Du *tracking*, des contre-mesures et de leur efficacité dans la publicité ciblée

Tracking, countermeasures and their effectiveness in targeted advertising

Robert Viseur[1]

[1] Service TIC, Université de Mons (UMONS), Belgique, 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 évolué 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 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.

ABSTRACT. From a Web of documents of uncertain commercial interest, driven by pioneers believing in knowledge sharing, the Web later evolved into a collaborative, real-time form that was made profitable by advertising. The latter has evolved towards targeted advertising, including behavioral advertising based on the massive collection of usage traces. These traces come from various tracking devices including IP addresses (IP tracking), the now known cookies or fingerprints (e.g. browser fingerprinting and canvas fingerprinting). While the collection was initially limited to the workstation (mainly through the browser), it was later extended to smartphones and connected objects. This led to the trace marketing and attention economy that digital natives were confronted with at an early stage. Various countermeasures were gradually deployed by users (parameterization, extensions, e.g. ad blockers), by anonymization services (e.g. VPN and proxy), by the publishers themselves or by the regulator (e.g. RGPD). This paper proposes, on the one hand, a presentation of the structuring of the online advertising sector followed by a state of the art on the tracking tools deployed there, on the other hand, an inventory and analysis of the countermeasures deployed as well as their effectiveness. We show in particular the rapid evolution of the techniques used and the heterogeneity of the coverage offered by a priori equivalent protective devices.

MOTS-CLÉS. marketing des traces, économie de l'attention, adtech, publicité programmatique, publicité comportementale, privacy, tracking, big data.

KEYWORDS. targeted advertising, attention economy, adtech, programmatic advertising, behavioral advertising, 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 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 législateur (p. ex. réglementation pour la protection des données à caractère personnel). 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

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

standardisés, et de liens sponsorisés (*search*) au sein des moteurs de recherche (Allary et Balusseau, 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 d'actions (historique de visites de sites web, des mots-clefs rentrés dans les moteurs de recherche…), 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 vente).

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].

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.

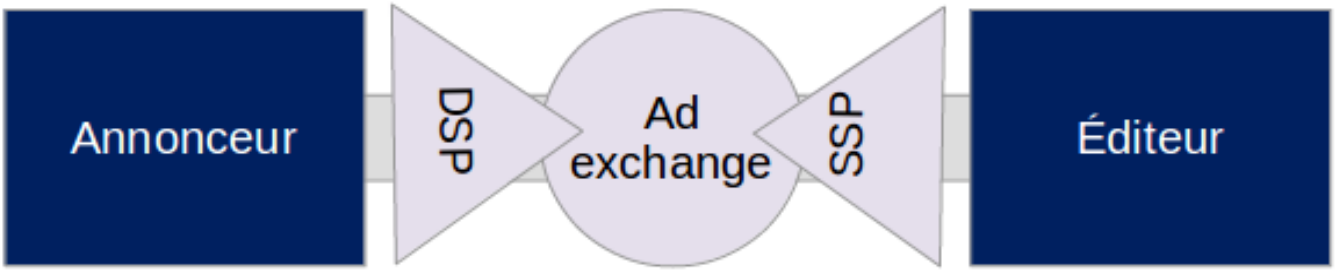

Figure 1. *Écosystème de la publicité programmatique (Allary et al., 2018)*

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 enchères en temps réel (RTB : *Real Time Bidding*) des espaces publicitaires disponibles (Renaud, 2017).

DSP	Ad exchange	SSP	Éditeur
Google Adwords		Google (*search*)	
Google Adwords		Google Adsense	Partenaires Adsense (*display*)
Search Ads 360 (ex-DoubleClick Search)	Google, Ads, Microsoft Advertising, Baidu...	Bing, Baidu, Google, Yahoo (*search*)	
Google Adwords, réseaux tiers	*DoubleClick Ad Exchange (Google)*	Google Display Network, Adsense, Youtube, portails, sites d'actualités….	
Facebook Ads			Facebook
AdRoll, AppNexus, Criteo, MediaMath...	Criteo, DoubleClick (Google), Rubicon Project...	AppNexus, OpenX, PubMatic, Rubicon Project...	Éditeurs (portails, sites d'actualités…)

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

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*). À 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'achats (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 concentrant plusieurs métiers de la publicité programmatique (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 inventaire récent des mécanismes de *tracking* est notamment proposé par Ishtiaq et al. (2017). 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. On parlera dans ce cas d'identifiants probabilistes (par opposition aux identifiants déterministes) car « *cette logique implique une marge d'erreur liée à l'application de règles statistiques* » (Allary et al., 2018). 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 faiblement peuplée et qu'il utilise une configuration atypique sur son terminal de connexion.

3.1. *Exploitation de l'adresse IP*

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 effet, 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.

