iE1.[596]

MINES ParisTech
Bibliothèque

Les empreintes numériques

Mémoire de troisième année du Corps des Mines

par Romain Delassus, Rémi Leblond et David Torrin
Ingénieurs des Mines, promotion 2010

2013

Avant-propos

• • •

Ce mémoire a été réalisé durant l'année scolaire 2012-2013, dans le cadre de la troisième et dernière année de formation du Corps des Mines.

Il a été élaboré en grande partie grâce à des entretiens, et nous souhaitons remercier ici toutes les personnes qui ont accepté de s'entretenir avec nous au long de cette année, et dont la rencontre fut riche et passionnante. Pour plus d'exhaustivité, nous les avons citées à la fin de ce mémoire.

Nous adressons un remerciement tout particulier à François Comet, ainsi qu'à notre pilote Jean-Pierre Dardayrol, pour leur aide précieuse.

Ce document est organisé en quatre parties. Dans la première, nous nous attacherons à identifier les caractéristiques faisant d'une information une donnée personnelle. Nous montrerons que ces caractéristiques dépendent de nombreux paramètres, rendant la notion même de donnée personnelle instable dans le temps et dans l'espace. Dans une époque marquée par une capacité de calcul toujours croissante et par une multiplication des données collectées et mises en relation, l'anonymisation de données nous apparait comme un mythe détournant dangereusement de l'analyse des risques potentiels. Ceci justifie d'élargir le débat sur la protection des données à l'ensemble des empreintes numériques, c'est à dire à toute donnée issue de l'utilisation d'un service Internet.

La deuxième partie se penchera sur les méthodes utilisées aujourd'hui par les entreprises pour produire, collecter et valoriser nos empreintes numériques. Par une analyse des modèles économiques actuels de l'économie numérique, nous montrerons que nos données servent de monnaie alternative dans cet écosystème. Nous décrirons en particulier le fonctionnement du marché de la publicité en temps réel, certainement le modèle le plus abouti de valorisation d'empreintes numériques à l'heure actuelle. Mais si ce marché innovant est fortement créateur de valeur, nous montrerons que son opacité a pour conséquence directe une surexploitation non efficace de nos empreintes.

La troisième partie sera consacrée à l'analyse des risques liés à l'exploitation de nos données : Big Brother et le risque de la société de surveillance bien sûr, mais plus particulièrement les risques liés à la protection des consommateurs.

La dernière partie présentera les diverses pistes, privées ou publiques, envisageables pour encadrer les pratiques d'exploitation des empreintes numériques. Nous montrerons que toutes ces pistes ont leurs limites : la politique réglementaire sera toujours en retard sur la technologie, et le laissez-faire du marché ne peut conduire, dans les conditions actuelles, qu'à une surexploitation des données dangereuse pour les consommateurs.

Enfin, nous présenterons deux pistes actuellement délaissées et peu explorées par les pouvoirs publics. La première est celle de l'éducation et de la formation des citoyens, moyen élégant de surmonter l'inefficacité actuelle du marché. Nous proposons la mise en place d'une véritable école française de recherche sur les empreintes numériques dont la mise en place dès aujourd'hui aurait un double intérêt. A court terme, ceci favorisait la réflexion des citoyens sur ces enjeux, réflexion stratégique à la vue de l'importance qu'ont pu prendre les lobbies dans les discussions autour du prochain règlement européen. Mais surtout à moyen terme, cette expertise deviendrait un enjeu économique exportable dans des domaines aussi variés que le service à la personne, les transports, la santé, l'environnement… La seconde est un soutien public volontariste aux initiatives respectueuses dont nous développons un exemple : la MarinièreBox. Plutôt que de considérer ces problématiques comme des états de fait subis que la société doit apprendre à gérer, nous pouvons choisir d'en faire une opportunité stratégique. Les modèles de protection restent en grande partie à définir et à développer, et nous nous trouvons à la naissance d'un secteur qui est voué à devenir très porteur à l'avenir.

Table des matières

INTRODUCTION **5**

DONNEES PERSONNELLES ET DONNEES NON PERSONNELLES ? **7**

DONNEE PERSONNELLE OU DONNEE SENSIBLE ? 8
LE MYTHE DE L'ANONYMISATION 9
LA DIFFERENCE ENTRE NOS IDENTITES NUMERIQUES ET CLASSIQUES 11
LA VIE PRIVEE, ET PUIS C'EST TOUT ? 12

COMMENT SONT PRODUITES LES EMPREINTES NUMERIQUES, ET POUR QUELS USAGES ? **13**

L'ECONOMIE NUMERIQUE DOMINEE PAR LE TOUT GRATUIT 13
Nos données sont-elles la nouvelle monnaie d'échange de l'économie numérique ? *14*
Payons-nous notre boite e-mail avec le contenu de notre correspondance ? *14*
Les modèles d'affaires multi-faces *15*
Et mon profil Facebook, combien vaut-il ? *16*
DE LA COLLECTE A L'IDENTIFICATION 19
Des empreintes souvent issues d'une contribution bénévole *19*
Les différents modes de collecte *21*
Nos données sont mortes, seuls les traitements ont de la valeur *23*
De multiples méthodes d'identification *24*
LES UTILISATIONS ET MONETISATIONS ACTUELLES 26
Quelles informations sont récoltées ? *26*
Le marché de la publicité en temps réel *27*
La création de nouveaux services *31*
Vendre des données ? *32*

ON NE FAIT PAS D'OMELETTE SANS CASSER DES ŒUFS **34**

BIG BROTHER IS WATCHING YOU 34
Vers la disparition de la vie privée ? *34*
La vie privée, un problème de vieux cons ? *35*
LA PERSONNALISATION ANONYME 37
Big Brother, l'arbre qui cache la forêt *37*
Le behavioural pricing *38*
Le narcissisme 2.0 ..... *45*
UN RISQUE POUR LE *BUSINESS* 50
La CNIL, fossoyeuse des start-ups innovantes ? *50*
Un marché de la publicité en difficulté *51*
Emergence de modèles respectueux de la vie privée *53*

ALORS FINALEMENT, QUE DOIT-ON FAIRE ? **55**

REGULATION PAR LE MARCHE 56

Autorégulation : laissons faire la main invisible du marché *56*

Concurrents indirects et problématiques concurrentielles *57*

LE POUVOIR DES UTILISATEURS 58

Des utilisateurs souverains ? *58*

L'importance de la sensibilisation *59*

Les solutions de contournement *59*

LES INITIATIVES PUBLIQUES 60

Le rôle de l'Etat *60*

Appréhender les limites de la voie réglementaire *60*

FAUT-IL PLUS DE REGULATION ? 62

DE L'IMPORTANCE DE LA FORMATION ET DE LA RECHERCHE 63

Pourquoi former les utilisateurs ? *63*

Comment former les utilisateurs ? *64*

L'ECOLE DE LA SECURITE, NOUVEAU DOMAINE DE L'EXCELLENCE FRANÇAISE ? 65

CONCLUSION **69**

ANNEXE. PERSONNES RENCONTREES **70**

Introduction

20.06.2025, 06h30.

Dès son réveil, Gaspard met ses web glasses et se branche sur sa chaîne d'information personnelle. Pratique, ça, une chaîne de télévision qui prend en compte ce qu'il aime et ce qui l'intéresse et ne lui transmet que des informations compatibles. Merci Google pour ce service gratuit, que de temps gagné !

Pause pub. *Ariel, la première lessive 5 en 1 qui lave sans eau !* Ça tombe bien, Gaspard est à court de lessive depuis hier, c'est sa machine qui lui a envoyé un mail. C'est vraiment bien Ariel. Il n'achète plus que ça. Ah tiens, le prix a encore augmenté en revanche. Comme celui du pain et de l'électricité hier. C'est pénible ! Heureusement qu'il vient de recevoir une augmentation de salaire la semaine dernière.

A bien y réfléchir, il faudra qu'il en parle à ses amis, qui n'ont pas l'air de se plaindre, eux. Presque comme s'ils ne s'étaient pas rendu compte de ces changements. Enfin, il attendra le prochain afterwork Facebook, parce que sortir, ces derniers temps, il évite. Son émission d'info lui fait un rapport tous les matins sur les cochonneries qui traînent dehors, et ça ne lui donne vraiment pas envie. Surtout qu'il n'a pas pu renouveler sa couverture santé. Il n'a d'ailleurs pas très bien compris pourquoi à la dernière renégociation de contrat, sa mutuelle a augmenté de 400%. C'est bête, il voulait justement aller voir un cardiologue, son père vient de faire un infarctus.

