas directement la valeur du cookie dans la Table “http_redirect”. Toutefois, il existe 11 demandes de redirections lorsque les cookies sont déclinés et 33 lorsqu’ils sont acceptés. Pour rappel, une redirection signifie que, lorsque le navigateur a demandé au site lematin.ch de pouvoir charger sa page, le serveur a répondu en incluant également une demande de redirection vers un autre site. Ceci indique que certains sites profitent du surf

d'un utilisateur vers lematin.ch pour envoyer des informations à d'autres serveurs. Ainsi, comme illustré dans le Tableau 11, on voit que Casalemedia et DoubleClick s'échangent mutuellement des informations sur l'utilisateur.

La colonne "old_request_id" permet de lier ces redirections à la requête d'origine et facilite le traçage à travers les différentes Tables de OpenWPM.

old request url	old request id	new request url
https://dsum-sec.casalemedia.com/rsum?ixi=1&cm_dsp_id=85&cb=https%3A%2F%2Fcm.g.doubleclick.net%2Fpixel%3Fgoogle_nid%3Dcasale_media2_dbm%26google_cm%26google_sc%26google_hm%3D	327	https://cm.g.doubleclick.net/pixel?google_nid=casale_media2_dbm&google_cm&google_sc&google_hm=ZIG8rXOAF65hn2pseTdsgAA

Tableau 11 : Extrait de la Table "http_redirect" de OpenWPM pour lematin.ch

	Browser ID	Host	Cookie name	Cookie value
1.	1642466324 (tripadvisor.com)	.adnxs.com	uuid2	7559368050676657941
2.	1642466324 (tripadvisor.com)	.fr.tripadvisor.ch	TASID	0EB5C5C921B742849D11F8BEF7865F07

Tableau 12 : Extrait de la Table "javascript_cookie" de OpenWPM pour tripadvisor.ch

Le Tableau 12 (ligne 1) montre qu'un des sites à qui Tripadvisor envoie une requête est adnxs, identifié comme distributeur de publicité sur internet (Ghostery GmbH, 2023), appartenant maintenant à Microsoft (Hercher, 2021). Lorsque la valeur du cookie est entrée comme filtre dans la Table "http_request", on trouve 11 entrées dont les URL's contiennent adnxs, pubmatic (PubMatic, 2023) ou casalemedia. Ainsi la communication entre ces entreprises est bien présente.

D'une manière peut-être plus surprenante car moins visible de prime abord, lorsque la valeur du cookie que l'on trouve à la ligne 2 est entrée dans la Table "http_request", on trouve 51 entrées dont 16 ont tripadvisor.com comme URL's et le reste des URL's aussi diverses que jscache, Criteo, bat.bing ou DoubleClick pour ne citer que celles-ci ! Tandis que Criteo, bat.bing et DoubleClick sont des entreprises publicitaires connues et dont nous avons déjà discuté, l'adresse jscache.com mérite qu'on s'y attarde. En effet, cette adresse est probablement celle d'un serveur de stockage de fichiers javascript (Stockhan, 2023). Si on pousse plus loin l'étude de cette URL, on y voit le mot clé pixelList et les noms "facebook", "clicktripz", "criteo", "google", "bing", "iheartradio", "tiktok", "dv360". Ce sont tous les pixels cachés dans la page de Tripadvisor et qui demandent l'envoi d'informations sur l'utilisateur lors du chargement de la page volontairement visitée.

Cette valeur de cookie ne donne à nouveau pas d'entrées directement dans la Table "http_redirect". Toutefois, comme pour lematin.ch, des redirections ont bien lieu pour ce site. Elles sont au nombre de 87. Sans montrer la Table en entier, j'ai par exemple pu observer

des demandes entre des partenaires Google, comme DoubleClick et Adservice (montré dans le Tableau 13), DoubleClick et Casalemedia ou encore Cs.media et Amazon-adservices (pas montrés ici).

old request url	old request id	new request url
https://ad.doubleclick.net/ddm/activity/src=5153226;type=invmedia;cat=ta_us0;dc_lat=;dc_rdid=;tag_for_child_directed_treatment=;tfua=;npa=;gdpr=\$%7BGDPR%7D;gdpr_consent=\$%7BGDPR_CONSENT_755%7D;ord=4582880554700.993?	236	https://adservice.google.com/ddm/fls/z/src=5153226;type=invmedia;cat=ta_us0;dc_lat=;dc_rdid=;tag_for_child_directed_treatment=;tfua=;npa=;gdpr=\$%7BGDPR%7D;gdpr_consent=\$%7BGDPR_CONSENT_755%7D;ord=4582880554700.993

Tableau 13 : Extrait de la Table “http_redirect” de OpenWPM pour tripadvisor.ch

Ces deux url’s montrent des valeurs communes qui permettent à DoubleClick et Casalemedia d’identifier cet utilisateur et recouper des informations sur lui.