3.2. *Exploitation des cookies*

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. On parlera de *first-party cookies* pour les *cookies* créés par le site web visité et de *third-party cookies* pour les *cookies* générés par un site tiers (p. ex. régies publicitaires ou outils d'analyse

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

de fréquentation). Ces *cookies* peuvent être produits par différents dispositifs incluant les pages web (p. ex. fonction *setcookie* en PHP), les scripts Javascript (via la propriété *document.cookie*) et les *web beacons* (p. ex. pixels invisibles).

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 après sa suppression en s'appuyant sur un autre dispositif de stockage, soit un *cookie* Flash (en réalité, un objet local partagé ou LSO, pour *Local Shared Objects*), 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 (d'autant que sa prise en charge par Adobe prendra fin au 31 décembre 2020). Les *evercookies* Javascript ont été imaginés par Samy Kamkar en 2010[8]. Ils s'appuient sur différents mécanismes de stockage permettant la création de *cookies* extrêmement persistants.

Les entreprises actives 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).

3.3. *Exploitation des empreintes*

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[9] 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).

3.4. *Exploitation des identifiants mobiles*

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*

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

8 Cf. https://github.com/samyk/evercookie.

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

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).

Les méthodes de *fingerprinting* ont également été adaptées aux 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). Les techniques de *fingerprinting* se sont donc diversifiées au fil du temps (Ishtiaq et al., 2017). Pour être pleinement efficaces, elles nécessitent une veille régulière et de complexes expérimentations (Kobusińska et al., 2018).

3.5. *Exploitation des applications*

Le poste de travail fait l'objet d'une collecte de données grâce au système d'exploitation voire aussi des applications installées. Les activités de télémétrie sont ainsi critiquées, en particulier dans le cas de Windows 10 (mais aussi des versions postérieures à Windows 7 après mise à jour). Cette fonctionnalité permet aux administrateurs systèmes, par l'analyse des fichiers de télémétrie, de comprendre les causes d'un incident ou l'origine d'une cyberattaque (Hang et al., 2020) et à Microsoft, de comprendre la cause d'un dysfonctionnement mais aussi la nature des usages permettant d'orienter les choix lors de l'ajout de fonctionnalités. S'il a fait l'objet de débats parfois houleux, ce système n'est en aucun cas prévu pour le *tracking* commercial. Il est paramétrable (4 niveaux) et a fait l'objet d'efforts d'anonymisation documentés par les équipes de Microsoft (Ding et al., 2017). Si les cas de collecte à partir des applications sont plus rarement évoqués, ils ne sont pas inexistants comme le montrent certaines affaires de revente de données à caractère personnel éventuellement anonymisées (cf. par exemple la filiale Jumshot de l'éditeur antivirus Avast).

Par ailleurs, l'utilisation d'Internet s'est substantiellement déplacée vers les terminaux mobiles (*smartphones*, tablettes...). Ces derniers représentent ainsi en 2020 plus de 50% du trafic web mondial (source : Statcounter). De plus, ils présentent des particularités techniques et sont soumis à une activité très importante de collecte de données via des *trackers* publicitaires intégrés aux *apps*. Ces *trackers* héritent des droits accordés aux applications lors de leur installation et bénéficient ainsi d'un statut privilégié comparé aux *trackers* utilisés pour les navigateurs (Stevens et al., 2012). Binns et al. (2018) montrent que 88,44% des *apps* Android possèdent au moins un *tracker* Alphabet Inc. (Google), pour la publicité ciblée mais aussi parfois l'analyse de performance, 42,55%, au moins un *tracker* Facebook et 22,75%, au moins un *tracker* Microsoft. *Last but not least*, via Android et ses logiciels pré-installés tels que Google Maps, Google dispose d'autres canaux de collecte de données, notamment pour la géolocalisation des terminaux (Nitot, 2016 ; Khatoon et al., 2017). Les applications mobiles accèdent en pratique à de nombreuses données à caractère personnel, incluant notamment la localisation et les contacts (Khatoon et al., 2017), susceptibles d'être ensuite partagées avec des tiers, y compris pour des *apps* liées à la santé (Blenner et al., 2016), alors même que les politiques de confidentialité, pour autant que l'on prenne la peine de les lire, ne sont pas exemptes de problèmes (Yu et al., 2018). Plusieurs composants distincts contribuent dès lors à la collecte de données : le système d'exploitation (et les éventuelles applications imposées), les applications tierces et les *trackers* (publicitaires et/ou *analytics*) inclus dans ces applications.

4. Inventaire des contre-mesures

Les consommateurs disposent de plusieurs contre-mesures face au *tracking* incluant les navigateurs web, soit par leur configuration, soit par leur extension, le choix d'une plate-forme davantage respectueuse de la vie privée et l'utilisation d'outils d'anonymisation. La législation (RGPD) offre par ailleurs une protection générique imposant au minimum l'information de l'utilisateur quant aux traitements de données mis en œuvre.