En ce moment, Gaspard se sent vraiment en phase avec la société. C'est simple, l'immense majorité des gens pensent comme lui. Il est d'accord avec tous les articles des journalistes et même avec les commentaires postés en réponse par d'autres lecteurs. C'est un sentiment bien agréable. Bien plus agréable en tout cas que de discuter avec Gérard au travail, qui passe son temps à chercher des conspirations américano-économico-sociales pour exploiter les gens. C'est sans doute ces sites pirates non référencés par Google que Gérard consultait qui lui ont tourné la tête. D'ailleurs, ça fait un certain temps qu'il ne l'a plus vu, ce collègue.

Pour se changer les idées, Gaspard va vite faire un tour sur Amazon, son site de vente en ligne préféré. Amazon, c'est tellement plus pratique que les quelques magasins physiques qui subsistent ! On sait qu'on ne va pas perdre de temps à regarder des produits qui ne nous intéressent pas, le tri est fait par des algorithmes qui connaissent nos goûts sur le bout des doigts ! Toutes les séries comiques américaines sont là, dès la première page. Ah tiens, en revanche, il ne trouve pas ce best-

seller mondial dont sa mère lui a tant parlé. C'est étrange, il croyait que ça faisait le buzz en ce moment. Tant pis.

Il est 7h30, c'est l'heure d'aller au travail. Gaspard vérifie que sa puce s'est bien rechargée cette nuit. Cette puce, il doit la porter en permanence depuis qu'un matin, Google lui a demandé de choisir entre ça et payer pour sa boite mail. Gaspard ne sait pas trop à quoi ça peut bien leur servir, mais bon après tout il n'a rien à cacher, non ? Il se dit d'ailleurs que ses parents sont bien idiots d'avoir choisi de payer. Vraiment pas 3.0, ceux-là…

Vous trouvez l'histoire de Gaspard caricaturale ? Vous avez raison, elle l'est.

Encore que… Qu'est-ce qui fait l'actualité en ce moment ? Les difficultés de l'administration Obama à justifier les moyens employés par ses services de sécurité. Ceux-ci ont accès non seulement aux relevés d'appels téléphoniques des citoyens, mais aussi apparemment à des portes dérobées dans les serveurs des géants américains de l'Internet. Il faut, nous dit-on, trouver un équilibre entre protection de la vie privée et lutte contre le (cyber)-terrorisme.

De l'autre côté de l'Atlantique, l'Europe travaille à un règlement sur la protection des données personnelles qui fait l'objet d'un lobbying américain sans précédent (plus de 4000 amendements déposés, record absolu) !

Le CEO de Google publie un livre dans lequel il explique que les nouvelles technologies sont le meilleur moyen d'asseoir la puissance américaine sur le reste du monde. Julien Assange lui répond depuis l'ambassade équatorienne que ces nouvelles technologies signent la mort de la vie privée et que même les Etats vertueux ne sauront résister à la tentation croissante de contrôler leurs citoyens.

Pendant ce temps, Facebook, Google, et plus largement une myriade d'acteurs de l'Internet proposent des services gratuits en contrepartie de données personnelles qu'ils stockent *ad vitam aeternam* dans leurs propres serveurs.

Les data brokers sillonnent Internet à la recherche de données monnayables en espèces sonnantes et trébuchantes. Des ad exchanges, véritables places de marché de données personnelles, s'installent.

Google lit vos mails, Facebook scrute vos amitiés. Et leurs algorithmes progressent si vite que l'anonymisation semble appartenir au passé.

Alors, l'histoire de Gaspard est-elle complètement invraisemblable ?

Données personnelles et données non personnelles ?

Les informations du monde « pré-internet » côtoient de nouvelles typologies de données. Parmi toutes ces données, lesquelles peut-on considérer comme ayant un caractère personnel ? L'article 2 de la loi Informatique et Libertés[1] apporte une première réponse :

> *« Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. »*

Et si toutes les données étaient progressivement en passe de devenir personnelles ?

---

[1]La loi n°78-17 relative à l'informatique, aux fichiers et aux libertés du 6 janvier 1978 est la loi française qui réglemente aujourd'hui l'exploitation des données à caractère personnel. Elle a

Donnée personnelle ou donnée sensible ?

Le concept de donnée personnelle renvoie au concept d'**information discriminante**, qui est une information permettant d'appliquer à un individu en particulier un traitement spécifique en fonction de ce qui le caractérise.

Dans cette approche, la loi Informatique et Libertés, article 8.1, liste certains types de données comme particulièrement sensibles :

> *« Il est interdit de collecter ou de traiter des données à caractère personnel qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle de celles-ci. »*

Cette approche par typologie des données atteint toutefois rapidement ses limites, le niveau de sensibilité d'une information étant en constante évolution, et dépendant fortement de la culture du pays, de l'époque, du progrès technique, etc.

Aux Etats-Unis par exemple, les statistiques sur les origines ethniques ou sur les pratiques religieuses sont socialement acceptées. Ces études sont au contraire illégales en France, ce qui s'explique en partie par l'histoire de notre pays.

De même, les données génétiques et les données biométriques ont intégré en 2012 le cercle des données sensibles dans le projet de règlement européen, alors qu'elles n'étaient pas dans la directive 95/46/CE du 24 octobre 1995.

Même s'il est important, ne serait-ce que symboliquement, de conserver une protection particulière sur ce sous-ensemble des données sensibles, la définition exacte de ce sous-ensemble est tout sauf consensuelle.

La question des données sensibles montre que la notion de degré de sensibilité d'une donnée est complexe, et que l'on ne peut pas définir les données à caractère personnel par ce seul moyen.

Le mythe de l'anonymisation

Si on ne peut pas classer les données par degré de sensibilité, peut-être peut-on identifier une typologie par degré de personnalisation. La plupart des lois en matière de protection des données personnelles mettent ainsi l'accent sur les données identifiantes ou identifiables.

Ainsi, la directive 95/46/CE considère[2]

> *« que les principes de la protection doivent s'appliquer à toute information concernant une personne identifiée ou identifiable; que, pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en œuvre, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne; que les principes de la protection ne s'appliquent pas aux données rendues anonymes d'une manière telle que la personne concernée n'est plus identifiable;*

Ce qui apparaît, présenté sous cette forme théorique, comme du bon sens s'avère extrêmement complexe dans la réalité : des données a priori anonymes peuvent finalement révéler un "caractère personnel". Les lobbyistes cherchant à réduire l'impact du projet de règlement européen l'ont d'ailleurs bien compris[3]. Ainsi, des amendements qu'ils proposent introduisent par exemple la notion de « donnée pseudonymisée », une sorte de donnée à mi-chemin entre la donnée personnelle et la donnée anonyme, qui sortirait alors de fait du champ d'application des protections prévues par le règlement.

Une des premières idées des législateurs pour protéger les informations personnelles et les libertés individuelles des citoyens a été de recourir à des techniques d'anonymisation. En nettoyant les bases de données de certains champs identifiants, il serait ainsi possible de garantir l'impossibilité d'identifier précisément les individus dont ces bases sont l'objet. Il existerait donc deux types bien distincts de bases de données : d'un côté des bases de données personnelles et de l'autre des bases de données anonymes.

---

[2]Considérant 26 de la directive 95/46/CE du Parlement européen et du Conseil

[3]Un lobbying particulièrement intense, mené par les banques, assurances et géants de l'Internet, a lieu à Bruxelles autour du projet de règlement européen, lobbying que la Commissaire Européenne Viviane Reding a qualifié de « *féroce, absolument féroce* ».

Pourtant, que nous apprend l'histoire à ce sujet ? Que lorsque des bases de données censées être anonymisées ont été rendues publiques, il a souvent été possible de casser l'anonymat, et ceci même si l'anonymisation avait été garantie par des chercheurs spécialistes du sujet.

