Marketing des traces : du *tracking*, des contre-mesures et de leur efficacité

Robert Viseur[1]

1. FWEG - Universite de Mons
Service de Technologies de l'Information et de la Communication
17, place Warocqué, B-7000 Mons
robert.viseur@umons.ac.be

RÉSUMÉ. D'un Web de documents à l'intérêt commercial incertain, porté par des pionniers croyant au partage des connaissances, le Web a par la suite évolué vers une forme collaborative et temps réel rentabilisée par la publicité. Cette dernière a évoluée vers la publicité ciblée incluant la publicité comportementale basée sur la collecte massive de traces d'usage. Ces traces proviennent de différents dispositifs de tracking incluant les adresses IP (IP tracking), les désormais connus cookies ou les empreintes (p. ex. browser fingerprinting et canvas fingerprinting). Si la collecte s'est au départ limitée au poste de travail (essentiellement au travers du navigateur), elle a pu par la suite s'étendre aux smartphones et objets connectés. En a découlé le marketing des traces et l'économie de l'attention auxquels les digital natives ont été précocement confrontés. Diverses contres-mesures ont été progressivement déployées par les utilisateurs (paramétrage, extensions, p. ex. bloqueurs de publicités), par des services d'anonymisation (p. ex. VPN et proxy), par les éditeurs eux-mêmes ou par le régulateur (p. ex. RGPD). Ce papier exploratoire propose, d'une part, une présentation de la structuration du secteur de la publicité en ligne suivie par un état de l'art sur les outils de tracking qui y sont déployés, d'autre part, un inventaire et une analyse des contre-mesures déployées ainsi que de leur efficacité. Nous montrons en particulier l'évolution rapide des techniques utilisées et l'hétérogénéité de la couverture offerte par des dispositifs protecteurs a priori équivalents.

Mots-clés : marketing des traces, économie de l'attention, adtech, publicité programmatique, publicité comportementale, privacy, tracking, big data.

1. Introduction

Mesguish et Thomas (2013) distinguent quatre âges du web. Le premier, s'étendant de 1994 à 1996, est baptisé « *Web des pionniers* ». Cette expression désigne le développement d'un Web encore réduit en taille alimenté par des pionniers technophiles. De 1996 à 2004, le « *Web des documents* » s'accompagne d'une explosion du nombre de sites permise par la facilité des nouveaux outils d'édition de contenu et alimentée par les débuts du commerce électronique. La recherche d'information passe par les annuaires ou par les moteurs de recherche. Cette période voit la naissance de l'entreprise Google. Le « *Web social* », parfois appelé Web 2.0, s'étend de 2004 à 2010. Il voit une implication plus importante des utilisateurs dans la création et l'enrichissement des contenus. Enfin, le « *Web temps réel* » se développe dès 2010 avec la part croissante des réseaux sociaux (audience) ainsi que le développement des *smartphones* et des tablettes. Les applications mobiles se développent au détriment du Web classique (documents, hyperliens, etc.). Cette évolution s'est accompagnée d'une mutation de la publicité en ligne sous des formes de plus en plus ciblées (Peyrat, 2009), jusqu'à la publicité comportementale cherchant à coller au plus près des centres d'intérêt immédiats des consommateurs tels que révélés par leur historique de navigation. Cette personnalisation avancée suppose un travail permanent de *tracking* (p. ex. *cookies*) et d'analyse de données (profilage) par les régies publicitaires (p. ex. Google et Facebook). Ce profilage des utilisateurs couplé à la connexion permanente (via le *smartphone*) conduisent à une nouvelle forme de capitalisme basé sur l'économie de l'attention. Le concept d'économie de l'attention a fait l'objet d'un effort de théorisation de la part d'Emmanuel Kessous (Kessous, 2011 ; Kessous, 2012). Ce dernier décrit la transition d'un marketing de segmentation vers un marketing des traces renforçant l'emprise des offreurs sur les consommateurs en l'absence d'un contrôle fort des données à caractère personnel[1] par les individus. Dans un monde où le coût de l'accès à l'information tend vers 0, l'objet rare n'est plus l'information mais bien l'attention. La généralisation des activités d'extraction de traces d'usage conduit à la mise en place d'un capitalisme de surveillance (Zuboff, 2019) couvrant à la fois les mondes virtuels (p. ex. moteurs de recherche) et réels (p. ex. objets connectés).

Le secteur de la publicité en ligne a donc sensiblement évolué depuis ses débuts seconde moitié des années quatre-vingt-dix. Il a en particulier bénéficié des principales tendances technologiques liées à la transformation numérique, *cloud computing*, *big data* et *machine learning* en tête. A titre d'exemple, l'entreprise française Criteo possédait en 2015 plus de 10.000 serveurs répartis dans 6 centres de données permettant de traiter jusqu'à 800.000 requêtes HTTP par seconde (Clapaud, 2015). Il en a résulté une réorganisation progressive du secteur faite de concentration (p. ex. Google) mais aussi de spécialisation de certains acteurs plus

[1] Ces contributions ont été écrites avant la mise en œuvre par l'Union européenne du Règlement Général de Protection des Données (RGPD).

petits. Sur le plan du *tracking*, de nouvelles techniques apparaissent (p. ex. *device fingerprinting*) tandis que d'autres deviennent obsolètes compte tenu de l'apparition de nouvelles techniques ou de la diffusion de contre-mesures efficaces (p. ex. blocage par défaut du *canvas fingerprinting*). Face à cette débauche de mécanismes de pistage numérique et à l'omniprésence de la publicité, le secteur a cependant dû faire face à des réactions issues des consommateurs (p. ex. bloqueurs de publicités), des associations militantes (cf. Framablog, 2017) ou du législateurs (p. ex. réglementation pour la protection des données personnelles). Il existe donc un besoin pour un état des pratiques qui soit à jour en matière de publicité en ligne et d'outils de *tracking* prenant en compte leur efficacité au regard de la diffusion de contre-mesures technologiques (p. ex. bloqueurs de publicités) ou légales (p. ex. RGPD).