4.1. *Configuration du navigateur*

La Fondation Mozilla, qui produit le navigateur Firefox, met en avant depuis plusieurs années son engagement pour le respect de la vie privée. Ce navigateur en est actuellement à la version 81.0.1 (publiée le 01 octobre 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, les 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[10]. Troisièmement, Firefox permet le blocage de *trackers* au travers de la fonctionnalité *Enhanced Tracking Protection* (ETP) accessible depuis la barre d'adresse[11]. Parmi les options offertes par les navigateurs web, le mécanisme « *Do Not Track* » (DNT[12]), s'il a suscité quelques espoirs de mise en place d'un mécanisme d'*opt-in* concernant la collecte de données à caractère personnel, a finalement été abandonné avec notamment, d'une part, la fermeture du groupe de travail dédié au sein du W3C (*Tracking Protection Working Group*) et, d'autre part, l'abandon pur et simple par Safari à partir de la version 12.1 (au profit de la technologie propriétaire ITP[13]). Pour un historique plus complet du mécanisme DNT, lire Fleishman (2019).

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 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. *Choix de la plate-forme*

Le choix de la plate-forme n'est pas neutre du point de vue de la propension à tracer ou non les utilisateurs. Apple, retiré du marché de la publicité (iAD) depuis 2016, tend ainsi à se distinguer sur le plan du respect de la vie privée. D'une part, le navigateur Safari, proposé par Apple, s'est précocement distingué par sa fonctionnalité baptisée *Intelligent Tracking Prevention* (ITP) offrant

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

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

12 Cf. https://www.w3.org/TR/tracking-dnt/.

13 Cf. https://developer.apple.com/documentation/safari-release-notes/safari-12_1-release-notes.

un dispositif anti-pistage similaire à l'ETP de Firefox et conduisant à un taux de blocage des *cookies* particulièrement élevé sur les plates-formes Apple. D'autre part, Apple a annoncé le déploiement dans son système d'exploitation iOS 14 d'un mécanisme explicite de consentement pour l'accès à l'IDFA dans les applications (« *Asking Permission to Track* »), suscitant notamment la réprobation de Facebook et un report de la décision[14].

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[15], 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 (Ishtiaq et al., 2017 ; Mulazzani et al., 2013 ; 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). Les *evercookies* Javascript ont cependant été utilisés pour tracer les utilisateurs de Tor (Kobusińska et al., 2018). La combinaison des *evercookies*, du *fingerprinting* et de listes d'adresses IP a par exemple pu permettre la conduite d'attaques sophistiquées de désanonymisation sur les utilisateurs de Tor (Mulazzani et al., 2013).

4.5. *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 que des entreprises[16]. Elle met d'ailleurs progressivement en place une politique de la carotte et du bâton faite de sanctions ou de récompenses appliquées aux citoyens sur base d'un système complexe de crédit social au sein duquel chaque citoyen chinois est associé à un score de réputation (Liang et al., 2018 ; 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 à caractère personnel.

Le RGPD[17] repose notamment sur des principes de consentement éclairé et de proportionnalité des données collectées au regard des finalités du traitement telles que communiquées à l'utilisateur. La notion de données à caractère personnel est large puisqu'elle inclut des données directement

14 Cf. https://developer.apple.com/app-store/user-privacy-and-data-use/ et https://www.facebook.com/business/news/preparing-our-partners-for-ios-14-launch/ pour plus d'informations.

15 Une étude publiée par VPNpro (2019) révélait ainsi que plusieurs dizaines de services VPN étaient en réalité hébergés par des entreprises chinoises tandis qu'une proportion importante était opérée par des entreprises situées hors Union européenne.

16 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.

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

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. Google LLC a ainsi d'ores et déjà fait l'objet d'une sanction de 50 millions d'euros décidée par la CNIL le 21 janvier 2019 (CNIL, 2019). Et près de 150 milles plaintes et signalements ont été envoyés aux autorités de protection des données au cours des 12 premiers mois du dispositif[18].

5. Discussion

Nous discuterons ici trois points spécifiques : la sensibilisation des utilisateurs face à la collecte de données, l'efficacité des différentes contre-mesures proposées et les conséquences de l'entrée dans une ère post-*cookies*.

5.1. *Sensibilisation des utilisateurs*

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.