Au milieu des années 1990, une agence gouvernementale du Massachusetts, le Group Insurance Commission, publie les dossiers médicaux des fonctionnaires de l'Etat ainsi que de leurs familles. Comme les champs explicitement identifiants avaient été supprimés (nom, adresse, numéro de sécurité sociale etc), le gouverneur William Weld pu assurer à ses administrés que leur anonymat serait préservé. Pour seulement 20 dollars, Latanya Sweeney, doctorante en informatique, se procura alors les listes électorales de l'état du Massachusetts dans le but de croiser ces informations avec celles contenues dans les dossiers médicaux. Elle parvint ainsi à envoyer au gouverneur son propre dossier médical. La combinaison de la date de naissance, du sexe et du code postal était en effet suffisante pour identifier le gouverneur de façon unique.

La plupart des techniques de désanonymisation fonctionnent en croisant la base à priori anonymisée avec suffisamment d'autres données disponibles. Et à mesure qu'Internet grandit, le volume de données disponibles explose. Toutes les 5 minutes, 13 millions de recherches sont effectuées sur Google[4]. 500 millions de mails sont envoyés[5]. Et 2 500 nouveaux utilisateurs se connectent à Internet pour la première fois[6].

La définition actuelle des données à caractère personnel serait suffisamment « plastique » car elle comporte la notion de donnée directement ou indirectement identifiante. Mais comment le régulateur peut raisonnablement suivre le rythme du progrès technique de ré-identification ?

« Nous sommes absolument dépassés par les applications technologiques, tous les jours il en arrive de nouvelles sur notre bureau. »[7]

Dans une époque marquée par une capacité de calcul toujours croissante, par une multiplication des données collectées et mises en relation, il n'apparaît plus approprié de chercher à classifier les données en fonction de leur « personnalisation ».

[4]Valable pour décembre 2012. Source : comScore qSearch.
[5]Valable pour l'année 2012. Source : The Radicati Group, *Email Statistics Report, 2012-2016*.
[6]Valable du 01/01/12 au 31 juin 2012. Source : Internet World Stats.
[7]Alex Türk, président de la CNIL de 2004 à 2011

La différence entre nos identités numériques et classiques

On pourrait, à ce stade, être tenté d'arguer qu'il restera toujours des situations dans lesquelles nos données ne pourront jamais être désanonymisées. N'est-ce d'ailleurs pas à cela que sert la fonctionnalité de navigation privée sur Internet ?

L'exemple de l'adresse IP est à ce titre intéressant. L'adresse IP (*Internet Protocol*) est un numéro, unique, attribué de façon provisoire ou permanente, sous lequel un appareil est connecté à Internet. Ce numéro étant unique, il permet indiscutablement d'identifier un appareil connecté au réseau. Pourtant, se pose encore aujourd'hui la question de savoir si l'IP doit être considérée comme une donnée à caractère personnel ou non. La CNIL, suivie par la moitié des tribunaux français, considère que l'IP est une donnée personnelle[8]. Mais ceci n'est pas le cas de la Cour d'appel de Paris :

> *« L'adresse IP ne permet pas d'identifier le ou les personnes qui ont utilisé cet ordinateur, puisque seule l'autorité légitime pour poursuivre l'enquête (police ou gendarmerie) peut obtenir du fournisseur l'accès d'identité de l'utilisateur. »*[9].

Une solution définitive sera peut-être apportée par le projet de règlement européen, mais on voit bien que la difficulté réside dans le fait que l'adresse IP identifie un appareil et non une personne. Comme nous le verrons au travers de nombreux exemples dans la suite de ce document, il suffit bien souvent de pouvoir identifier un appareil pour être en mesure d'appliquer un traitement discriminant à celui qui l'utilise.

Il n'est pas nécessaire de savoir si c'est M. X ou Mme Y qui est derrière l'ordinateur, du moment qu'il est possible de recomposer une « identité virtuelle » de l'utilisateur.

---

[8]Communiqué CNIL, 2 août 2007
[9]Cour d'appel de Paris, 13ème chambre, section B. Arrêt du 27 avril 2007

La vie privée, et puis c'est tout ?

Une grande partie du débat sur les données personnelles porte précisément sur la définition juridique du caractère personnel d'une donnée.

Cette discussion est évidemment utile en matière de protection de la vie privée, et doit être effectuée à la lumière des difficultés que posent notamment les avancées technologiques, ce que Paul Ohm résume par « *La facile ré-identification représente un changement radical non seulement dans la technologie mais également dans notre compréhension de la vie privée* ». Les frontières juridiques qui délimitent les données à caractère personnel sont donc amenées à bouger.

Le terme même de « donnée personnelle » appelle à la vigilance lorsqu'il s'agit de vie privée. Cependant, il a tendance à restreindre la réflexion aux questions de vie privée « classique ». A ce titre, il est intéressant de remarquer que la question des données personnelles sur Internet se cristallise essentiellement autour des réseaux sociaux, Facebook en tête, car c'est là que l'on retrouve les données avec lesquels nous sommes les plus familiers.

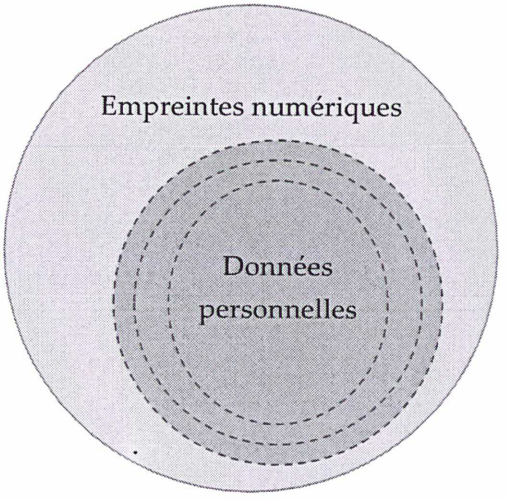


Figure 1 : Les empreintes numériques, incluant les données personnelles aux frontières fluctuantes

Cependant, nous voulons démontrer dans ce mémoire que les données personnelles ne sont pas seulement un enjeu de libertés individuelles, mais qu'il faut également prendre en compte les problématiques économiques sous-jacentes.

Il nous faudra élargir le débat en considérant les empreintes numériques, que nous définissons comme l'ensemble des données issues des utilisateurs sur Internet.

Comment sont produites les empreintes numériques, et pour quels usages ?

L'économie numérique dominée par le tout gratuit

Vous l'avez sûrement remarqué, notre sujet d'analyse a fait l'objet de nombreuses attentions ces derniers temps.

Les polémiques lancées par les révélations sur les écoutes à grande échelle menées par différents gouvernements ont ainsi fait beaucoup de bruit. Certains députés français mais aussi la commission européenne se sont penchés sur le sujet. L'une de ces initiatives nous intéresse particulièrement : il s'agit du rapport français sur la **fiscalité du numérique**[10], commandé par le gouvernement et publié en début d'année 2013. Ce rapport, surnommé « Colin – Collin » du nom des deux auteurs, évoque en effet la possibilité d'instaurer une taxe dont l'assiette serait la collecte des données numériques.

Le raisonnement des deux auteurs rejoint le nôtre et part d'un constat indéniable : sur Internet il y a un nombre incroyable de services gratuits. Très peu d'utilisateurs sont aujourd'hui prêts à payer pour une boite aux lettre électronique, et même les industries culturelles, bien qu'historiquement très réticentes vis-à-vis de ces modèles gratuits, développent aujourd'hui des services de « streaming[11] », voir même de téléchargement de musiques enregistrées, financés par la publicité.

Or il semble nécessaire que ces entreprises aient identifié une contrepartie valorisable pour que celles-ci voient une pertinence économique à développer des tels services gratuits.

---

[10] http://www.redressement-productif.gouv.fr/rapport-sur-fiscalite-secteur-numerique

[11] Le *streaming* désigne un mode de transmission de données en flux continu, dès que l'internaute sollicite le fichier.

Nos données sont-elles la nouvelle monnaie d'échange de l'économie numérique ?