On peut finalement procéder à la même analyse pour le site doodle.com. Les lignes 1 à 3 du Tableau 14 reflètent une navigation ayant accepté les cookies et les lignes 4 et 5 les ayant refusés.

	Browser ID	Host	Cookie name	Cookie value
1.	1642482422 (doodle.com, Y)	.adnxs.com	uuid2	6666510192071048688
2.	1642482422 (doodle.com, Y)	.ads.yieldmo.com	ptran	6666510192071048688
3.	1642482422 (doodle.com, Y)	.doodle.com	_ga_VYXRTXP3Z1	GS1.1.1686227748.1.0.1686227748.0.0.0
4.	1642482422 (doodle.com, Y)	.doodle.com	eupubconsent-v2	CPtCerAPtCerAAcABBEN DHCsAH [...]
5.	1485801276 (doodle.com, N)	.doodle.com	eupubconsent-v2	CPsraHAPsraHAAcABBE NDGCgAAAAAH [...]

Tableau 14 : Extrait de la Table “javascript_cookie” de OpenWPM

Les lignes 1 et 2 partagent le même cookie bien que la requête ne vienne pas de la même entreprise. Ces deux hôtes sont des entreprises de publicité.

Lorsque que la valeur du cookie adnxs (6666510192071048688) est entrée dans la table “http_request”, il en ressort 10 entrées ayant comme URL’s adnxs, yieldmo mais aussi casalemedia et pubmatic. Il est à noter que la réponse à la requête de Pubmatic n’a pas

abouti pour des questions de confidentialité. Cela se voit au statut de la réponse et au texte associé (*status = 400, texte = Request failed due to privacy signal*).

Comme la plupart des sites, Doodle fait appel à Google Analytics pour obtenir des informations statistiques sur l'utilisation de son site. On le voit avec le nom du cookie de la ligne 3.

En utilisant la première partie de la valeur de ce cookie, on trouve 13 entrées dans la table "http_request". Parmi lesquelles cinq sont en lien avec google et les autres avec des fournisseurs de publicité (Outbrain et DoubleClick).

Les redirections depuis Doodle sont drastiquement différentes selon si l'on accepte ou non les cookies. Ainsi, lorsqu'ils sont déclinés, seules 3 redirections sont enregistrées contre 142 lorsque les cookies sont acceptés !

Les cookies des lignes 4 et 5 sont tronqués pour une meilleure visibilité. Le nom du cookie amène à penser qu'ils sont en lien avec un consentement pour de la publicité. Le cookie de la ligne 4 est retrouvé dans 173 requêtes, ayant des url pointant vers doodle.com mais également les sites publicitaires habituels. De manière rassurante, le cookie de la ligne 5 ne correspond à aucune entrée dans la table "http_request".

En partant de l'identité d'une requête qui a été redirigée, choisie au hasard, (*request_id = 366*), j'ai pu trouver une valeur de cookie (*COOKIE IDE=AHWqTUIJs6UW3yCFj64iY7GA*) utilisée dans 19 requêtes, toutes en lien avec DoubleClick. Les en-têtes montrent qu'une seule requête émane directement de Doodle. Toutes les autres viennent de DoubleClick, Yieldmo, Casalemedia, Pubmatic ou Openx. Ce sont majoritairement des chargements d'images, probablement de la publicité ciblée.

5.4. Comparaison avec Lightbeam

L'utilisation de OpenWPM permet une étude plus approfondie des liens entre les sites que Lightbeam. En effet, OpenWPM enregistre bien plus d'informations. Par exemple, tandis que Lightbeam se contente d'enregistrer les URL's, OpenWPM enregistre en plus les en-têtes des différents échanges entre sites et les décortique de telle manière à ce que nous puissions facilement repérer qui est l'hôte, le référent (l'URL d'où vient la demande), le type de ressource, le statut de la requête etc... . On peut dès lors facilement déterminer si la requête demande le chargement d'une image ou d'un script.

J'ai sur cette base réalisé une statistique des sites tiers apparaissant le plus souvent. Il faut se rendre compte que ces valeurs seront forcément différentes de celles établies avec Lightbeam puisque, d'une part le surf n'est pas le même et d'autre part, j'ai filtré l'ensemble de la Table "http_request" avec les mots clés des entreprises et pas uniquement l'URL. Or comme je viens de l'écrire, Lightbeam ne permet une recherche que sur l'URL. Ainsi, si le nom de l'entreprise apparaît dans l'en-tête et pas dans l'URL alors je ne le verrai pas dans Lightbeam tandis que je le verrai grâce à OpenWPM.

Les résultats sont présentés dans le Tableau 15.