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

18 Cf. https://ec.europa.eu/info/sites/info/files/infographic-gdpr_in_numbers.pdf.

5.2. *Efficacité des contre-mesures*

Le Tableau 3 propose une synthèse de contre-mesures courantes et analyse leur efficacité au regard des techniques de *tracking* et des contre-contre-mesures (*sic*) potentiellement mises en place par les éditeurs de sites web (cf. « Réaction de l'éditeur »). En pratique, le navigateur Firefox permet la mise en place, à la configuration, d'un large éventail de dispositifs pour limiter le pistage incluant l'effacement des *cookies*, la détection des calculs d'empreintes, le blocage de *trackers* et l'envoi d'en-têtes HTTP « *Do Not Track* ». Il peut en outre être complété par des extensions spécialisées, par exemple dans le filtrage des publicités.

	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	±	±	?	✔	✔
Réaction de l'éditeur	Aucune mais inconfort….	Détection et blocage⑤	Détection et blocage⑤	Détection et blocage⑤	Application partielle⑥

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

①Les *cookies* peuvent être facilement refusés et effacés, de manière manuelle ou automatique, à l'aide d'un navigateur web. Configurer l'acceptation des *cookies* demande cependant du temps, que le recours aux « *cookie walls* » tend à aggraver. D'où sans doute le découragement de certains utilisateurs qui en viennent parfois à les accepter systématiquement (p. ex. extension « *I don't care about cookies* »). La navigation privée permet de systématiser l'effacement des *cookies* à la fermeture de l'onglet de navigation.

②Les bloqueurs de publicité permettent de limiter 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, ils ne bloquent pas d'autres types de *trackers* (p. ex. Google Analytics). Pour ces derniers, des extensions spécialisées doivent être installées au cas par cas (p. ex. extensions « Désactivation de Google Analytics » et Ghostery). Firefox, depuis la version 67.0.1, permet par ailleurs de configurer le blocage de *trackers* via la fonctionnalité *Enhanced Tracking Protection* (ETP)[19].

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

③ Le *canvas fingerprinting* est bloqué par Firefox depuis la version 58. L'énumération d'extensions (*plugins*) y est par ailleurs limitée[20]. Certaines extensions luttent également contre cette technique (p. ex. Privacy Badger).

④Le 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.

⑤La 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 et l'utilisation de Tor.

⑥L'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). L'impact réel du RGPD sur la collecte de données reste ainsi débattu. Sørensen et Kosta (2019) notent par exemple que la réduction constatée, sur certaines catégories de sites web, du nombre de serveurs tiers rencontrés par un internaute peut s'expliquer par différents facteurs incluant l'entrée en vigueur du règlement européen mais aussi des phénomènes de concentration autour de certaines plates-formes publicitaires dans l'écosystème *adtech* (p. ex. *header bidding*, cf. IAB, 2018).

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* »). La politique d'Apple s'avère cependant plus claire et visible dans la pratique : selon une étude FlashTalking (cf. « *Cookie Rejection Report 2020* »), le taux de blocage sur iOS serait ainsi de 68 % contre 30 % sur Android et 7 % chez Microsoft.

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	✓	✓

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

Fonctionnalités de blocage		Chrome	Firefox
		(« Désactivation de Google Analytics » + Ghostery, Privacy Badger...)	(Ghostery, Privacy Badger...)
Do Not Track		✓ (désactivé)	✓ (désactivé)

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

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 utilisées, les contre-mesures disponibles disposent d'une portée et d'une réelle efficacité, même si elles peuvent elles-mêmes être contrées par les éditeurs de sites web (p. ex. entrave à l'affichage d'une page en cas d'utilisation d'un bloqueur de publicités). Le tableau 5 compare les principaux navigateurs du marché (ont été exclues des solutions, souvent orientées *privacy*, mais faiblement diffusées, comme Brave ou Vivaldi). Deux sont *open source* : Firefox (sous licence MPL) et Chromium (sous licence BSD principalement). Plusieurs recourent à des technologies *open source* mais sont publiés sous licence propriétaire. Il s'agit de Safari, avec le moteur de rendu Webkit (sous licences LGPL et BSD), ainsi que de Chrome, Edge (depuis 2019) et Opera (depuis 2013), avec le navigateur Chromium. Ce dernier, intégrant le moteur de rendu Blink, *fork* de WebKit, contribue à asseoir le pouvoir de Google sur le Web. Le choix des internautes en matière de navigateur web ne reflète pas nécessairement l'investissement des éditeurs en matière de *privacy*. La montée en puissance de Chrome, servi par la puissance commerciale de Google, s'est ainsi faite au détriment des différents navigateurs Microsoft (Internet Explorer et Edge) mais aussi de Firefox. Deux autres ont fait de la *privacy* un cheval de bataille (Safari et Firefox) alors que deux autres dépendent au contraire des revenus tirés de la publicité (Google et Opera). La plupart sont régulièrement mis à jour (p. ex. sécurité). L'innovation sur les navigateurs peut porter sur le moteur de rendu (Gecko, Webkit, Blink...) ou sur les fonctionnalités du navigateur. Trois navigateurs se distinguent sur ce plan : Firefox, Chrome et Opera. Le cas d'Opera mérite un commentaire plus approfondi. Ayant abandonné son moteur de rendu (Presto) pour Chromium, Opera a réussi à maintenir un rythme rapide d'innovation en ajoutant de nombreuses fonctionnalités originales. Ces dernières incluent notamment un VPN. Opera met aussi en avant la sécurisation de la vie privée alors que ses revenus dépendent de la publicité (source : rapport annuel 2019 d'Opera Limited) et que l'entreprise dépend aujourd'hui d'un fond chinois (Golden Brick Capital), suscitant la perplexité de certains observateurs quant à la confiance à accorder à ce logiciel. La position de Microsoft reste pour sa part assez floue. L'éditeur a cherché à se relancer dans la course en faisant tardivement le choix de Chromium comme base technique et conserve des activités dans le domaine de la publicité en ligne.