Ce raisonnement permet d'introduire la notion de **travail gratuit**[12] que l'on retrouve dans le rapport sur la fiscalité du numérique. Vous en avez déjà peut-être entendu parler, reformulé de manière quelque peu négative, sous la forme suivante :

> *« Si vous ne payez pas le service, c'est que vous êtes le produit »*

En effet, la proposition de valeur faite par ces entreprises de l'économie numérique est originale et particulière, dans le sens où celle-ci n'est jamais payée en flux monétaire. Ceci pose un certain nombre de problèmes, notamment fiscaux puisque des produits créés sur le territoire français ne sont jamais valorisés dans le PIB et échappent à toute assiette d'imposition. Il est alors intéressant de remarquer que l'assise de la taxe que les rapporteurs proposent est justement la collecte de données numériques !

Tout semble indiquer que ce sont nos données qui, sur Internet, jouent le rôle de monnaie d'échange.

Payons-nous notre boite e-mail avec le contenu de notre correspondance ?

Prenons, pour fixer les idées, un exemple concret - celui des boites aux lettres électroniques - et reformulons notre conclusion précédente : payerions-nous notre boite mail avec le contenu de notre correspondance ?

La gestion d'un service e-mail a en effet un coût strictement positif[13], et la compétition étant forte sur ce type de service, il est difficile d'imaginer que toutes ces entreprises (Google, Yahoo, Microsoft,

---

[12] Le terme « travail » n'est ici pas à comprendre au sens littéral : ce travail gratuit désigne la contrepartie troquée contre le service fourni gratuitement (regarder une publicité sur YouTube, donner son carnet d'adresse à LinkedIn ou encore laisser google analyser ses mails…)

[13] Ce coût prend la forme d'achat d'espace de stockage, de développement et de maintenance du logiciel, etc.

Orange, SFR, Free etc.) développeraient ce service purement bénévolement. Mais alors pourquoi développent-elles de tels services ?

Une première idée pourrait être qu'il faille replacer ce service dans un contexte et un marché plus global : de nombreux services après-ventes peuvent par exemple sembler gratuits lorsqu'on les considère seuls (ils ne sont généralement pas facturés à l'utilisation) alors que la réalité est que ces services, indispensables pour pouvoir vendre un produit, sont facturés aux consommateurs sous la forme d'abonnements inclus dans le prix du produit initial. Ceci pourrait expliquer l'intérêt que trouvent les fournisseurs d'accès à Internet à développer des services mails. Pourtant, dans le cas de Google, nous utilisons des boites Gmail depuis plusieurs années sans avoir jamais payé le moindre centime à Google, que ce soit pour ce service ou pour n'importe quel autre.

Ainsi, dans certains cas tout du moins, il est nécessaire que ces services génèrent une valeur réelle strictement positive pour les entreprises qui les développent. Dans le cas de Gmail, le flux d'empreintes numériques généré par le service est transformé par Google en valeur monétaire grâce à la publicité.

> Le contenu de notre correspondance a une valeur telle qu'elle nous permet à elle seule de payer notre boite email.

N.B. Il semble à ce stade nécessaire de clarifier notre position : les formulations que nous avons employées jusqu'ici peuvent sembler quelque peu critiques et dénonciatrices vis-à-vis de ce système. Il n'en est rien et ces formules n'ont été utilisées que dans le but d'éveiller votre intérêt de lecteur ! Nous pensons en effet, et nous vous le présenterons dans la suite de ce document, que ce système est fondamentalement créateur de valeur pour l'ensemble des participants. Il existe d'autre part des boites e-mails payantes, où le contenu de la correspondance n'est jamais monétisé. Ces services sont aujourd'hui essentiellement utilisés par les entreprises, mais le développement de modèles gratuits pour l'utilisateur permet, dans notre opinion, un formidable développement de ces nouveaux usages.

Les modèles d'affaires multi-faces

Les modèles économiques financés par la publicité, dont nous venons de vous donner un exemple, peuvent être décrits et analysés dans le cadre plus général des modèles d'affaires dits « multifaces ». De manière très schématique, il faut, pour stabiliser ce modèle où une entreprise fournit un service qui lui coûte à un utilisateur qui ne paye pas, introduire un troisième acteur pour jouer le rôle de

client. On parle alors de marché biface car il existe deux clientèles tout à fait différentes, quoique finalement interdépendantes.

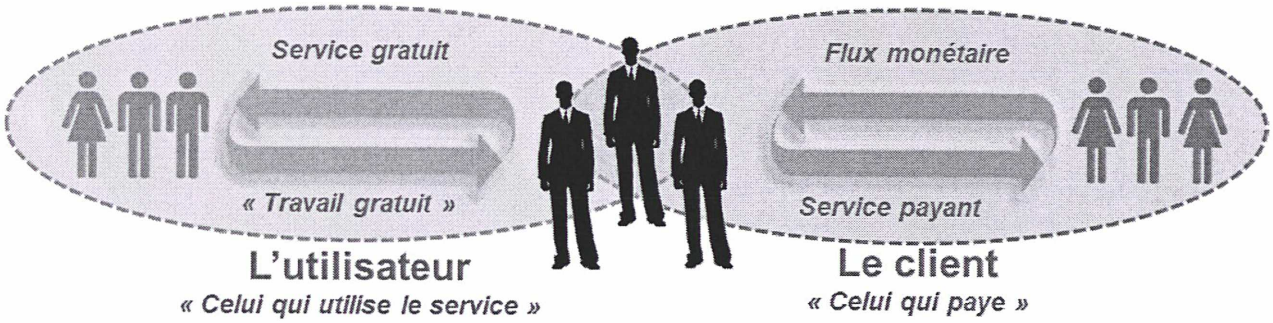


Figure 2 - Modèle d'affaire biface, schéma simplifié

Ces modèles ont notamment une caractéristique essentielle[14], celle de présenter des « effets de réseau croisés ». Formulé dans un langage d'économiste, cela veut dire que l'utilité qu'un agent d'un côté du marché retire de sa participation au service dépend du nombre de participants de l'autre côté du marché. Dans notre exemple de boite e-mail, plus il y a d'utilisateurs de la messagerie, plus des entreprises seront intéressés à investir dans de la publicité ciblée sur cette plateforme, et donc plus le service gagnera en fonctionnalités, attirant de nouveaux utilisateurs. L'existence même de ces externalités croisées pose d'importants problèmes concurrentiels, mais ceux-ci sortent quelque peu du cadre de notre analyse.[15]

Et mon profil Facebook, combien vaut-il ?

Vous n'êtes pas encore convaincus que vos données sont une monnaie d'échange sur Internet ? C'est bien dommage car les investisseurs, eux, semblent déjà l'être...

Penchons-nous désormais sur les dernières entrées en bourse d'entreprises de l'économie numérique, ou encore sur les acquisitions faites ces dernières années par les grands consolidateurs de marché que sont Google, Apple ou Facebook. Il apparaît clairement que toutes les valorisations

---

[14] Weyl, E Glen. 2010, *"A Price Theory of Multi-sided Platforms."* American Economic Review

[15] Des exemples de telles problématiques sont l'existence de monopole naturels, ou encore l'impact sur la neutralité du traitement de l'information (pour plus d'information à ce sujet, voir l'article de Nathalie Sonnac "*Médias et publicité ou les conséquences d'une interaction entre deux marchés*", Le Temps des médias n°6, printemps 2006, p.49-58).

ne se font pas sur les mêmes critères. Lorsque ces consolidateurs ne s'intéressent qu'à la technologie de l'entreprise ciblée, les prix d'achats d'entreprises semblent découler d'un rationnel financier classique de type « *make or buy* » (valorisation sur des montants équivalents à l'investissement nécessaire pour tout reconstruire) ou encore de type chiffre d'affaire supplémentaire généré, et les montants restent raisonnables aux yeux des observateurs extérieurs[16]. Au contraire, lorsque l'entreprise touche directement le grand public (réseaux sociaux par exemple), les montants s'envolent vers des niveaux qui suscitent de vives réactions et semblent incompréhensibles pour une partie des observateurs[17].

Prenons l'exemple de l'entreprise Facebook Inc. A l'heure où nous écrivons ces lignes, l'entreprise est valorisée à plus de $45 milliards[18] : comment les investisseurs arrivent-ils à de tels calculs ?