Ce papier exploratoire est décomposé en quatre sections. Dans une première section, nous proposons de dresser un panorama des pratiques avancées de publicité en ligne (publicité contextuelle, publicité comportementale, *retargeting*, publicité programmatique…). Elle sera suivie d'une section dédiée aux techniques de *tracking* que ces pratiques nécessitent. Dans une troisième section, nous dressons un inventaire des contre-mesures disponibles. Dans une quatrième section, et avant de conclure par les limitations et les perspectives de cette recherche préliminaire, nous discuterons la diffusion de ces contre-mesures et de leur efficacité.

2. Essor de la publicité programmatique

Le marketing en ligne s'appuie sur diverses techniques maintenant éprouvées : courriels commerciaux, réseaux sociaux numériques, référencement de sites internet... Parmi celles-ci, la publicité en ligne recourt principalement à la diffusion de bannières (*display*), dont les formats sont standardisés, et de liens sponsorisés (*search*) au sein des moteurs de recherche (Allary et al., 2018). Les transactions relatives aux bannières se sont pendant plusieurs années réalisées de gré à gré, conduisant surtout à la valorisation des espaces publicitaires présents dans les pages principales des sites web, dès lors entraînant de nombreux invendus parmi les espaces présents sur les pages secondaires (longue traîne). La valorisation de cet inventaire s'est dès lors ouvert aux réseaux publicitaires (*ad networks*, *affiliate networks* ; p. ex. Tradedoubler) offrant une rémunération moindre mais permettant d'améliorer substantiellement le taux de remplissage des espaces.

La publicité en ligne s'est progressivement sophistiquée avec la publicité ciblée. Peyrat (2009) en distingue trois variantes. La publicité personnalisée dites classique est adaptée « *en fonction des caractéristiques connues de l'internaute* » telles que son âge, son sexe ou sa localisation. Ces données sont fournies volontairement par l'internaute, par exemple lors de l'inscription sur un service. La publicité contextuelle est déterminée « *en fonction du contenu immédiat fourni à l'internaute* ». L'annonce affichée est donc adaptée au contenu de la page web sur laquelle elle est affichée. Le ciblage peut éventuellement être affiné grâce à la géolocalisation de l'internaute ou par la recherche d'information (requête) qui a

conduit à la page par le bais d'un moteur de recherche. La publicité comportementale est choisie « *en observant le comportement de l'internaute à travers le temps* ». En pratique, un profil individuel va être dressé sur base des d'actions (historique de visites de sites web, des mots-clefs rentrés dans les moteurs recherches…), permettant une adaptation des publicités proposées. Parmi les techniques éprouvées et diffusées, citons en particulier le *retargeting* (Allary et al., 2018 ; Lambrecht et al., 2013). Ce dernier permet l'affichage, sur des sites externes, d'une publicité liée à un produit proposé sur le site de l'annonceur et pour lequel l'internaute a, lors d'une visite sur le site, marqué un intérêt (visualisation d'une page, recherche par mot-clef, inclusion dans une liste d'envies ou un panier d'achats…). L'objectif est dès lors de raccompagner le prospect dans l'entonnoir de conversion (*funnel*) jusqu'à la concrétisation d'une action (p. ex. prise de contact ou ventes).

Plusieurs régies se sont spécialisées sur ces différentes techniques plus avancées. D'une part, Google a investi dès 2000 dans son service de publicité Google Adwords (rebaptisé Google Ads en 2018) permettant un affichage de publicités textuelles (liens sponsorisés[2]) adaptées aux mots-clefs soumis au moteur de recherche ainsi que, au travers de la régie Google Adsense, un affichage de publicités textuelles adaptées en fonction du contenu de la page web contenant l'espace publicitaire, des mots-clefs associés à la publicité (achetés aux enchères et facturés au CPC[3]), de la géolocalisation de l'internaute, de sa langue et de la plage horaire (Allary et al., 2018). D'autre part, la société française Criteo s'est différenciée par son service de *retargeting*[4] permettant la personnalisation des annonces en fonction des pages consultées (*retargeting* statique) ou d'un profil individuel dressé sur base de données comportementales exploitées par des algorithmes de *machine learning* (*retargeting* dynamique). Google a par la suite ajouté un service équivalent de remarketing dynamique à sa régie Google Ads[5].

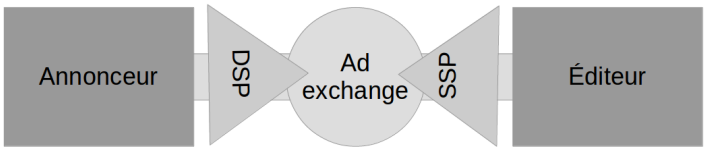

Figure 1. Écosystème de la publicité programmatique (Allary et al., 2013)

2 Ce type de produit publicitaire est classé dans le *Search Engine Advertising* (SEA), distinct du référencement naturel, soit le *Search Engine Optimization* (SEO), les deux étant regroupés dans le *Search Engine Marketing* (SEM).

3 CPC = *Cost per Clic* ou Coût par Clic.

4 Cf. https://www.criteo.com/fr/quest-ce-que-le-retargeting-votre-guide-complet/ pour plus de détails.

5 Cf. https://support.google.com/google-ads/answer/3124536 pour plus de détails.

Les pionniers comme Google ou Criteo ont ouvert la voie à une automatisation accrue de la publicité en ligne et ont conduit au développement de la publicité programmatique (Allary et al., 2018 ; cf. Figure 1). Cette dernière transforme la manière d'envisager une transaction commerciale entre un acheteur et un vendeur de publicité, au travers d'une place de marché (*ad exchange*) et grâce à la mise aux enchère en temps réel (RTB : *Real Time Bidding*) des espaces publicitaires disponibles (Renaud, 2017).