Nom du site tiers	lematin.ch Y	lematin.ch N	tripadvisor.ch	doodle.com Y	doodle.com N	Total
Googlesyndication	57	17	62	86	0	222
Pubmatic	37	3	60	121	0	221
Casalemedia	21	18	38	58	0	135
DoubleClick	37	10	46	31	0	128
Outbrain	0	0	80	11	0	91
Criteo	1	0	48	7	0	56
Rubicon	0	0	31	11	0	42
Bing	1	0	34	0	0	35
Adnxs	11	9	4	8	0	32
Adservice	4	0	8	2	0	14

Tableau 15 : Nombre d'occurrence de sites tiers par site visité

Ce tableau permet les observations suivantes :

- Googlesyndication tient à nouveau le haut du panier mais plus du tout de manière aussi claire qu'avec Lightbeam puisqu'ici le nombre d'occurrence n'est pas significativement plus grand que celui de Pubmatic. Ce dernier n'apparaissait d'ailleurs même pas comme acteur principal tant il était masqué par les occurrences de Googlesyndication et Doubleclick. Je nuancerai tout de même cette découverte par le fait que Pubmatic apparaît essentiellement lors du surf sur doodle.com, site visité par un seul utilisateur du panel. De plus, pour cet utilisateur, Pubmatic apparaît bien en-deçà de Googlesyndication, Doubleclick ou Criteo. Malgré tout, OpenWPM permet de révéler Pubmatic comme acteur non négligeable et ce, grâce à la recherche par les en-têtes des requêtes http. C'est une SSP américaine (PubMatic, 2023).
- Tripadvisor et Doodle, lorsque les cookies sont acceptés, utilisent une plus large palette d'acteurs publicitaires que lematin.ch. On peut imaginer que cela est dû à l'internationalité de ces sites vis-à-vis du site lematin, qui n'est qu'un site national, de surcroît suisse, donc à petite audience.
- Il est intéressant de constater que Doodle respecte le choix de ses utilisateurs puisqu'aucun publicitaire n'apparaît lorsque les cookies sont refusés.

Pour en revenir à Facebook et au fait qu'il est apparu comme *First Party* lors des tests avec Lightbeam / Thunderbeam quand bien même les utilisateurs de mon panel n'ont pas visité le site volontairement, il faudrait, pour en comprendre le mécanisme le voir apparaître sous la colonne "top_level_url" dans la Table "http_request" et observer l'en-tête associé. Ce n'est malheureusement pas le cas. Je ne suis donc pas en mesure d'expliquer ce phénomène ni s'il s'agit d'un biais dans ma recherche ou d'un défaut des extensions utilisées.

Les réseaux sociaux n'apparaissent que pour les sites lematin.ch et tripadvisor.ch, ce qui présume que doodle.com n'a aucun lien avec les réseaux habituels, que les cookies soient acceptés ou non. Par ailleurs, certains réseaux sont totalement absents de la base de données OpenWPM utilisée ici. C'est le cas pour "instagram", "youtube", "snapchat" et "twitter". On trouve cependant les valeurs suivantes en filtrant l'URL de la Table "http_request" avec les mots-clés "facebook", "linkedin", "tiktok" et "pinterest".

Browser ID	Facebook	TikTok	Pinterest	Linkedin
1312132910 (lematin.ch, Y)	8	4	0	0
1468245175 (lematin.ch, N)	9	4	0	0
1642466324 (tripadvisor.com)	9	44	5	1

Tableau 16 : Nombre d'occurrence des réseaux sociaux

Ces valeurs restent relativement basses en comparaison des valeurs trouvées pour les sites publicitaires. Le contraste avec les résultats obtenus par mon panel d'utilisateurs est grand. Je n'explique cela que par la nature des sites visités lors des surfs avec OpenWPM.

Il est intéressant de constater que pour toutes les requêtes, le type de ressources associé à Facebook ou TikTok est une image ou du code javascript. Même si cela ne prouve pas que des pixels espion sont présents sur ces sites, l'hypothèse qu'il s'agisse de ce type de traçage est réaliste. Je n'ai en tous les cas pas observé d'icônes visibles de Facebook ou TikTok sur ces sites.

Finalement je conclurai en disant qu'OpenWPM est un outil puissant pour une analyse pointue des sites internet et de leur comportement en matière de confidentialité, mais que selon le degré de recherche que l'on veut atteindre, Lightbeam est suffisant.

6. Proposition de discussion avec élèves

Le présent chapitre propose un plan de cours de 2 périodes, sous forme de travaux pratiques avec des élèves de première ou deuxième année en voie maturité. Idéalement ce cours devrait avoir lieu en début d'année scolaire afin de pouvoir effectuer un suivi des données, sur une période plus longue.

Je ne remets pas ici le détail de certains passages qui ont déjà été discutés dans ce présent rapport. Une fois la séquence testée, elle sera en libre accès pour mes collègues.