Nous ne prétendons pas nous pencher ici sur les méthodes de valorisations complexes réellement mises en œuvre mais il est bien souvent possible de les approximer grossièrement à l'aide de raisonnements simples. L'un d'entre eux consiste par exemple à valoriser une entreprise en fonction de résultats d'exploitation projetés sur les années futures ce qui permet d'évaluer le rendement du placement : si l'investisseur espère un rendement de 10% par an, il valorisera l'entreprise à hauteur de $^1/_{10\%}$ =10 fois son résultat d'exploitation espéré pour l'année suivante[19].

Suivant ce raisonnement, les investisseurs comparent ainsi le ratio Valorisation boursière / Résultat d'exploitation à celui d'autres entreprises similaires : plus ce ratio est petit, plus le rendement de l'investissement est important. Dans le cas de Facebook, le résultat d'exploitation en 2012 était de $538 millions, ce qui nous donne un ratio de 75, bien loin de la moyenne de 15 observée sur le S&P500. Pour être dans la moyenne du marché, le résultat d'exploitation de Facebook devrait être autour de $3 milliards...

---

[16] En 2013 par exemple : WiFiSlam acheté $20 millions par Apple pour sa technologie de géolocalisation en intérieur ou Wavii acheté $30 millions par Google pour sa technologie de Traitement automatique du langage.

[17] L'exemple d'Instagram, dont le rachat en 2012 par Facebook pour $1 milliard a fait couler beaucoup d'encre, est à ce titre assez caractéristique.

[18] Valorisation bien loin de sa valeur d'entrée de $68 milliards il y a un peu plus d'un an.

[19] Le raisonnement est ici volontairement extrêmement simplifié, mais en achetant x % de l'entreprise à ce prix, l'investisseur récupèrera un an plus tard x % du résultat d'exploitation soit très exactement 10 % de son investissement initial.

Il faut certes nuancer ce résultat en tenant en compte deux facteurs :

- Tout d'abord, l'investissement dans ce type de start-up high-tech est par nature risqué et imprévisible et le système de valorisation repose sur quelques rares mais très juteux succès ; ce qui a pour conséquence directe une tendance à la survalorisation du secteur dans son ensemble.

- D'autre part, le résultat d'exploitation de Facebook est en forte croissance : si l'on raisonne sur cinq ans par exemple, anticiper une croissance de 30% par an revient à un résultat d'exploitation moyen sur la période de 2,35 fois l'actuel, ce qui ferait dans le cas de Facebook passer le ratio de 75 à 32. Il est intéressant toutefois de remarquer qu'il faut attendre une croissance de plus de 50% par an sur les cinq prochaines années pour que notre raisonnement arrive à un ratio dans la moyenne de marché.

Tout ceci nous permet de douter très sérieusement du fait que Facebook Inc. soit valorisé en fonction de ses performances financières actuelles par les investisseurs du monde entier. Mais un autre raisonnement permet d'éclairer cette valorisation de manière beaucoup plus satisfaisante.

Nous avons vu que la valorisation actuelle de Facebook correspondait à un résultat d'exploitation annuel espéré par les investisseurs de $3 milliards. A la lumière du nombre d'utilisateurs actifs de la plateforme (plus de 1,1 milliards chaque mois), ce chiffre peut être reformulé sous la forme de moins de 3$ par an et par utilisateur actif. Une valeur qui semble bien plus raisonnable, surtout lorsque l'on sait que Google génère de l'ordre de 20$ de chiffre d'affaire annuel par utilisateur actif[20]...

En moyenne les investisseurs du monde entier pensent pouvoir facturer la location de l'accès à votre profil Facebook pour 3$ par an !

Ce n'est certes pas beaucoup mais vous pouvez toujours vous rattraper avec votre profil LinkedIn (valorisé à $20 milliards pour 200 millions de membres, ce qui monte la valorisation d'un profil à 7$ par an).

[20] http://www.webanalyticsworld.net/2011/01/revenue-per-unique-user-amazon-google.html

De la collecte à l'identification

Des empreintes souvent issues d'une contribution bénévole

La collecte des empreintes numériques est ainsi devenue la norme sur Internet, et leur exploitation est généralement régie par une clause des Conditions générales d'utilisation (CGU) du site.

Ces conditions, qui doivent être accessibles depuis n'importe quelle page du site, sont considérées acceptées par chaque utilisateur du site. Par exemple, les CGU de Google[21] informent l'utilisateur quelles empreintes vont être collectées lors de l'utilisation de ses services (Search, Gmail, YouTube etc.)

> *« Lorsque vous utilisez nos services ou que vous affichez des contenus fournis par Google, nous pouvons automatiquement collecter et stocker des informations dans les fichiers journaux de nos serveurs. Cela peut inclure [...] vos requêtes de recherche [...] votre numéro de téléphone [...] votre adresse IP. »*

Est-il pour autant raisonnable de penser que la majorité des internautes sont réellement conscients de ces conditions ? Certes, nul n'est censé ignorer la loi, mais une étude[22] suggère tout de même qu'un internaute moyen aurait besoin de 250 heures pour lire les CGU de tous les services qu'il utilise. La Cour d'appel de Pau est d'ailleurs allée dans ce sens dans un arrêt[23] concernant Facebook :

> *« La clause [...] est noyée dans de très nombreuses dispositions [...] en petits caractères [...] arrive au terme d'une lecture complexe de douze pages format A4 pour la version papier [...] et la clause attributive de compétence doit être réputée non écrite. »*

---

[21]http://www.google.com/intl/fr/policies/privacy/
[22]McDonald, A. et Cranor, L. « The Cost of Reading Privacy Policies. » *I/S: A Journal of Law and Policy for the Information Society*. 2008.
[23]Cour d'appel de Pau, 1ère chambre. Arrêt du 23 mars 2012

Ceci est d'autant plus difficile pour les utilisateurs à suivre que les CGU ainsi que les paramètres par défaut évoluent très fréquemment, comme le montre l'évolution des paramètres de Facebook.

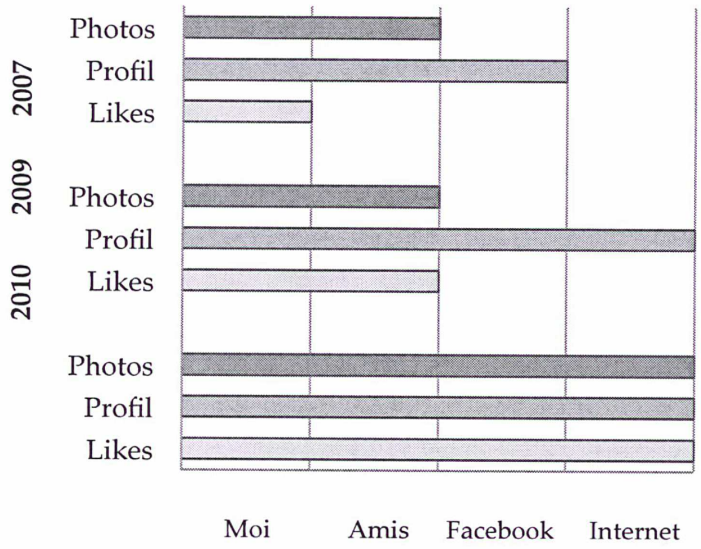


Figure 3 : Evolution des paramètres de confidentialité sur Facebook

Dans le but de pallier ce manque de transparence, le régime applicable aux cookies a été modifié par le « Paquet Télécom »[24]. Auparavant, un site qui désirait insérer un cookie en informait ses utilisateurs dans les CGU. Désormais, sauf certains cas particuliers, l'internaute doit exprimer son consentement préalable, explicite et spécifique (« *opt-in* ») pour que le site puisse insérer un cookie sur son appareil. Il faut cependant reconnaître que cette obligation n'est pour l'instant quasiment pas respectée en France, la CNIL admettant que « *la majorité des sites sont dans l'illégalité* »…

« Oui, j'ai lu les conditions générales d'utilisations » est peut-être le plus grand mensonge d'Internet.

[24] Directive n°2009/136/CE du 25 novembre 2009, transposée en droit français par l'ordonnance n° 2011-1012 du 24 août 2011.