	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 %	⋆ ⋆ ⋆	⋆ ⋆ ⋆	✓	⋆

Tableau 5. *Comparaison des navigateurs.*

Les extensions pour les navigateurs sont nombreuses et peuvent être évaluées suivant différents critères tels que la couverture, la compatibilité et le confort d'utilisation (cf. Tableau 6 pour quelques exemples). La colonne « Intérêt » reflète la couverture (variété des menaces prises en charge) et l'efficacité (efficacité du filtrage) des solutions évaluées. 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 de l'extension ainsi que de ses liens avec le marché de la publicité ciblée (p. ex. Adblock Plus). Par ailleurs, deux technologies émergent : celles utilisant des listes noires de *trackers* (p. ex. Ghostery) et celles, fonctionnant par apprentissage, capables de détecter dynamiquement les scripts problématiques, apportant donc un filtrage plus nuancé (p. ex. Privacy Badger[21]).

21 Cf. https://privacybadger.org/#faq.

Extension	Automa-tique	Couverture	Configu-ration	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	Détection (épisodique) par les sites	⋆ ⋆ ⋆

Tableau 6. *Efficacité des extensions.*

5.3. *Conséquences de l'ère post-cookie*

Comme nous l'avons montré, l'utilisation des *cookies* à des fins de ciblage publicitaire est de plus en plus contrariée par les éditeurs de navigateurs web. En effet, l'environnement Apple se distingue par un taux de blocage très élevé des *cookies* tandis que Safari et Firefox ont déployé des dispositifs de filtrage des *cookies* tiers, que Google a annoncé ne plus supporter dans Chrome à partir de 2022. Dès 2022, c'est donc plus ou moins 90% du trafic qui fera l'objet d'un filtrage des *cookies* tiers, une menace notamment documentée dans le rapport annuel de Criteo, licorne française spécialisée dans la publicité programmatique. Les conséquences portent sur trois dimensions : le ciblage de la publicité, la mesure de performances et la concurrence sur le marché de la publicité digitale.

Tout d'abord, le ciblage de la publicité en ligne repose en grande partie, pour la publicité web du moins, sur l'utilisation de *cookies*, en ce inclus des *cookies* tiers permettant le partage d'informations entre acteurs du secteur. Le rejet systématique des *cookies* entraîne donc un risque de réduction de la pertinence des publicités du fait de l'impossibilité de suivre le cheminement des utilisateurs sur le Web et d'en déduire leurs préférences à court, moyen ou long terme. Cette tendance impose donc aux entreprises du secteur de réfléchir à des solutions alternatives : meilleure valorisation des données propriétaires (*first-party*) mais aussi développement de techniques plus insidieuses comme le recours aux identifiants matériels (p. ex. *device* ID) ou aux actions marketing (p. ex. concours et tests de personnalité). Ensuite, la mesure des performances des actions marketing en ligne s'appuie également sur les *cookies*. Cela concerne des outils *analytics* comme Google Analytics mais aussi les régies publicitaires dès lors qu'elles mesurent l'efficacité d'une campagne ou d'un support promotionnel (p. ex. calcul du *capping* et du *reach*[22]). Le rejet des *cookies* conduit ainsi à une surestimation du taux de couverture d'une campagne du fait de la surévaluation du nombre de visiteurs uniques et à un risque d'exposition excessive des utilisateurs à une même campagne en

22 Le *capping* permet de fixer une limite au nombre d'impressions d'une publicité pour un *cookie* donné (Allary et al., 2018). Le *reach* désigne pour sa part la couverture d'une campagne, c'est-à-dire le pourcentage d'individus appartenant à la cible ayant été exposé à la campagne publicitaire.

l'absence d'historique. Enfin, l'abandon des *cookies* impacte fortement les conditions de concurrence sur le marché publicitaire.