Le plan de cours est le suivant :

- Introduction au sujet de la protection de la sphère privée sous forme de discussion / débat ;
- Introduction aux cookies ;
- Introduction à l'extension Lightbeam ;
 - Installation de l'extension
 - Démonstration de quelques liens entre les sites issus de cette recherche ;
 - Enregistrement des données privées à l'aide de Lightbeam ;
 - Export dans Excel ;
- Dépouillement des résultats individuels ;
 - Détection des sites essentiels ;
 - Détection des sites les plus traceurs ;
 - Détection des réseaux sociaux les plus présents ;
- Travail en groupe ;
 - Mise en commun de résultats obtenus ;
 - Présentation par un des élèves du groupe à l'ensemble de la demi-classe ;
- Proposition de protection.

6.1. La sphère privée c'est quoi ? (15 min)

Je commence par discuter de ce qu'est la sphère privée. Il existe un grand nombre de sources parlant de ce sujet, mais afin de ne pas alourdir le propos j'ai choisi une définition simple issue de Wiktionnaire : " Partie de la vie strictement réservée à une personne et dont elle décide des limites". (*Sphère Privée — Wiktionnaire*, 2023)

Je mets ensuite cette définition en lien avec les réseaux sociaux et les données laissées sur le net, de manière volontaire ou non, voire même de manière consciente ou non.

Selon le temps dont je dispose et de l'intérêt des élèves, je débattrai de l'article suivant qui, même s'il date un peu et n'est pas à jour du point de vue des réseaux sociaux récents, soulève des points intéressants de contradiction entre sphère privée et réseaux sociaux : [Qu'est-ce que la sphère privée ? » Les réseaux sociaux sur internet.](#)

Les questions à poser sont : Que signifie cette définition pour les élèves ? Que pensent-ils de la sphère privée ? Est-ce que ce concept leur parle ? Comment gèrent-ils leur sphère privée ? Quelles sont leurs habitudes face aux politiques de confidentialité ?

Si les élèves sont en OS philo/psycho, je propose de faire des liens avec la sociologie et de travailler de manière interdisciplinaire avec les collègues enseignant cette branche.

6.1.1. Pourquoi se protéger ?

Une fois le débat sur la sphère privée terminé, j'en commence un nouveau au sujet de l'importance de se protéger. En général, les élèves disent que leurs données personnelles ne sont pas importantes, que des tierces personnes ne peuvent pas être intéressées par ce qu'ils.elles pensent ou encore qu'ils.elles ne risquent pas de se les faire voler car ils.elles ne sont pas célèbres.

Je pense qu'il est important de casser ce mythe selon lequel seules les données de personnes célèbres sont intéressantes. Bien au contraire, les données de Monsieur et Madame tout le monde sont une mine d'or pour les publicitaires qui s'affairent autour du web. À ce moment du cours, j'introduis la notion d'apparente gratuité du web. Par le passé, lorsque nous désirions avoir une information, nous devions nous procurer un livre, un journal, avoir un abonnement dans une bibliothèque etc.... Bref, nous devions payer. Aujourd'hui, on nous fait croire que tout est gratuit. Or il n'en est rien. Les sites offrent une prestation pour laquelle ils font croire qu'ils ne sont pas payés. C'est un leurre. Quand il n'y a pas de produits, c'est nous le produit. Maintenir un site à jour a un coût. Dès lors, si le site n'est pas payant c'est l'utilisateur qui est le produit. Cela signifie que le site va monnayer d'une manière ou d'une autre les données divulguées par l'utilisateur.

Les données servent par exemple à proposer du contenu ciblé. Comme expliqué dans ce rapport, recevoir des propositions de manière récurrente va finir par influencer notre comportement et potentiellement induire une envie d'achat. Si ce n'est pas un achat qui est proposé, cela peut être des articles vantant les mérites de tels ou tels partis politiques ou sociétés. À nouveau, il est question d'influence et de voir son libre arbitre biaisé.

6.2. Les cookies (5 min)

On ne peut pas parler de traçage sans parler des cookies. Sans entrer dans le détail présenté dans ce rapport, je mentionne ce que sont les cookies, les différents types qui existent, leur utilité et le risque qu'ils représentent s'ils sont utilisés sans réglementations.

À ce moment, je demande aux élèves quelles sont leurs habitudes en matière d'acceptation ou de refus des cookies.

Si l'ensemble des élèves les accepte, je demande à quelques-un.e.s de les refuser afin d'être en mesure d'observer les effets éventuels.

6.3. Lightbeam (25 min)

Une fois cette introduction faite, je présente l'extension Lightbeam, ce qu'elle enregistre, et comment on l'installe, sans entrer dans le détail du fonctionnement.

La première chose que je leur montre est le graphe observé après quelques clics en choisissant délibérément certains sites pour des caractéristiques bien précises, par exemple~:

www.synonymo.fr ;

www.geeksforgeeks.org ;

www.watson.ch ;

www.deepl.com ;

www.google.ch.

Synonymo et geekforgeek.s parce qu'ils partagent des sites tiers. Watson car c'est un site que les jeunes utilisent comme source d'informations. À cette occasion, je ne manque pas de mettre en évidence les nombreux liens vers des publicitaires et des réseaux sociaux. Le site de traduction www.deepl.com car, non seulement, les élèves ont tendance à l'utiliser, mais en plus c'est un site isolé, qui fonctionne avec ses propres serveurs et statistiques (ce qui ne signifie pas qu'il ne garde pas nos données, mais il ne les partage pas). Finalement Google, pour sa popularité et son omniprésence lors de nos recherches sur internet.