Les différents modes de collecte

Les données soumises volontairement par l'utilisateur

Pour rendre le service proposé à l'utilisateur, une application Internet peut lui demander une saisie d'informations (par exemple un formulaire d'inscription, ou bien une requête sur un moteur de recherche). Ainsi, un utilisateur moyen de Google produit 98 requêtes par mois[25], et un internaute français possède un compte sur 15 sites en ligne en moyenne :

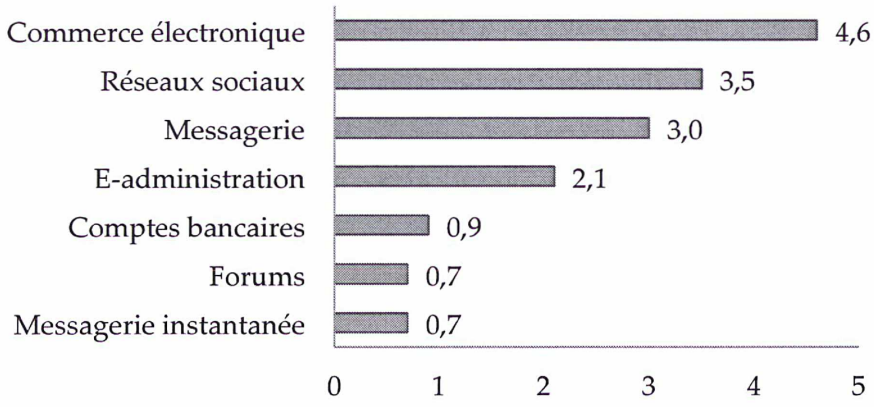


Figure 4 : Nombre de comptes en ligne pour un internaute français "moyen"[26]

L'évolution vers le web 2.0 a été marquée par une augmentation significative de la contribution des utilisateurs. L'internaute est en effet devenu un acteur actif du web : il interagit, partage et échange. Les réseaux sociaux se construisent autour des données soumises volontairement par les membres : publications, photographies, relations, etc. Ainsi, un internaute actif sur Facebook publie en moyenne 123 *likes*[27] et 7 photos[28] sur le réseau social.

---

[25]Source : comScore qSearch, décembre 2012

[26]Source : *Baromètre de la confiance des Français dans le numérique*, ACSEL et Caisse des Dépôts, octobre 2011, ainsi que *Observatoire des réseaux sociaux*, IFOP, novembre 2012.

[27]Un *Like* permet à l'utilisateur de Facebook de montrer qu'il aime un contenu. Source : *Facebook's Growth In The Past Year*, mai 2013.

[28]Source : *Facebook Annual Report* pour 2012.

Les données captées indirectement

Certaines données captées sont produites par l'utilisateur et collectées lors de l'utilisation d'une application, sans pour autant être nécessaires au fonctionnement direct de celle-ci.

Ainsi, suivant les utilisations, tous les types d'informations disponibles peuvent être collectés : historiques de navigation, adresses IP, clics, conversations, cookies, données de géolocalisations, etc.

L'extension Collusion[29] permet de rendre visibles les sites tiers qui captent ces données. Au fur et à mesure de la navigation, Collusion représente sur un graphe les sites visités par l'utilisateur ainsi que les sites tiers qui sont également impliqués dans la collecte de données en étant connectés aux sites visités.

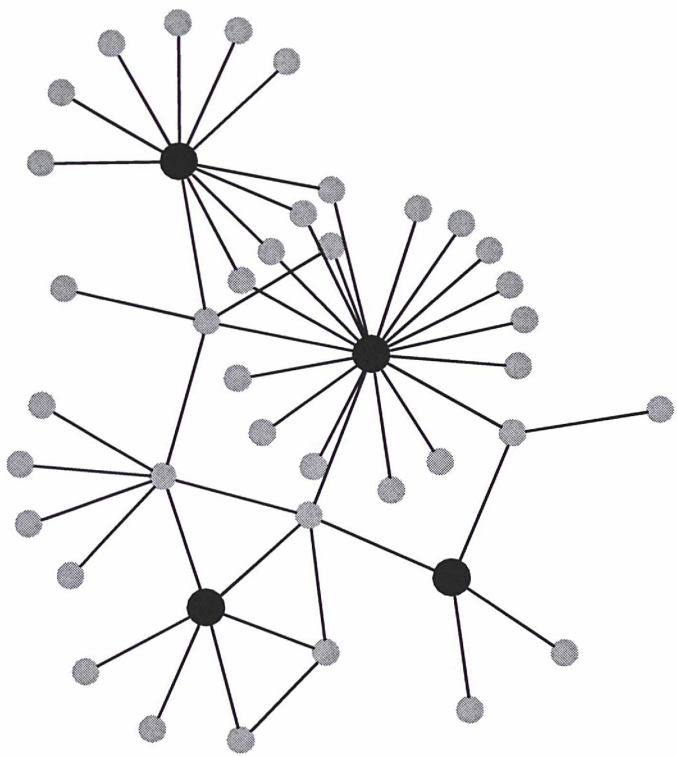


Figure 5 : Graphe obtenu par Collusion après avoir visité leMonde.fr, lesEchos.fr, Amazon.fr et eBay.fr. Les points noirs sont les 4 sites visités, les points gris sont les sites tiers.

[29]Add-on expérimental de la fondation Mozilla, disponible sur le navigateur Firefox.

Les données déduites par recoupement entre différentes bases

Enfin, certaines données sont déduites par recoupement entre différentes bases de données, qui peuvent être publiquement disponibles sur Internet ou propriétaires. Par exemple, 123people crée automatiquement des pages de profils en agrégeant les informations disponibles sur les individus : nom, biographie, photographies, informations issues des réseaux sociaux, actualités etc.

Nos données sont mortes, seuls les traitements ont de la valeur

Nos empreintes personnelles forment une nouvelle monnaie d'échange dans l'économie numérique. Mais, comme toute autre monnaie, pour que ces empreintes soient valorisées par les entreprises, encore faut-il pouvoir les transformer ou les échanger en monnaie classique. Pour cela, pas de secret : pour qu'une empreinte ait la moindre valeur, il faut nécessairement la lier à une personne physique identifiée, ou tout au moins à un appareil connecté.

> Savoir qu'un article sur un site de e-commerce a été vu 100 fois aujourd'hui est nettement moins intéressant que de savoir que, sur ces 100 fois, 6 utilisateurs ont regardé le produit plus de 10 fois.

Nous nous intéressons dans cette partie à la mise en œuvre technique de ces méthodes de « *tracking*[30] » des utilisateurs. Ces méthodes peuvent être classifiées en deux grands types :

- le **first-party tracking** où l'identification est effectuée directement par le site ou l'application avec lequel l'utilisateur interagit ;
- le **third-party tracking** où cette identification est effectuée par une entité extérieure[31].

Ce second type de méthode permet un suivi de l'utilisateur dans la durée et sur de multiples sites.

---

[30]Nous appelons dans la suite méthode de « tracking » toute méthode permettant de rattacher une action sur internet à un appareil connecté.

[31]Le terme « third party » fait référence au fait que l'utilisateur est le « second-party ».

De multiples méthodes d'identification

Les cookies

Les cookies sont des fichiers de texte qu'un site peut stocker à l'intérieur de votre navigateur, dans le but de le redemander plus tard. Le cas typique d'utilisation est lorsqu'un site stocke votre identifiant de compte pour vous permettre une connexion automatique à chaque visite. Ces fichiers ont obligatoirement une date d'expiration, mais en pratique celle-ci peut être fixée à plusieurs dizaines d'années…

Le tracking par cookie est très utilisé et les développeurs de navigateurs ont implémenté une certaine sécurité sur ce type de pratiques : tout d'abord parce que ces fichiers sont assez facilement accessible à l'utilisateur qui peut les effacer à tout moment. Ensuite au niveau de l'URL, un domaine comme *.lemonde.fr ne pouvant avoir accès qu'aux cookies ayant été créées à partir de ce même domaine.

Toutefois de nombreuses pages web incluent des images, des iFrames… qui font référence à des noms de domaines externes. Ce sont ces références qui permettent la mise en place de « *third-party cookies* » permettant de suivre l'utilisateur sur plusieurs sites et domaines différents.