Tableau 1. Concentration et spécialisation des acteurs de la publicité programmatique (basé sur Allary et al., 2018 & Weide, 2018).

<table>
<tr><th>DSP</th><th>Ad exchange</th><th>SSP</th><th>Éditeur</th></tr>
<tr><td colspan="2">Google Adwords</td><td colspan="2">Google (search)</td></tr>
<tr><td colspan="2">Google Adwords</td><td>Google Adsense</td><td>Partenaires Adsense (display)</td></tr>
<tr><td>Search Ads 360 (ex-DoubleClick Search)</td><td>Google, Ads, Microsoft Advertising, Baidu...</td><td colspan="2">Bing, Baidu, Google, Yahoo (search)</td></tr>
<tr><td>Google Adwords, réseaux tiers</td><td>DoubleClick Ad Exchange (Google)</td><td colspan="2">Google Display Network, Adsense, Youtube, portails, sites d'actualités….</td></tr>
<tr><td colspan="3">Facebook Ads</td><td>Facebook</td></tr>
<tr><td>AdRoll, AppNexus, Criteo, MediaMath...</td><td>Criteo, DoubleClick (Google), Rubicon Project...</td><td>AppNexus, OpenX, PubMatic, Rubicon Project...</td><td>Éditeurs (portails, sites d'actualités…)</td></tr>
</table>

La publicité programmatique modifie profondément la chaîne de valeur de la publicité en ligne et voit l'émergence de fonctions spécialisées (Allary et al., 2018). Premièrement, l'annonceur consolide ses données clients au sein d'un DMP (*Data Management Platform*), notamment alimenté par ses outils CRM (*Customer Relationship Management*) et ses outils *analytics* permettant le suivi de l'activité sur les sites web de l'entreprise, éventuellement complété par les données fournies par des partenaires ou des courtiers en données (cf. Allary et al., 2018, et Framablog, 2017, pour plus de détails). Il peut ensuite émettre des ordres d'achat d'espace publicitaire sur un DSP (*Demand Side Platform*). A l'extrémité de la chaîne, les éditeurs de sites web gèrent un inventaire d'espaces publicitaires disponibles et diffusent des demandes d'offres (*bid requests*) sur un SSP (*Supplier-Side Platform*). Au centre, une plate-forme d'échange publicitaire (*ad exchange*) organise la rencontre entre les ordres d'achat (offre) et les demandes d'offres (demande) au travers d'un mécanisme d'enchères en temps réel (RTB). L'annonceur le plus généreux remporte l'enchère et son annonce peut dès lors être affichée sur le site de l'éditeur. Cette opération, dont la durée totale est inférieure à la seconde, suppose l'évaluation de la valeur commerciale de l'internaute face à l'espace publicitaire mis aux enchères (Allary et al., 2018 ; Framablog, 2017). Dans le cas idéal, les acteurs

spécialisés au sein de cette chaîne sont capables d'échanger des données (interopérabilité) et de mettre en commun des données relatives aux profils individuels, ce qui suppose un fastidieux travail de réconciliation de *cookies* (*cookie syncing*) et de création d'identifiants uniques (UUID) au sein notamment des DMP. Face à ces écosystèmes ouverts se positionnent les écosystèmes (partiellement) fermés de Google et Facebook (cf. Tableau 1).

3. Inventaire des techniques de *tracking*

La publicité en ligne s'appuie sur divers mécanismes de collecte de données personnelles et de suivi de l'activité des internautes au fil de leur navigation. Ce suivi passe par l'utilisation d'outils de *tracking*, une pratique qualifiée par ses détracteurs de « *pistage numérique* » (p. ex. Framablog, 2017). Un premier inventaire des mécanismes de *tracking* a été réalisé récemment par Ishtiaq et al. (2017).

Le *tracking* des adresses IP, aussi appelée IP *tracking*, permet le suivi de la navigation d'un internaute sur base de l'adresse IP reçue par chaque terminal de consultation connecté par Internet (Debize et al., 2016). L'adresse IP peut être fixe mais est plus généralement dynamique (p. ex. changement d'adresse lors du redémarrage d'une box internet domestique). L'IP *tracking* permet donc le suivi de la navigation sur une période de temps limitée. Cette méthode de *tracking* fonctionne par contre quelque soit le terminal (ordinateur personnel, téléphone…) et le logiciel d'accès à Internet (navigateur, application mobile...). L'adresse IP permet aussi la géolocalisation du terminal à l'échelle du pays (avec une fiabilité proche de 100%) ou de la ville (avec une fiabilité ne dépassant pas 90%) (Koch et al., 2013). En particulier, les adresses IP sont distribuées par l'ICANN par lots, l'appartenance à un lot permet de connaître l'organisation ou le pays correspondant à l'adresse.

La technologie centrale du *tracking* sur le Web est le *cookie* HTTP. Le *cookie* est un ensemble de données renvoyé par un serveur web au navigateur web et que ce dernier stocke ensuite localement[6]. Seul le serveur ayant créé le *cookie* HTTP peut ensuite en relire le contenu. De plus, les *cookies* ont une durée de vie limitée. Par ailleurs, ils peuvent être refusés (au cas par cas ou de manière systématique) ou supprimés à l'initiative de l'utilisateur (via les paramètres de sécurité du navigateur). S'il peut être utilisé pour gérer la connexion ou la personnalisation sur un site web, le *cookie* permet aussi le *tracking* à des fins publicitaires, soit qu'il contienne un identifiant permettant l'identification de l'utilisateur (et donc le rapprochement avec des données personnelles conservées en base de données par la régie) soit qu'il contienne des données relatives à son historique de navigation.