Ensuite je leur accorde du temps, 10 minutes au maximum, pour surfer sur leurs sites préférés, accepter ou non les cookies et faire ce qu'ils font habituellement sur internet pour rendre leur expérience réelle et personnelle. Je ne leur laisse pas plus de temps car l'idée n'est pas qu'ils regardent des vidéos sur YouTube ou un autre réseau social mais bien qu'ils soient actifs et passent d'un site à un autre. À mon avis, ce moment est critique et je dois rester vigilante pour qu'ils fassent ce que je leur demande plutôt que de rester sur un site particulier. Selon leur enthousiasme, j'adapte les consignes et leur demande de faire une recherche particulière, par exemple trouver un lieu de vacances qui les attire, un hôtel et un restaurant en ne se basant pas uniquement sur Tripadvisor mais sur des recommandations venant de sources variées.

S'il reste du temps, les élèves choisissent un site sur lequel tous vont surfer. S'ils n'arrivent pas à se décider, je leur en impose un. Ceci permet d'établir le profil d'un site particulier en fonction du refus ou de l'acceptation des cookies.

Au bout de ces 10 minutes, je leur montre comment enregistrer les résultats et les exporter dans Excel. Selon l'année scolaire dans laquelle se trouve les élèves, je peux introduire quelques notions du langage de programmation json, et leur demander de créer eux-même un programme, dans python, permettant l'extraction de données d'un fichier.

6.4. Dépouillement des résultats (35 min)

Une fois le fichier Excel enregistré, ils doivent établir quelques statistiques, similaires à celles présentées dans ce rapport. Je commence par leur parler des sites essentiels au fonctionnement de la page et peut faire des liens avec les cookies présentés précédemment. Ensuite ils doivent :

- déterminer le nombre de sites tiers par site primaire visité ;
- corriger le nombre de sites tiers en fonction de sites essentiels ;
- déterminer le nombre de sites tiers non essentiels par site primaire visité ;
- établir un graphique du point précédent ;
- détecter le site qui fait appel au plus grand nombre de sites tiers ;
- détecter le réseau social qui apparaît le plus
- si cela s'applique :
 - déterminer le nombre corrigé de sites tiers pour le site primaire commun à tous ;
 - observer si des réseaux sociaux sont présents parmi ces sites ;
 - observer quels autres traceurs sont présents.

Une fois la partie individuelle terminée, les élèves se mettent par groupe de deux et :

- discutent entre élèves de l'impact d'accepter ou refuser les cookies ;
- réfléchissent à une manière de présenter les résultats ;
- présentent leurs résultats.

6.5. Comment se protéger ? (10 min)

Sur la base de ces résultats, je reprends la discussion au sujet de la sphère privée et leur demande ce qu'ils en pensent à présent et comment ils agiront à l'avenir.

L'utilisateur averti sait que ses données sont récoltées. Est-il conscient de ce que les plateformes en font ? Sait-il qu'il est considéré comme un produit marchand que les publicitaires cherchent à atteindre par tous les moyens ? Est-il conscient qu'en réalité internet n'est pas gratuit ? Que fait-il de son libre arbitre face à toutes les publicités qui lui sont soumises ? Autant de questions qui restent en suspens et que cette recherche n'adresse pas mais que j'aborderai dans le cadre du cours que j'ai élaboré grâce à cette étude, à l'intention des élèves du gymnase.

Je conclus la séquence en discutant des moyens existants pour se protéger.

7. Conclusion

Les objectifs de cette étude sont : obtenir des données collectées par de véritables utilisateurs, vulgariser les résultats pour en permettre l'accès à un large public et créer un cours pour des élèves du gymnase.

Naviguer sur internet est une activité quotidienne pour la plupart d'entre nous. Bien que cela puisse sembler banal et sans danger, chaque utilisateur laisse des traces de son passage, lesquelles servent de base aux acteurs publicitaires du web pour proposer du contenu ciblé. Seuls certains d'entre nous sont conscients de cette réalité. Certains laissent faire et acceptent les conséquences, d'autres refusent les cookies pensant que cela les prémunit de toute intrusion dans leur vie privée, d'autres encore utilisent des bloqueurs de publicité.

Un grand nombre d'études a été publié sur le sujet. La plupart sont d'un niveau technique élevé, ce qui empêche leur compréhension par une grande partie de la population. De plus, pour des questions de validation statistique, les données récoltées sont celles de robots imitant des humains surfant sur le net. Finalement, ces navigations s'effectuent sur des milliers de sites web durant des heures innombrables, sans interruption, ce qui n'est pas représentatif du comportement d'une personne réelle.