JavaScript

De nombreux sites utilisent aussi des fichiers JavaScript : ces fichiers contiennent des programmes que l'utilisateur exécute directement dans son navigateur. Ces programmes, présents aujourd'hui dans la quasi-totalité des sites, permettent notamment le développement d'applications très interactives (le programme s'exécutant directement sur l'ordinateur de l'utilisateur, les latences de chargement sont significativement réduites). Les programmes JavaScript peuvent envoyer des informations à leur serveur d'origine, mais les politiques de contrôle des navigateurs ne leur ouvrent qu'un accès assez limité aux données de l'utilisateur.

Fingerprinting

De manière à identifier un appareil connecté, il est possible d'exploiter toutes les informations transmises par les navigateurs et présentes dans les requêtes HTTP (adresse IP, user-agent string, polices installées, liens referer, préférences linguistiques, taille de l'écran etc.). Ces informations sont

utilisées avant tout pour simplifier la navigation, mais une étude récente[32] a montré que la combinaison de toutes ces informations permet d'identifier quasi-uniquement un appareil. L'étude empirique montre que ces méthodes permettent d'identifier un ordinateur unique parmi 290 000. L'ajout du préfixe IP, interdit en France, permet d'augmenter très significativement la précision de cette identification (la probabilité devient alors de 95%).

Le bouton like de Facebook

Le bouton like de Facebook est un bouton présent sur de très nombreux sites autres que Facebook. Mis en place par le site partenaire lui-même, il permet aux visiteurs, d'un simple clic, de partager la page visitée à leurs réseaux.

Ce bouton n'est pas une simple image, mais le résultat de l'exécution d'un programme JavaScript. Lorsque vous chargez une page contenant un bouton like, Facebook est averti de votre visite et vous envoie l'image du bouton correspondant. Ceci permet notamment d'ajouter des informations comme le nombre de personnes ayant « liké » la page, voir même des photos de vos amis si certains d'entre eux ont déjà partagé cette page.

Deux cas de figures se présentent : si vous êtes connectés à votre compte Facebook (ce que Facebook sait en demandant votre cookie à votre navigateur), Facebook sait immédiatement que vous, utilisateur n° XXX, venez de visiter la page concernée. Sinon Facebook peut stocker l'information de votre visite pour vous identifier dans le futur (lorsque vous vous connecterez par exemple).

[32] Peter Eckersley « How Unique Is Your Web Browser? » et http://www.panopticlick.eff.org

Les utilisations et monétisations actuelles

Quelles informations sont récoltées ?

Données légales

La loi oblige certains sites Internet à collecter et conserver un certain nombre d'empreintes numériques sur ses propres clients. Sur décision d'un juge, la puissance publique peut ensuite demander l'accès à ces informations dans le but de détecter des fraudes ou d'identifier des utilisateurs (pour la sécurité nationale et lutte contre le terrorisme par exemple).

Ergonomie du parcours client

Un certain nombre d'empreintes sont également collectées par les sites dans le but de personnaliser l'expérience des utilisateurs (articles déjà lus, maintien du caddie d'achat entre différentes sessions etc.).

Données analytiques

Par données analytiques, nous désignons tous les types de mesures agrégées, comme par exemple l'analyse de trafic. Bien qu'il soit techniquement possible pour un éditeur de site d'effectuer ces mesures par lui-même (ou d'utiliser des logiciels open-sources permettant de conserver ses propres données), beaucoup d'éditeurs utilisent des logiciels extérieurs, comme Google Analytics, plus simple à mettre en œuvre. Ces méthodes de *third party tracking* permettent une valorisation complète de ces données analytiques par l'éditeur du logiciel d'analyse.

Profilage marketing

Comme Facebook, la plupart des réseaux sociaux ont créé des boutons, similaires au bouton *Like*, qui leur permettent de suivre les utilisateurs sur la majorité de leur parcours sur Internet. Ces derniers ne communiquent pas sur les informations qu'ils récoltent, mais rien n'empêche techniquement d'obtenir l'information de visite d'un utilisateur, même lorsque celui-ci ne clique pas sur le bouton et même lorsque l'utilisateur n'est pas connecté ou n'a pas de compte sur le réseau social.

Tracking des appareils mobiles

Les appareils mobiles comme les smartphones disposent aujourd'hui de réelles capacités de calcul et sont équipés de nombreux capteurs (micro, caméra, GPS, accéléromètre etc). Ils regorgent d'informations personnelles comme le numéro de téléphone, la position géographique, le carnet d'adresses etc. De nombreuses polémiques concernant ces données ont explosé ces dernières années, comme par exemple lorsque certaines applications (Facebook mobile, entre autres) envoyaient à leurs propres serveurs l'intégralité du carnet d'adresses du téléphone.

Piste d'évolution future : « Reality et Physical Mining »

Le *physical mining* désigne les techniques d'exploration de données (*data mining*) appliquées aux données du monde physiques (celles issues des capteurs de nos smartphones par exemple). Le but de ces techniques est d'analyser les relations humaines et sociales à partir des données de localisation des utilisateurs, du journal et des temps d'appels, des données vidéos etc.

Il ne s'agit ici pas de science-fiction : une étude d'Acquisti sur la convergence de la reconnaissance faciale, des réseaux sociaux et du *data mining* montre ainsi qu'il est possible d'utiliser ces techniques de *physical mining* pour extraire des données très sensibles sur une personne (identifier quelqu'un dans la rue, désanonymiser les pseudonymes des sites de rencontre grâce à une photo de profil…).

Le marché de la publicité en temps réel

Dans la publicité traditionnelle (télé, radio, presse etc), les contrats de publicité se négocient de gré à gré entre les acheteurs (annonceurs ou agences) et les éditeurs.

L'explosion du nombre de sites web a engendré un accroissement important de l'inventaire publicitaire disponible. Les éditeurs ont alors cherché à revendre l'inventaire qu'ils n'arrivaient pas à vendre directement aux annonceurs. Après plusieurs tentatives infructueuses, c'est le RTB (pour *Real-Time Bidding*, ou enchères en temps réel) qui s'est imposé.

Ce concept, popularisé par Google à travers son service *AdWords* d'achats de mots-clés par enchère, consiste à proposer chaque impression publicitaire à des annonceurs à travers une vente aux enchères en temps réel (au plus une centaine de millisecondes pour afficher la publicité). Cette technique s'est aujourd'hui beaucoup popularisée, et est appelée à prendre une part grandissante dans la publicité en ligne.

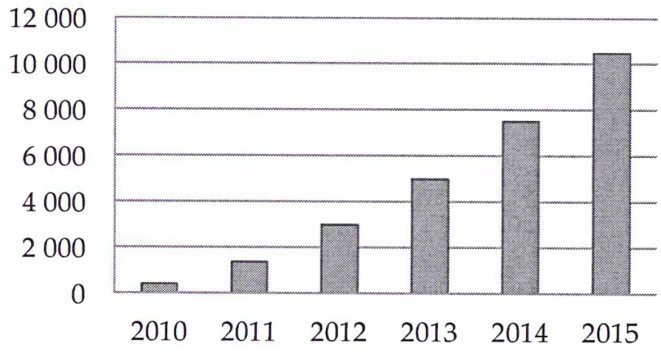


Figure 6 : Dépenses (milliards de $) publicitaires via le RTB[33]

Les acteurs

Comme sur les marchés financiers, trois acteurs principaux sont impliqués dans le RTB :

- ***Ad Exchange***[34] : place de marché automatisée qui permet l'achat et la vente d'espaces publicitaires en temps réels. L'*Ad Exchange* est financé par une commission sur chaque transaction.
- ***DemandSide Platform***[35] **(DSP)** : plateforme utilisée pour acheter les espaces publicitaires sur les *Ad Exchanges*. Ainsi en France par exemple, les agences medias ont des équipes (les *trading desks* – Vivaki pour Publicis, Affiperf pour Havas) qui pilotent l'achat de publicité en ligne pour leurs annonceurs.
- ***SupplySide Platform***[36] **(SSP)** : plateforme utilisée par les éditeurs pour commercialiser leurs espaces publicitaires invendus sur les *Ad Exchanges*.

Autours de ces trois types principaux gravite un nombre incroyable d'acteurs, chacun tentant d'attaquer ce marché en apportant sa propre innovation technologique. Il s'agit par exemple des *AdServer*[37], permettant l'hébergement et la distribution de la publicité en ligne, ou encore des *Data Management Platform* (DMP), vendant des données agrégées.