Catégorie	Poids	Google	Facebook	Microsoft	Autres
Search	45 %	± 90 %	0 %	< 10 %	< 5 %
Display (social)	22 %	0 %	± 90 %	< 5 %	<10 %
Display (hors social)	18 %	± 80 %	0 %	nc	nc
Autres	15 %	0 %	0 %	0 %	± 100 %

Tableau 7. *Estimation des parts de marché (publicité digitale). Chiffres repris ou estimés par l'auteur sur base de données publiées par l'Observatoire Syndicat des régies Internet et PwC (en partenariat avec l'Udecam).*

En effet, le marché de la publicité digitale est marqué par une double réalité. D'une part, il est dominé par deux acteurs (Facebook et Google) détenant ensemble 75 % du marché français (cf. Tableau 7). Les autres acteurs du marché sont très fragmentés et collaborent grâce à l'échange et la synchronisation de *cookies* tiers (*cookies syncing*). Facebook et Google présentent cependant des situations distinctes, qui vont faire l'objet d'une explicitation dans les deux paragraphes suivants.

Les revenus de Facebook proviennent pour 98,5 % de la publicité en ligne (source : rapport annuel 2019). Pour le ciblage, l'entreprise s'appuie en grande partie sur les données collectées au sein de son propre réseau (incluant Facebook mais aussi Instagram et Messenger) fort de 2,5 milliards (hors Instagram) d'utilisateurs actifs mensuels. Le profilage est donc réalisé à l'aide de données propriétaires (*first-party*) et peut s'appuyer sur une identification déterministe. Entre autres choses, l'utilisation des boutons sociaux (p. ex. *Likes*) permet une segmentation efficace des utilisateurs (p. ex. classification dichotomique, cf. Kosinski et al., 2013). L'entreprise est dès lors peu impactée par la transition vers une ère post-*cookies*. Les scandales liés à l'exploitation indue des données (p. ex. Cambridge Analytica ; Isaak et Hanna, 2018) et l'intérêt croissant pour la *privacy* en Europe (p. ex. RGPD et remise en cause du *Privacy Shield*[23]) ont un impact à long terme potentiellement plus important.