Dans mon étude j'ai analysé les données de 10 personnes ayant enregistré leurs navigations pendant une semaine grâce à une extension (Lightbeam ou Thunderbeam) installée sur leur navigateur. J'ai complété les résultats à l'aide d'OpenWPM, outil enregistrant l'intégralité des échanges qui se produisent lors du chargement d'une page web.

Bien que la taille de l'échantillon ne permette pas une validation statistique, les résultats issus de mon analyse sont le fruit de d'expériences web vécues par des humains.

La partie théorique de ce rapport, qui se veut accessible à tous, résume le fonctionnement des cookies et des protocoles de transfert régissant internet. Elle détaille également les techniques de traçage utilisées par les publicitaires, présents sur tous les sites ou presque, que nous visitons.

L'analyse des résultats a permis de constater que le nombre de sites tiers contactés à l'insu des utilisateurs ne dépend pas seulement de l'acceptation ou du refus des cookies mais également du type de sites visités et des actions y sont menées. En effet, les sites d'informations et de commerce en ligne tendent à abreuver les internautes de publicités même s'ils ont refusé les cookies. J'observe tout de même une diminution du nombre de sites contactés selon si la navigation se fait en acceptant ou en refusant les cookies. En moyenne, cette valeur passe de 10 à 13 sites contactés en arrière-plan à 3 à 6, lorsque les cookies sont refusés.

Cette étude a confirmé la présence de traçage sur les sites étudiés, révélant des échanges d'informations entre des entreprises publicitaires. La compagnie publicitaire qui ressort le plus souvent est Google et ses différents partenaires. Les réseaux sociaux, spécifiquement Facebook, Youtube et Tiktok, sont également présents comme sites tiers, recueillant des informations sur nos activités en ligne.

Grâce à cette étude, j'ai acquis de nouvelles compétences dans les différents sujets abordés, ce qui m'a permis de développer un cours à l'intention des élèves du gymnase. Ce cours aborde les thèmes des cookies, du traçage et du respect de la sphère privée et vise à accroître leur degré de connaissance sur ces sujets tout en éveillant leur curiosité. Les exemples concrets de cette étude constituent une base solide pour la création de matériel pédagogique dont je pourrai m'inspirer et dont je pourrai faire profiter mes collègues.

Remerciements

Je tiens à exprimer ma profonde gratitude envers mon directeur de projet individuel, Linus Gasser, pour sa précieuse supervision tout au long de cette étude. Ses conseils avisés, son temps généreusement consacré et ses nombreuses relectures ont grandement enrichi ce travail de recherche. Sans son soutien, ce projet n'aurait pas abouti.

Je tiens à remercier grandement les participants de cette étude, qui ont accepté de partager leurs actions sur le net. Leur contribution volontaire a été essentielle pour la collecte des données. Je suis particulièrement reconnaissante de leur participation et de leur confiance. Sans eux, ce projet n'aurait simplement pas pu voir le jour.

Je souhaite également adresser mes remerciements à mon experte, Sandra D. Siby, qui m'a aidée à installer et à comprendre le fonctionnement d'OpenWPM.

Mes remerciements vont également à Chiara Tanteri et Gaia Barazzetti qui m'ont guidées dans l'écriture et la soumission d'une demande de consentement au comité éthique de l'EPFL. Leur expertise et soutien ont facilité la mise en place de cette étude dans le respect des normes éthiques.

Enfin, je souhaite remercier Olivier Levêque pour sa lettre de soutien.

Je suis sincèrement reconnaissante envers toutes les personnes mentionnées ci-dessus, ainsi que toutes celles qui ont contribué de près ou de loin à la réussite de cette étude.

References

- Andreessen, M. (2023, March 30). *HTTP cookie*. Wikipedia. Retrieved April 15, 2023, from https://en.wikipedia.org/wiki/HTTP_cookie
- Bashir, M. A., Arshad, S., Robertson, W., & Wilson, C. (2016, August). Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. *Proceedings of Usenix Security*, 481-496. Retrieved November, 2022, from <https://personalization.ccs.neu.edu/Projects/Retargeting/>
- Baudry, B., & Laperdrix, P. (2015, September 01). *Le fingerprinting : une nouvelle technique de traçage*. Ed-Diamond.com. Retrieved April 12, 2023, from <https://connect.ed-diamond.com/MISC/misc-081/le-fingerprinting-une-nouvelle-technique-de-tracage>
- Bingler, S., West, M., & Wilander, J. (2023, April 5). *Cookies: HTTP State Management Mechanism*. IETF HTTP Working Group. Retrieved April 12, 2023, from <https://httpwg.org/http-extensions/draft-ietf-httpbis-rfc6265bis.html#name-table-of-contents>
- CookieDatabase. (2023, June 27). *Cookie: _fbp*. Cookiedatabase.org. Retrieved July 4, 2023, from https://cookiedatabase.org/cookie/facebook/_fbp/
- CryptPad*. (2023, May 03). CryptPad: Collaboration suite, encrypted and open-source. Retrieved May 11, 2023, from <https://cryptpad.fr/>
- Edwards, J. (2022, December 6). *What is a Tracking Pixel? How Web Beacons Work*. CHEQ. Retrieved April 11, 2023, from <https://cheq.ai/blog/what-are-tracking-pixels/>
- Englehardt, S., & Narayanan, A. (2016). Online tracking: A 1-million-site measurement and analysis. *Proceedings of ACM CCS 2016*. https://senglehardt.com/papers/ccs16_online_tracking.pdf
- Fouad, I., Bielova, N., Legout, A., & Sarafijanovic-Djukic, N. (2020). Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *Proceedings on*