Le caractère potentiellement éphémère des *cookies* a conduit au développement de techniques pour en assurer la persistance. On parle alors de *cookie respawning*, d'*evercookie* voire de *cookie* zombie. Le principe consiste à récréer un *cookie* HTTP

[6] Cf. https://developer.mozilla.org/fr/docs/Web/HTTP/Cookies pour plus d'informations sur le fonctionnement technique des *cookies* HTTP.

après sa suppression en s'appuyant sur un autre dispositif de stockage, soit un *cookie* Flash (en réalité, un objet local partagé ou LSO), soit un mécanisme de stockage persistant dans le navigateur tel qu'IndexedDB (Acar et al., 2014). Le Flash n'étant plus utilisé que par moins de 3 % des sites web[7], le premier dispositif peut être considéré comme caduc.

Le *tracking* par *cookies* a été complété par diverses méthodes de calcul d'empreintes (*fingerprinting*), utilisables avec les navigateurs web (Acar et al., 2014), mais aussi avec les *smartphones*. S'agissant des navigateurs web, la technique consiste à exploiter l'extrême variété des configurations des navigateurs (*user agent* mais aussi liste des polices ou des extensions installées) et, plus largement, des postes de travail (système d'exploitation, modèle de carte graphique, version de pilote de carte graphique...). Le *browser fingerprinting* permet ainsi de calculer l'empreinte d'un navigateur sur base des spécificités précitées[8] tandis que le *canvas fingerprinting* exploite les différences (minimes) de rendu graphique. Plus précisément, le *canvas fingerprinting* consiste à transformer en image *lossless*, avec l'API canvas du navigateur web, une chaîne de caractères constituant un pangramme parfait (de manière à maximiser la diversité de rendu), puis à récupérer cette image avec la méthode Javascript `ToDataURL`, et enfin à transformer l'image en chaîne de caractères en utilisant le codage *base64*. Selon Acar et al. (2014), environ 5 % des sites classés dans le Top 100000 Alexa utilisaient le *canvas fingerprinting*, contre 2 % environ pour les sites issus du Top 1000 Alexa. Parmi les utilisateurs connus citons la société AddThis, dont les *widgets* sont largement diffusés et permettent une excellente couverture, soit 97,2 % selon Acar et al. (2014), de la population étasunienne. Firefox met en œuvre un blocage par défaut du *canvas fingerprinting* depuis Firefox 58 (publié le 23 janvier 2018).

En pratique, les méthodes de *fingerprinting* ont également été utilisées avec les téléphones (*device fingerprinting*). Elles s'appuient par exemple sur l'exploitation des données issues du suivi du rythme de décharge de la batterie (*battery fingerprinting*) ou des capteurs de mouvements (Chen et al., 2017 ; Das et al., 2018). Par ailleurs, les terminaux mobiles ont fait l'objet d'une nouvelle méthode de suivi : le *tracking* par ID (Allary et al., 2018 ; Reichgut, 2016). Ainsi, chaque terminal iOS (IDFA : *Identifier For Advertising*), Android (GAID : *Google Adversting ID*) ou Windows (WAID : *Windows Advertising ID*) possède un identifiant unique et non permanent, donc différent d'un numéro de téléphone ou d'un numéro de série, permettant le suivi du terminal (Al-Kabra et al., 2019).

La collecte de données à caractère personnel peut aboutir à l'identification d'un individu au cours de sa navigation, soit de manière directe (authentification, ID…), soit de manière indirecte par croisement d'informations. Par exemple, à l'extrême, la géolocalisation d'une adresse couplée à une empreinte de navigateur peut conduire à l'identification d'un individu particulier si sa demeure est localisée dans une zone

[7] Cf. https://w3techs.com/technologies/details/cp-flash pour un suivi des statistiques d'utilisation.

[8] Cette technique peut notamment être testée avec le site Panopticlick développé par l'*Electronic Frontier Foundation* (EFF).

faiblement peuplée et qu'il utilise une configuration atypique sur son terminal de connexion. Les acteurs actifs dans la publicité en ligne, et en premier lieu les régies publicitaires, recourent par ailleurs à la synchronisation de *cookies* (*cookie syncing*) de manière à regrouper les informations collectées par différents serveurs (Acar et al., 2014 ; Papadopoulos et al., 2019). Cette activité est en particulier essentielle dans le contexte de la publicité programmatique (Allary et al., 2018).

4. Inventaire des contre-mesures

4.1. Configuration du navigateur

Par souci de simplification, cette section portera sur le navigateur Firefox. Ce dernier en est actuellement à la version 73.0 (publiée le 11 février 2020). Premièrement, la configuration du navigateur permet de limiter l'utilisation des *cookies* par les sites consultés, soit que l'utilisateur les refuse au fur et à mesure, soit que l'utilisateur en bloque certains de manière systématique, soit qu'il les supprime périodiquement. Au sein de Firefox, ces opérations peuvent être configurées dans l'onglet « Vie privée et sécurité ». Deuxièmement, la navigateurs offrent généralement une fonctionnalité de navigation privée. Cette dernière permet une navigation sans enregistrement des cookies et de l'historique de navigation au-delà de la session courante[9]. Troisièmement, certains navigateurs offrent des fonctionnalités avancées de blocage de *trackers*. C'est notamment le cas de Firefox avec la via la fonctionnalité *Enhanced Tracking Protection* (ETP) accessible depuis la barre d'adresse[10].