23 Cf. Arrêt remis par la Cour de justice de l'Union européenne (CJUE) le 16 juillet 2020 (Document 62018CJ0311).

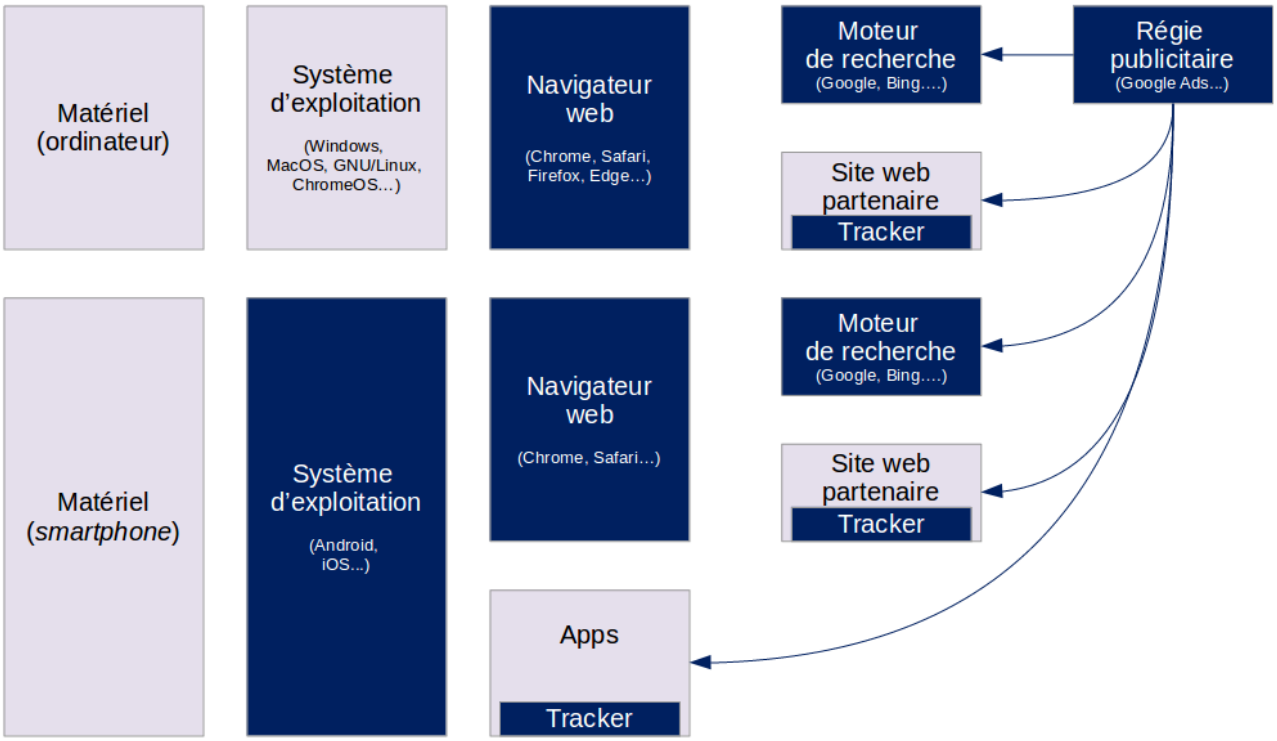

Figure 2. *Contrôle de Google sur le marché publicitaire.*

La part des revenus publicitaires dans le chiffre d'affaires de Google s'élevait à 83,9 % en 2019 (source : rapport annuel 2019 d'Alphabet Inc.), en légère baisse comparativement aux deux années précédentes. Google dispose d'une position dominante sur le marché des liens sponsorisés dans les moteurs de recherche du fait de la position dominante de son moteur de recherche. Il peut également s'appuyer sur ses autres services, notamment Youtube (Google Web Search et Youtube sont les deux sites les plus visités au monde selon le Top Sites Alexa). Le suivi des utilisateurs recourt à un mélange d'identification déterministe (compte Google, *device* ID, *cookie*...) et probabiliste (adresse IP…). Google peut donc s'appuyer sur un volume important de données propriétaires et dépend moins que les acteurs indépendants de l'utilisation des *cookies*, tout particulièrement des *cookies* tiers. Au-delà de cette seule dimension, Google sécurise par ailleurs à la fois l'expérience publicitaire (par le contrôle du logiciel de navigation) et la chaîne de collecte de données (par le contrôle du système d'exploitation mobile et du navigateur). La Figure 2 met en évidence (blocs bleus foncés) les composantes impactant l'expérience publicitaire sur lesquelles Google dispose d'une position dominante (système d'exploitation mobile, navigateur web, *trackers analytics*, publicitaires et mobiles...). Les contextes du poste de travail et du *smartphone* y sont distingués de manière à mettre en évidence la forte emprise de Google sur les terminaux mobiles.

Conclusion

Notre recherche 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 l'activité des utilisateurs sur Internet (en particulier pendant 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 deux limitations. Premièrement, l'analyse des possibilités de configuration ou d'extension des navigateurs s'est focalisée sur le navigateur Firefox, avec une

extension à Safari et à Chrome. Une analyse similaire devrait donc être réalisée pour les éditeurs d'autres navigateurs web (p. ex. Microsoft Edge et Opera). L'angle mort est cependant limité puisque les parts de marché combinées de Chrome, Safari et Firefox dépassent 80 % que ce soit en France ou au niveau mondial (source : Statcounter). Deuxièmement, alors que l'utilisation d'Internet s'est substantiellement déplacée vers les terminaux mobiles (*smartphones*, tablettes…), 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).

Six points nous paraissent constituer des perspectives intéressantes à cette recherche. Premièrement, si elle prend en compte le poste de travail et les *smartphones*, elle n'englobe pas la question du *tracking* au niveau des objets connectés. Ces derniers incluent notamment des dispositifs portables de suivi de la condition physique (p. ex. Fitbit, Garmin et Withings) conçus explicitement pour les collectes de données (Zhou et Piramuthu, 2014) dont l'usage peut être facilement détourné. Comment dès lors les utilisateurs peuvent-ils contrôler valablement les traitements réalisés sur ces données à caractère personnel ? Deuxièmement, nous avons vu que l'utilisateur était un acteur clef dans la modération des activités de *tracking* mais que ce dernier n'était pas toujours conscient de la nature de ce traçage. Comment dès lors pourrait-on élaborer une méthodologie outillée permettant de mesurer simplement 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 ? Troisièmement, l'entrée en application du RGPD a vu l'émergence des *cookie walls* souvent fastidieux à configurer, conduisant à l'émergence d'un consentement de moins en moins éclairé. La recherche de techniques de *tracking* alternatives confrontent par ailleurs les consommateurs à des concepts peu familiers. Comment dès lors conserver ou développer le caractère éclairé du consentement lors de la navigation en ligne ? Quatrièmement, le capitalisme de surveillance (Zuboff, 2019) a été, au travers du *tracking*, essentiellement analysé du point de vue des pratiques commerciales liées à la publicité digitale. La dimension sécuritaire et, surtout, la collaboration étroite entre les acteurs privés et publics a été peu ou prou développée. L'exemple du Social Credit System chinois montre cependant la perméabilité existant entre ces deux sphères et les contributions du privé tant pour le support technique que pour l'apport de nouvelles données (Liang et al., 2018). Dans le contexte sécuritaire actuel (terrorisme, pandémie...), et malgré les gardes-fous constitutionnels ou réglementaires existant en Europe et dans une moindre mesure aux États-Unis, un tel risque de création d'une infrastructure de surveillance d'état est-elle envisageable ? Cinquièmement, nous avons vu que la protection des citoyens et l'opposition à certains outils de *tracking* (les *cookies*, en particulier les *cookies* tiers) pouvaient paradoxalement conduire à un renforcement des GAFA(M) sur les marchés locaux de la publicité digitale à la fois sur un plan purement commercial (position dominante du point de vue de la part de marché) et en matière de sécurisation de l'accès aux données sur l'ensemble de la chaîne de valeur publicitaire. Comment peut-on optimiser et, si possible, mesurer cette emprise sur le marché de la publicité digitale ? Sixièmement, le rejet de la publicité en ligne par les utilisateurs menace un modèle économique traditionnel des éditeurs de contenus. Quels sont les modèles économiques alternatifs (p. ex. micro-donations ponctuelles ou récurrentes) praticables par ces éditeurs ? Que sait-on de leur efficacité ? S'accompagnent-ils d'un arrêt (ou d'une diminution) du *tracking* ?