- Privacy Enhancing Technologies*, 2020(2), 499-518.
<https://petsymposium.org/2020/files/papers/issue2/popets-2020-0038.pdf>
- Fouad, I., Santos, C., Legout, A., & Bielova, N. (2021, May). Did I delete my cookies? Cookies respawning with browser fingerprinting. *Computing Research repository*.
<https://doi.org/10.48550/arXiv.2105.04381>
- Gaudiaut, T. (2022, December 15). *Infographie: Les géants de la publicité en ligne*. Statista. Retrieved May 16, 2023, from
<https://fr.statista.com/infographie/28966/principales-societes-plateformes-selon-les-ventes-mondiales-de-publicite-en-ligne-en-2022/>
- Ghostery GmbH. (2023, April 03). WhoTracks.me - Bringing Transparency to Online Tracking. Retrieved April 17, 2023, from <https://whotracks.me/>
- Glossary of Marketing and Data Terms*. (2023, April 5). Treasure Data. Retrieved June 11, 2023, from <https://www.treasuredata.com/glossary/>
- GoldSparrow. (2021, June 23). *Googlesyndication*. Supprimer les logiciels espions & les logiciels malveillants avec SpyHunter. Retrieved April 18, 2023, from
<https://www.enigmasoftware.fr/googlesyndication-supprimer/>
- Happy. (2023, May 30). *Best SSPs for Publishers in 2023 – A Comprehensive List*. Headerbidding.co. Retrieved July 13, 2023, from
<https://headerbidding.co/best-ssp-for-publishers/>
- Hercher, J. (2021, December 21). *Xandr, Formerly AppNexus, Is Now Formerly AT&T, After Its Acquisition By Microsoft*. AdExchanger. Retrieved July 13, 2023, from
<https://www.adexchanger.com/online-advertising/xandr-formerly-appnexus-is-now-formerly-att-after-its-acquisition-by-microsoft/>
- INTEGRAL AD SCIENCE*. (2023, March 12). IAB France. Retrieved April 19, 2023, from
<https://www.iabfrance.com/membre/integral-ad-science>
- Jackson, T. (1996, February 12). This Bug in Your PC is a Smart Cookie. *Financial Times*.
- Jonoxia. (2012, 02 01). *Favicon not always found · Issue #39 · mozilla/lightbeam*. GitHub. Retrieved May 11, 2023, from <https://github.com/mozilla/lightbeam/issues/39>

Klassen, C. (2023, April 17). Start. Retrieved May 23, 2023, from <https://www.lightbeam.chikl.de/>

La Rédaction JDN. (2019, January 11). *Ce que veulent dire les chiffres dans les cookies _ga d'Universal Analytics*. JDN. Retrieved July 4, 2023, from <https://www.journaldunet.fr/web-tech/tutoriels-analytics/1203131-ce-que-veulent-dire-les-chiffres-dans-les-cookies-ga-d-universal-analytics/>

The Largest Independent Sell-Side Ad Platform. (2023, July 6). Magnite. Retrieved April 19, 2023, from <https://www.magnite.com/?lang=fr>

L'extension Lightbeam pour Firefox n'est plus prise en charge | Assistance de Firefox. (2019). Retrieved April 12, 2023, from <https://support.mozilla.org/fr/kb/extension-lightbeam-firefox-plus-prise-en-charge>

Microsoft. (2023). Bing. Retrieved April 18, 2023, from <https://www.bing.com/>

Musa, M. B., & Nithyanand, R. (2022). ATOM : Ad-network tomography. *Proceedings on Privacy Enhancing Technologies*, 2022(4), 295-313. <https://doi.org/10.56553/popets-2022-0110>

Papadopoulos, P., Kourtellis, N., Markatos, E. P., & Association for Computing Machinery. (2019, May). Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. *WWW '19: The World Wide Web Conference*, 1432–1442. <https://doi.org/10.1145/3308558.3313542>

Parlement Européen. (2016, May 24). *Le règlement général sur la protection des données - RGPD*. CNIL |. Retrieved June 22, 2023, from <https://www.cnil.fr/fr/reglement-europeen-protection-donnees>