4.2. Installation d'extensions

Les navigateurs modernes permettent généralement l'installation d'extensions (*plugins*). Parmi les extensions populaires, citons les bloqueurs de publicités (p. ex. AdBlock Plus ou uBlock Origin). Les filtres mis en œuvre peuvent cependant dépendre du bloqueur utilisé. Édité par la société eyeo GmbH, AdBlock Plus filtre ainsi par défaut les serveurs publicitaires mis sur liste noire par la communauté EasyList mais laisse par contre passer des « *publicités acceptables* » c'est-à-dire conformes aux critères du Comité Publicité Acceptable d'où le service AdBlock Plus tire ses revenus… Parmi les « clients » de ce comité citons la société Criteo. Cette dernière ne manque d'ailleurs pas de mentionner (discrètement) sa porosité aux bloqueurs sur son site commercial (« *recover ad-blocked impressions with our ability to serve Acceptable Ads* »). La protection offerte par les bloqueurs de publicités varie donc d'une solution à l'autre.

Au côté des bloqueurs de publicités, d'autres extensions spécialisées sont proposées. Citons en particulier Ghostery. Ghostery s'appuie sur une base de

[9] Cf. https://support.mozilla.org/fr/kb/navigation-privee-naviguer-avec-firefox-sans-enregistrer-historique pour plus de détails.

[10] Cf. https://blog.mozilla.org/blog/2019/06/04/firefox-now-available-with-enhanced-tracking-protection-by-default/ pour plus de détails.

données de *trackers* (plus de 4500) classés par catégories (publicité, *analytics*, réseaux sociaux...) pour permettre, sur la plupart des navigateurs web du marché, le filtrage des *trackers* (ou de catégories de *trackers*) sélectionnés dans les paramètres de configuration de l'outil (par exemple, les boutons sociaux ne sont pas supprimés par défaut).

4.3. Protection par la législation

La protection des données à caractère personnel présente des approches distinctes en fonction des pays et des cultures. Trois pôles majeurs tendent ainsi à se dégager : les États-Unis, la Chine et l'Union européenne (Demiaux, 2018). Le modèle étasunien de régulation des données personnelles est davantage centré sur la primauté de la liberté individuelle, voire associe la *privacy* à un comportement de dissimulation et à une source d'inefficience (Rochelandet, 2010). La Chine permet pour sa part la collecte massive au profit tant de l'état[11] que des entreprises. Elle met d'ailleurs progressivement en place un régime de sanctions aux « mauvais » citoyens sur base d'un système de crédit social au sein duquel chaque citoyen chinois est associé à un score de réputation (Raphaël et al., 2019). Le modèle européen a divergé du modèle étasunien à partir des années soixante-dix en érigeant la protection des données à caractère personnel au rang de liberté fondamentale. Cette conception a conduit à la mise en application à partir du 25 mai 2018 du Règlement sur la Protection des Données Personnelles (RGPD). Le modèle européen prend en compte l'asymétrie de pouvoir entre les grands organismes et les citoyens, et veille au consentement éclairé des citoyens confrontés à la collecte de données personnelles.

Le RGPD[12] repose notamment sur des principes de consentement éclairé et de proportionnalité des données collectées au regard des finalités du traitement tels que communiquées à l'utilisateur. La notion de données à caractère personnel est large puisqu'elle inclut des données directement nominatives (telles que le nom et le prénom) et des données indirectement nominatives (Banck, 2018). Sont donc notamment couverts par le règlement les identifiants, les données de localisation, les adresses IP ou les *cookies* relatifs à une personne physique identifiable directement ou indirectement. Cette définition volontairement très large réduit sensiblement la marge de manœuvre des entreprises, obligées de demander à l'utilisateur une autorisation explicite et préalable à toute collecte de données à caractère personnel, désormais dans l'incapacité d'agir dans l'ombre sans risquer un constat de violation du règlement suivi d'une amende pouvant aller jusqu'à 10 millions d'euros ou 2 % du chiffre d'affaires annuel mondial de l'exercice précédent.

[11] Nous ne développerons pas dans cet article la question de la collecte de données par les états et, en particulier, par les États-Unis. Nous renvoyons donc au chapitre 17 « *Cybersécurité : dimension géostratégique et politique* » de Debize et al. (2016) qui y consacrent un important développement.

[12] Nous renvoyons à Banck (2018) pour une présentation complète mais synthétique du RGPD.

4.4. Anonymisation de la connexion

L'anonymisation de la connexion peut être mise en œuvre avec un niveau croissant d'efficacité par l'utilisation d'un *proxy*, d'un VPN ou d'un client Tor. Le *proxy* permet de masquer l'adresse IP du client car il expose sa propre adresse IP (Savchenko et al., 2015). Les VPN apportent en plus un chiffrement de la communication. Leur utilisation suppose de s'inscrire sur un serveur VPN à la fiabilité avérée, ce qui implique généralement le paiement d'un abonnement mensuel (p. ex. Ghostery Midnight). Quant à Tor, il repose sur une solution décentralisée et chiffrée s'appuyant sur un réseau de nœuds *proxy* (Savchenko et al., 2015). En outre, le navigateur Tor inclut différents mécanismes de lutte contre les *evercookies*, le *canvas fingerprinting* et le *cookie syncing* (Acar et al., 2014).

5. Discussion

L'utilisation de contre-mesures efficaces par les internautes suppose une conscience minimale des mécanismes de *tracking*. Ils se révèlent malheureusement sous informés. Si les internautes ont connaissance de l'existence de la collecte de données, la nature de cette dernière leur est souvent inconnue (Morey et al., 2018 ; cf. Tableau 2). Ainsi, les trois quarts des internautes ignorent la collecte de leur localisation alors que cette dernière peut être obtenue au travers des informations GPS (*smartphone*) ou de l'adresse IP du terminal. Dans le même ordre d'idée, selon une étude Connected Life 2017, « seuls » 29 % des Belges, 34 % des Français et 30 % des Européens utiliseraient un *adblocker*, contre 18 % des internautes dans le monde. Suire (2016) laisse par ailleurs entendre que l'installation d'un bloqueur chez les étudiants découle davantage d'un sentiment d'agacement face aux intrusions publicitaires que d'un rejet des pratiques de collecte massive de données à caractère personnel.