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*, pp. 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 N°. 15/801,971.

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

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

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.

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*, pp. 23-31.

Blenner, S. R., Köllmer, M., Rouse, A. J., Daneshvar, N., Williams, C., & Andrews, L. B. (2016). *Privacy policies of android diabetes apps and sharing of health information*. Jama, 315(10), pp. 1051-1052.

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, pp. 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/.

CNIL (2019). *Le Conseil d'État valide la sanction prononcée à l'encontre de la société Google LLC* . CNIL. En ligne : https://www.cnil.fr/fr/le-conseil-detat-valide-la-sanction-prononcee-lencontre-de-la-societe-google-llc.

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*, pp. 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.

Ding B., Kulkarni J., & Yekhanin S. (2017). *Collecting telemetry data privately*. In Advances in Neural Information Processing Systems, pp. 3571-3580.

Fleishman G. (2019). *How the tragic death of Do Not Track ruined the web for everyone*. Fast Company. En ligne : https://www.fastcompany.com/90308068/how-the-tragic-death-of-do-not-track-ruined-the-web-for-everyone.

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/.

Han J., Park J., Chung H., & Lee S. (2020). *Forensic analysis of the Windows telemetry for diagnostics*. arXiv preprint arXiv:2002.12506.

IAB (2018). *Header bidding and auction dynamics White paper*, IAB Europe, Auguts 2018. En ligne : https://www.iab.it/wp-content/uploads/2018/09/IAB-Europe_Header-Bidding-and-Auction-Dynamics-White-Paper_August-2018-1-compressed.pdf.

Isaak J., & Hanna M.J. (2018). *User data privacy: Facebook, Cambridge Analytica, and privacy protection*. Computer, 51(8), pp. 56-59.

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), pp. 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.

Khatoon A., & Corcoran P. (2017). Privacy concerns on Android devices. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 149-152.

Kobusińska A., Pawluczuk K., & Brzeziński J. (2018). *Big Data fingerprinting information analytics for sustainability*. Future Generation Computer Systems, 86, pp. 1321-1337.

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), pp. 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), pp. 561–576.

Liang F., Das V., Kostyuk N., & Hussain M. M. (2018). *Constructing a data-driven society: China's social credit system as a state surveillance infrastructure*. Policy & Internet, 10(4), pp. 415-453.

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, pp. 64-74.

Mulazzani M., Reschl P., Huber M., Leithner M., Schrittwieser S., Weippl E., & Wien F. C. (2013). Fast and reliable browser identification with javascript engine fingerprinting. In *Web 2.0 Workshop on Security and Privacy* (W2SP).

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.

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, pp. 696-700.

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/.

Sørensen J., & Kosta S. (2019). Before and after gdpr: The changes in third party presence at public and private european websites. In *The World Wide Web Conference*, pp. 1590-1600.

Stevens R., Gibler C., Crussell J., Erickson J., & Chen H. (2012). Investigating user privacy in android ad libraries. In *Workshop on Mobile Security Technologies* (MoST).

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

VPNpro (2019). *The few behind the many: hidden VPN owners unveiled*. En ligne : https://vpnpro.com/wp-content/uploads/Infographic-VPNpro-97-VPN-products-run-by-just-23-companies.pdf.

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

Yu L., Luo X., Chen J., Zhou H., Zhang T., Chang H., & Leung H. K. (2018). *PPChecker: Towards Accessing the Trustworthiness of Android Apps' Privacy Policies*. IEEE Transactions on Software Engineering.

Zhou W., & Piramuthu S. (2014). Security/privacy of wearable fitness tracking IoT devices. In *2014 9th Iberian Conference on Information Systems and Technologies* (CISTI) (pp. 1-5). IEEE.

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