Patil, H. (2018, January 21). *Ultimate Guide to HTTP Cookies. What every web developer needs to know...* | by Harshal Patil. webf. Retrieved April 13, 2023, from <https://blog.webf.zone/ultimate-guide-to-http-cookies-2aa3e083dbae>

promo10g20. (2011). *Qu'est-ce que la sphère privée ? » Les réseaux sociaux sur internet : une dissolution de la sphère privée ?* Description de controverses – Mines Paris. Retrieved June 7, 2023, from

https://controverses.minesparis.psl.eu/prive/promo10/promo10_G20/quest-ce-que-la-sphere-privee/index.html

PubMatic. (2023, July 9). Programmatic Digital Marketing | Advertising Technology Company. Retrieved July 10, 2023, from <https://pubmatic.com/>

Schmitt, S. (2023, February 05). *France Marketing Technology Landscape Supergraphic*. Retrieved May 16, 2023, from

<https://samuelschmitt.com/fr/france-marketing-technology-landscape-supergraphic/>

Schroeder, S., & Tyler, J. (2023, May 14). *The Top DSPs for 2023*. Playwire. Retrieved July 13, 2023, from <https://www.playwire.com/blog/top-dsps>

SimilarwebLTD. (2023, April 19). Website Traffic - Check and Analyze Any Website | Similarweb. Retrieved April 21, 2023, from <https://www.similarweb.com/>

sphère privée — *Wiktionnaire*. (2023, April 29). Wiktionnaire. Retrieved June 7, 2023, from https://fr.wiktionary.org/wiki/sph%C3%A8re_priv%C3%A9e

Stockhan, K. (2023, January 28). *Js Cache*. JavaScripting.com. Retrieved July 4, 2023, from <https://www.javascripting.com/view/jscache>

Thomas, N., & Martin, C. (2023, May 10). *Top Data Management Platforms in 2023*. Playwire. Retrieved July 13, 2023, from <https://www.playwire.com/blog/top-dmp-partners>

Thunderbeam-Lightbeam for Chrome. (2021, October 5). Thunderbeam-Lightbeam for Chrome. Retrieved May 11, 2023, from <https://chrome.google.com/webstore/detail/thunderbeam-lightbeam-for/hjkajeglckopdkbggdiajobpilgccgnj?hl=en-GB>

Trevisani, J., & Martin, C. (2022, December 3). *The Best Ad Exchanges for Publishers in 2023*. Playwire. Retrieved July 13, 2023, from <https://www.playwire.com/blog/best-ad-exchanges-for-publishers>

TutorialsPoint Contributor. (2023, March 12). *HTTP Tutorial*. Tutorialspoint. Retrieved May 4, 2023, from <https://www.tutorialspoint.com/http/index.htm>

Urban, T., Tatang, D., Degeling, M., Holz, T., & Pohlmann, N. (2020, October). The Unwanted Sharing Economy: An Analysis of Cookie Syncing and User Transparency under GDPR. *Proceedings of the 15th {ACM} Asia Conference on Computer and Communications Security*. <https://doi.org/10.1145%2F3320269.3372194>

URI - Glossaire MDN : définitions des termes du Web | MDN. (2022, September 21). MDN Web Docs. Retrieved May 7, 2023, from <https://developer.mozilla.org/fr/docs/Glossary/URI>

W3C contributeurs. (2010, January). Same Origin Policy - W3C - Web Security. Retrieved April 13, 2023, from https://www.w3.org/Security/wiki/Same_Origin_Policy

Wikipedia - Criteo. (2022, September 16). *Criteo*. Wikipédia. Retrieved April 18, 2023, from <https://fr.wikipedia.org/wiki/Criteo>

Wikipedia - DoubleClick. (2023, January 30). *DoubleClick*. Wikipédia. Retrieved April 18, 2023, from <https://fr.wikipedia.org/wiki/DoubleClick>

Wikipédia - HTTP. (2023, April 12). *Liste des codes HTTP — Wikipédia*. Wikipédia. Retrieved May 4, 2023, from https://fr.wikipedia.org/wiki/Liste_des_codes_HTTP

Wikipedia - Magnite. (2023, February 2). *Magnite Inc*. Wikipedia. Retrieved April 19, 2023, from https://en.wikipedia.org/wiki/Magnite_Inc

Wikipedia - MCI. (2023, March 25). *MCI Inc*. Wikipedia. Retrieved April 12, 2023, from https://en.wikipedia.org/wiki/MCI_Inc.

Wlosik, M., & Sweeney, M. (2021, August 12). *First-Party & Third-Party Cookies: What's the Difference?* Clearcode. Retrieved May 9, 2023, from <https://clearcode.cc/blog/difference-between-first-party-third-party-cookies/>

Zawadziński, M. (2018, April 5). *What is Cookie Syncing and How Does it Work?* Clearcode. Retrieved May 4, 2023, from <https://clearcode.cc/blog/cookie-syncing/>

Annexes

[HREC - Application](#)

[HREC - Formulaire de consentement](#)

[HREC - Formulaire d'information](#)