Tableau 2. Prise de conscience de la collecte de données (Morey et al., 2018).

Types de données	**Pourcentage d'individus conscients de partager ce type de données**
Liste d'amis sur les réseaux sociaux	27 %
Localisation	25 %
Recherche sur le web	23 %
Historique de communication (p. ex. archive de *chat*)	18 %
Adresses IP	17 %
Historique de navigation	14 %

Tableau 3. Évaluation de l'efficacité des contre-mesures.

	Firefox (configuration)	**Firefox (nav. priv.)**	**Bloqueur de publicité**	**Tor (client)**	**RGPD**
Portée	Générale	Générale	Publicité②	Générale	Juridique
*Cookie*①	✔	✔	✔	✔	✔
Evercookie	✘	?	✔	✔	✔
Browser fingerprinting	±③	±③	✔	✔	✔
Canvas fingerprinting	✔③	✔③	✔	✔④	✔
Adresse IP	✘	✘	✔	✔	✔
Historique de recherche	±	±	na	✔	✔
Réaction de l'éditeur	Aucune mais inconfort….	Détection et blocage⑤	Détection et blocage⑤	Détection et blocage⑤	Application partielle⑥

① Les *cookies* peuvent être facilement refusés et effacés, de manière manuelle ou automatique, à l'aide d'un navigateur web. La navigation privée permet de systématiser l'effacement des *cookies* à la fermeture de l'onglet de navigation.

② Les bloqueurs de publicité permet de bloquer l'affichage de la publicité mais aussi la collecte de données par le *tracker* (*tag*) en interdisant l'exécution du script Javascript correspondant. Par contre, il ne bloque pas d'autres types de *trackers* (p. ex. Google Analytics). Pour ces derniers, des extensions spécialisées doivent être installées au cas par cas ; Firefox, depuis la version 67.0.1, permet par ailleurs de configurer le blocage de *trackers* via la fonctionnalité *Enhanced Tracking Protection* (ETP)[13].

③ Le *canvas fingerprinting* est bloqué par Firefox depuis la version 58. L'énumération d'extensions (*plugins*) y est par ailleurs limitée[14].

④ Le client Tor inclut des contre-mesures permettant de lutter efficacement contre le *canvas fingerprinting* et le *cookie syncing* (Acar et al., 2014). De plus, Tor permet l'anonymisation de la connexion.

⑤ La détection d'une contre-mesure permet à l'éditeur de site web d'éventuellement bloquer l'affichage du contenu. La détection est notamment possible pour les bloqueurs de publicités, la navigation privée ou l'utilisation de Tor.

⑥ L'application du RGPD incombe uniquement aux organismes établis en Europe ainsi qu'aux organismes établis hors Union européenne traitant les données de citoyens européens. De plus, l'efficacité réelle dépend du caractère réellement éclairé du consentement de l'utilisateur, de l'activité de détection des infractions, des plaintes déposées et de la capacité (réduite) des autorités de contrôle nationales (p. ex. CNIL en France).

13 Cf. https://blog.mozilla.org/blog/2019/06/04/firefox-now-available-with-enhanced-tracking-protection-by-default/ pour plus de détails.

14 Cf. https://bugzilla.mozilla.org/show_bug.cgi?id=757726 pour plus de détails.

Le Tableau 3 propose une synthèse de contre-mesures courantes et analyse leur efficacité au regard des techniques de *tracking* et des contre-mesures potentiellement mises en place par les éditeurs de sites web. En pratique, le navigateur Firefox permet la mise en place, à la configuration, d'un large éventail de dispositifs pour limiter le pistage (effacement des *cookies*, détection des calculs d'empreintes, blocage de *trackers*, envoi d'entêtes HTTP « *Do Not Track* »…). Les mêmes fonctionnalités tendent à se retrouver sous Chrome (cf. Tableau 4). Cependant, intégré dans un écosystème plus large permettant à Google de collecter des données et de déployer ses services de publicités ciblées, Google Chrome organise une certaine perméabilité aidant l'entreprise à ainsi préserver son modèle d'affaires (p. ex. filtrage des « *publicités intrusives ou trompeuses* »).

Tableau 4. Comparaison de Chrome et Firefox (lutte contre le pistage).

Fonctionnalités de blocage		**Chrome**	**Firefox**
Popups		✓ (activé)	✓ (activé)
Cookies		✓ (configurable)	✓ (configurable)
Publicité	Natif	✓ (désactivé)	✗
	Extensions	✓ (Adblock Plus, Ghostery...)	✓ (Adblock Plus, Ghostery…)
Trackers	Natif	✗	✓ (désactivé)
	Extensions	✓ (« Désactivation de Google Analytics » + Ghostery, Privacy Badger...)	✓ (Ghostery, Privacy Badger...)
Do Not Track		✓ (désactivé)	✓ (désactivé)

La collecte non désirée de données peut s'apparenter à un problème de sécurité car elle viole la confidentialité des données à caractère personnel. Bien utilisée, les contre-mesures disponibles disposent d'une portée et d'une réelle efficacité, même si elles peuvent elles-mêmes s'exposer à des contre-mesures de la part notamment des éditeurs de sites web (p. ex. entrave à l'affichage d'une page en cas d'utilisation d'un bloqueur de publicités). Compte tenu de la puissance commerciale de Google, le choix des internautes en matière de navigateur web ne reflète cependant pas l'investissement des éditeurs en matière de *privacy* (cf. Tableau 5 ; statistiques : Statcounter) alors même que la dépendance de Google au modèle d'affaires publicitaire est régulièrement rappelée (cf. par exemple Nitot, 2016).

Tableau 5. Comparaison des navigateurs.

	Open source	**Diffusion**	**Mise à jour**	**Innovation**	**Publicité ciblée**	***Privacy***
Firefox	✓	5 %	⋆⋆⋆	⋆⋆⋆		⋆⋆⋆
Chromium	✓	< 0,1 %	⋆⋆	⋆⋆		⋆⋆
Chrome		60 %	⋆⋆⋆	⋆⋆⋆	✓	⋆
Safari		15 %	⋆⋆⋆	⋆⋆		⋆⋆
Internet Explorer		< 2 %	⋆	⋆		⋆⋆
Edge		< 5 %	⋆⋆⋆	⋆⋆		⋆(⋆)
Opera		2,5 %	⋆⋆⋆	⋆⋆⋆	✓	⋆

Le extensions pour les navigateurs sont diversifiées et diffèrent suivant différents critères : couverture, compatibilité et confort d'utilisation (cf. Tableau 6 pour quelques exemples). Deux points ressortent. D'une part, à côté de solutions partielles existent des solutions globales permettant de configurer le filtrage de différents types de *trackers* (p. ex. Ghostery). D'autre part, à l'instar des navigateurs web, l'effectivité du filtrage est dépendante du modèle d'affaires de l'éditeur et de ses liens avec le marché de la publicité ciblée (p. ex. Adblock Plus) !

Tableau 6. Efficacité des extensions.

Extension	**Automatique**	**Couverture**	**Configuration**	**Compatibilité**	**Désagrément(s) connu(s)**	**Intérêt**
Adblock Plus	✓	*Trackers* publicitaires	✓	Chrome, Firefox, Internet Explorer, Safari, Edge, Opera…	Acceptation de la « *publicité acceptable* », détection par les sites	⋆
Ublock Origin	✓	*Trackers* publicitaires	✓	Chrome, Safari, Firefox, Chromium	Refus d'inclusion dans Chrome Web Store	⋆⋆
Ghostery	✓	*Trackers* publicitaires, *trackers analytics*...	✓	Chrome, Firefox, Safari, Edge, Opera, Cliqz (Firefox)	Détection (épisodique) par les sites	⋆⋆⋆
Privacy Badger	✓	*Trackers* (dont publicitaires)	✓	Chrome, Firefox, Opera	Ralentissement de la navigation (?)	⋆⋆

6. Conclusion

Notre recherche exploratoire nous a permis de décrire les évolutions du secteur de la publicité et d'expliquer le besoin en techniques avancées de *tracking* permettant le suivi individualisé de la navigation mais aussi la création de profils et la mise en commun de données via la synchronisation de *cookies*. Sur base d'un inventaire complet des techniques de *tracking*, nous avons pu proposer une analyse de l'efficacité de ces mécanismes de *tracking* ainsi que des contre-mesures déployées pour limiter ou bloquer la collecte de données à caractère personnel. Nous avons notamment montré l'hétérogénéité de la protection offerte par des contre-mesures à première vue équivalentes (p. ex. bloqueurs de publicités).

Notre recherche souffre d'au moins quatre limitations. Premièrement, l'analyse des possibilités de configuration ou d'extension des navigateurs s'est focalisée sur le navigateur Firefox. Or, la Mozilla Foundation, qui le produit, se distingue par ses engagements en faveur de la protection de la vie privée[15]. Une analyse similaire devrait donc être réalisée pour les éditeurs d'autres navigateurs web (p. ex. Apple Safari et Microsoft Edge), en particulier ceux édités par des entreprises dont le modèle d'affaires dépend de la publicité ciblée (p. ex. Google Chrome et Opera). Deuxièmement, le poste de travail fait l'objet d'une collecte de données au départ du système d'exploitation voire aussi des applications installées. Les activités de télémétrie liées à l'amélioration de la qualité sont ainsi critiquées (p. ex. télémétrie sous Windows 10) tandis que certaines applications sont épinglées pour la revente de données à caractère personnel éventuellement anonymisées (cf. antivirus Avast[16]). Troisièmement, notre recherche est centrée sur le poste de travail et les navigateurs web qui y sont utilisés. Or, l'utilisation d'Internet s'est substantiellement déplacée vers les terminaux mobiles (*smartphones*, tablettes…). Ces derniers présentent des particularités techniques et sont soumis à une activité très importante de collecte de données via des *trackers* publicitaires intégrés aux *apps* (Binns et al., 2018) ainsi que, *last but not least*, *via* des logiciels pré-installés tels qu'Android ou Google Maps (Nitot, 2016). Quatrièmement, l'analyse des outils de *tracking* investigue peu les techniques récentes dédiées aux terminaux mobiles et aux objets connectés de type identifiant unique non permanent. Cette pratique, largement diffusée pour les terminaux mobiles, intéresse également dans le cadre du développement de la publicité programmatique sur les télévisions connectées et s'inscrit dans une réflexion plus large sur le remplacement des *cookies* tiers dont le filtrage par les navigateurs (p. ex. Firefox et Safari) est de plus en plus fréquent (Sluis, 2020).

Parmi les perspectives, outre le travail sur les limitations précitées, citons l'élaboration d'une méthodologie outillée permettant de mesurer l'exposition d'un individu à la collecte de données à caractère personnel compte tenu de ses usages de dispositifs connectés et des éventuelles contre-mesures mises en œuvre.

---

15 Cf. https://www.mozilla.org/fr/firefox/privacy/ pour plus de détails.

16 Cf. https://www.zdnet.fr/actualites/oui-avast-vend-des-donnees-personnelles-et-cela-se-savait-39898209.htm pour plus de détails.

Bibliographie

Acar G., Eubank C., Englehardt S., Juarez M., Narayanan A. & Diaz C. (2014), The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, p. 674-689.

Al-Kabra R., Bodiga P. K., Dahlstrom N., Sinha R., Morrow J., Drake A., & Phan C. (2019). *Ascertaining network devices used with anonymous identifiers*. U.S. Patent Application No. 15/801,971.

Allary J. & Balusseau V. (2018). *La publicité à l'heure de la data. Adtech et programmatique expliqués par des experts*, Dunod.

Baudry B. & Laperdrix P. (2015). *Le fingerprinting : une nouvelle technique de traçage*, MISC, n°081, septembre 2015. En ligne : https://connect.ed-diamond.com/MISC/MISC-081/Le-fingerprinting-une-nouvelle-technique-de-tracage.

Banck A. (2018). *RGPD : la protection des données à caractère personnel*, Gualino.

Binns R., Lyngs U., Van Kleek M., Zhao J., Libert T. & Shadbolt N. (2018). Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, p. 23-31.

Broussard G. (2019). *Internalisation programmatique en France : taux d'adoption, avantages, degrés et types de fonction d'achat intégré par rapport à l'Europe*, Interactive Advertising Bureau (IAB), avril 2019.

Chen J., Fang Y., He K. & Du R. (2017). Charge-Depleting of the Batteries Makes Smartphones Recognizable, *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*, Shenzhen, p. 33-40. DOI : 10.1109/ICPADS.2017.00016.

Clapaud A. (2015). *Criteo, une architecture Big Data unique au monde*, Le Journal du Net, 10 mars 2013. En ligne : https://www.journaldunet.com/solutions/cloud-computing/1151178-criteo-une-architecture-big-data-unique-au-monde/.

Das A., Acar G., Borisov N. & Pradeep, A. (2018). The Web's Sixth Sense: A Study of Scripts Accessing Smartphone Sensors. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, p. 1515-1532.

Debize T., Anzala-Yamajako A., Soullié A., Billois G., Kokos A., Wolfhugel C. & Bloch L. (2016). *Sécurité informatique: Pour les DSI, RSSI et administrateurs*, Eyrolles.

Demiaux V. (2018). *De la CNIL au RGPD : 40 ans de protection des données (interview)*, L'Histoire, 25 mai 2018. En ligne : https://www.lhistoire.fr/entretien/de-la-cnil-au-rgpd%C2%A0-40-ans-de-protection-des-donn%C3%A9es.

Framablog (2017). *Comment les entreprises surveillent notre quotidien*, Framablog, 25 octobre 2017. En ligne : https://framablog.org/2017/10/25/comment-les-entreprises-surveillent-notre-quotidien/.

Ishtiaq A., Abbasi S. H., Aleem M., & Islam M. A. I. (2017). *User tracking mechanisms and counter measures*. International Journal of Applied Mathematics Electronics and Computers, 5(2), p. 33-40.

Kessous E. (2011), L'économie de l'attention et le marketing des traces, *Actes du colloque Web social, communautés virtuelles et consommation*.

Kessous E. (2012). *L'attention au monde. Sociologie des données personnelles à l'ère numérique*, Armand Colin.

Koch R., Golling M. & Rodosek G. D. (2013). Advanced geolocation of IP addresses. In *International Conference on Communication and Network Security (ICCNS)*, p. 1-10.

Kosinski M., Stillwell D. & Graepel T. (2013). *Private traits and attributes are predictable from digital records of human behavior*, PNAS April 9, 110(15) p. 5802-5805. En ligne : https://www.pnas.org/content/110/15/5802.

Lambrecht A. & Tucker C. (2013). *When Does Retargeting Work? Information Specificity in Online Advertising*. Journal of Marketing Research, 50(5), p. 561–576.

Mesguish V. & Thomas A. (2013). *Net recherche 2013*. De Boeck. ISBN : 978-2-8041-8228-1.

Morey T., Forbath T. & Schoop A. (2018). *Données clients : concevoir un système transparent de confiance*, Harvard Business Review, Printemps 2018, p. 64-74.

Nitot, T. (2016). *surveillance:///*, C&F éditions. ISBN : 978-2-915825-65-7.

Papadopoulos P., Kourtellis N. & Markatos E. (2019). Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*, p. 1432-1442.

Peyrat B. (2009). *La publicité ciblée en ligne*, CNIL. En ligne : https://www.cnil.fr/sites/default/files/typo/document/Publicite_Ciblee_rapport_VD.pdf.

Raphaël R. & Xi L. (2019). *Quand l'état organise la notation de ses citoyens. Bons et mauvais chinois*, Le Monde diplomatique, janvier 2019. En ligne : https://www.monde-diplomatique.fr/2019/01/RAPHAEL/59403.

Reichgut M. (2016), *Advertiser ID Tracking And What It Means For You*, Forbes, 16 mai 2016. En ligne : https://www.forbes.com/sites/onmarketing/2016/05/16/advertiser-id-tracking-and-what-it-means-for-you/#c8d03a118bf0.

Renaud J.-F. (2017). *Les achats programmatiques : comprendre les enjeux*, Gestion, 2017/2 (Vol. 42), p. 106-109. DOI : 10.3917/riges.422.0106. En ligne : https://www.cairn.info/revue-gestion-2017-2-page-106.htm.

Rochelandet F. (2010). *Économie des données personnelles et de la vie privée*, La Découverte, Paris.

Savchenko, I.I., Gatsenko, O.Y. (2015). *Analytical review of methods of providing internet anonymity*. Aut. Control Comp. Sci. 49, 696–700 (2015).

Sluis S. (2020). *Post-Cookie Apocalypse, IAB Unveils 'Project Rearc'*. AdExchenger, 11 février 2020. En ligne : https://www.adexchanger.com/ad-exchange-news/post-cookie-apocalypse-iab-unveils-project-rearc/.

Suire R. (2016). *GénérationY, GénérationZ, Génération A-nalphanète ? Portrait d'une cohorte d'étudiants en 2016*. M@rsouin, Université de Rennes.

Weide K. (2018), *Worldwide Digital Advertising Software Market Shares, 2017: Despite Intense M&A Activity, Still a Fragmented Market*, IDC, septembre 2018.

Zuboff S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. ISBN : 978-1610395694.