---

[33]Source : IDC, « *Real-Time Bidding in the United States and Worldwide* », 2012.
[34]*Ad Exchange* importants : Yahoo Right Media, DoubleClick.
[35]DSP importants : Appnexus, MediaMath, Turn.
[36]SSP importants : Pubmatic, Rubicon, Improve Digital.
[37]*AdServers* importants : Smart Adserver, 24/7 Real Media.

Le processus du RTB

Il est à ce stade intéressant de décrire le processus du RTB, tant il permet de réaliser quels sont les chemins empruntés par nos empreintes numériques. L'affichage d'une publicité ciblée sur Internet peut ainsi être décomposé en 4 étapes : la demande d'affichage, la vente aux enchères, l'affichage et finalement le *cookie matching*.

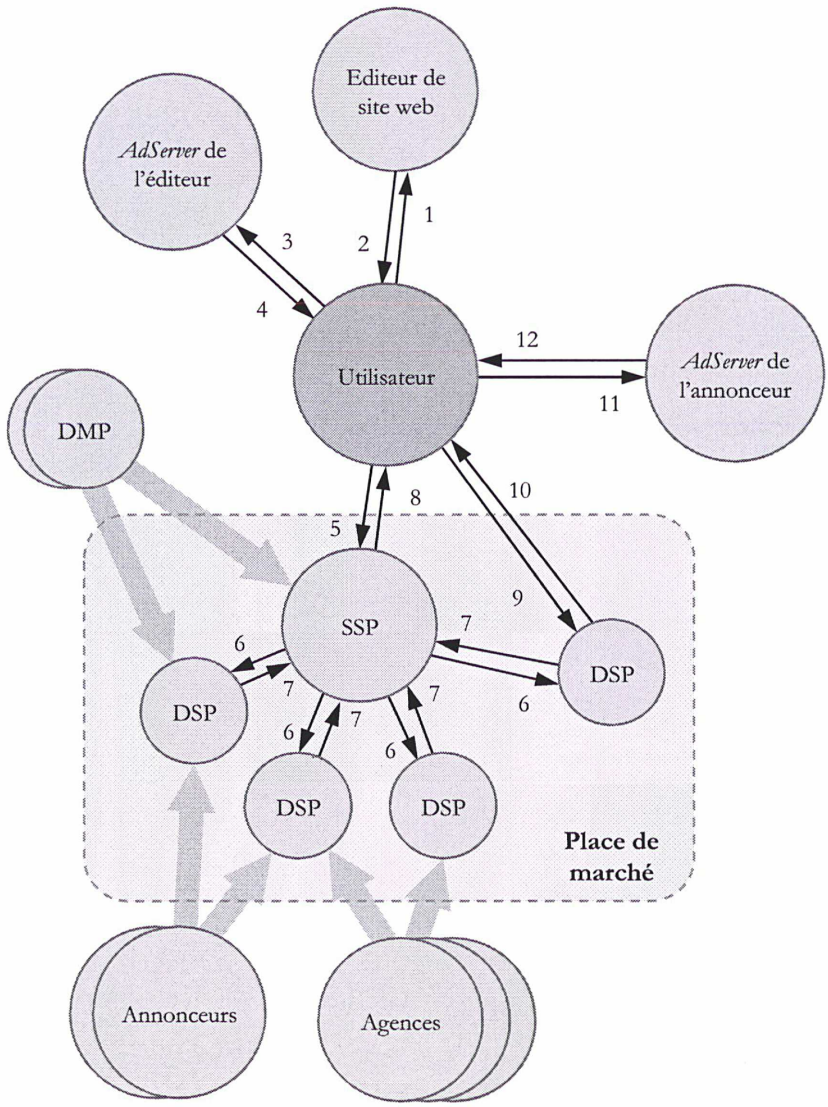


Figure 7 - Place de marché RTB

A. La demande d'affichage de publicité

1. L'utilisateur demande une page web via son navigateur.
2. Le serveur de l'éditeur de la page renvoie le contenu de la page demandé sous forme de code HTML. Au niveau de l'encart publicitaire, l'éditeur place un JavaScript qui indique au navigateur les coordonnées de son *AdServer*, pour connaître le contenu à insérer.
3. Le navigateur interroge l'*AdServer* de l'éditeur.
4. L'*AdServer* répond avec un « *ad tag* » qui contient les coordonnées du SSP à appeler.
5. Le navigateur appelle le serveur SSP.
   Si il existe, le SSP identifie l'utilisateur en lisant le cookie (nommé $cookie_{SSP}$) qu'il a pu placer auparavant.

B. La vente aux enchères sur la place de marché

6. Le SSP lance les enchères via un *call* (une offre) à tous les DSPs avec qui il est en relation, en envoyant le $cookie_{SSP}$ ainsi que les caractéristiques de l'encart (URL, format, etc.).
7. Les DSP cherchent à reconnaître l'utilisateur.
   - Si c'est la première fois que le SSP lui envoie un *call* pour $cookie_{SSP}$, le DSP n'est pas en mesure de le reconnaître, et il ne répondra à priori pas à l'enchère du SSP. En effet, le DSP ne connaît alors rien du profil de l'utilisateur : son *bid* (son enchère) serait alors sûrement trop bas.
   - Si il a déjà eu un *call* pour $cookie_{SSP}$, alors le DSP est en mesure d'identifier l'utilisateur en faisant correspondre le $cookie_{SSP}$ et son propre cookie (nommé $cookie_{DSP}$) (voir à partir de l'étape 13 pour le processus de *cookie matching*). Grâce au profil que le DSP a établi de l'utilisateur, il décide de *bidder* ou non.
8. Le SSP reçoit les enchères des différents DSPs qui ont choisi de *bidder*, et choisit le DSP le mieux offrant. Il répond alors au navigateur d'appeler le DSP gagnant.

C. L'affichage de l'encart publicitaire

9. Le navigateur interroge le DSP gagnant.
10. Le DSP répond avec les coordonnées de l'*AdServer* de l'annonceur.
11. Le navigateur interroge l'*AdServer* de l'annonceur.
12. L'*AdServer* de l'annonceur envoie l'encart de publicité à afficher.

Il faut environ 50 millisecondes pour passer de l'étape 1 et l'étape 12, afin que l'encart de publicité puisse s'afficher quasi-instantanément pour l'utilisateur. Le processus pourrait s'arrêter là. Cependant, lorsque les DSPs décident ou non d'enchérir à l'étape 7, ils ont besoin d'identifier l'utilisateur pour savoir si ils sont intéressés par lui servir une publicité. Or l'utilisateur est identifié par le SSP par un $cookie_{SSP}$ (celui qui est envoyé pour l'enchère), et par le DSP par un $cookie_{DSP}$. Il faut permettre à chaque DSP de faire correspondre (« *matcher* ») ces deux cookies.

D. Le cookie matching

13. Lorsque l'encart publicitaire a été servi à l'utilisateur, le SSP indique au navigateur d'appeler tous les DSPs (et pas uniquement celui qui a gagné l'enchère – il peut y en avoir une cinquantaine).
14. Chaque DSP est alors en mesure de lire le cookie$_{DSP}$ qu'il a pu placer précédemment (ou peut le placer si c'est la première fois qu'il se fait appeler par l'utilisateur). Il peut alors associer son cookie$_{DSP}$ au cookie$_{SSP}$ qui lui été envoyé par le SSP pour identifier l'utilisateur. Ainsi, la prochaine fois que le SSP lancera des enchères pour cet utilisateur en envoyant le cookie$_{SSP}$, chaque DSP saura qu'il s'agit de l'utilisateur qu'il identifie dans ces bases de données par cookie$_{DSP}$.

Le lien avec les empreintes numériques

Les empreintes numériques sont au cœur de la technologie de RTB. En effet, ce sont les empreintes numériques de l'utilisateur qui permettent aux annonceurs de décider quelle publicité lui servir. En pratique, si le profil d'un utilisateur montre qu'il a un intérêt pour l'automobile, un annonceur qui cherche à placer une campagne de lancement d'une nouvelle voiture misera plus cher qu'un annonceur en lien avec un autre sujet pour remporter l'espace publicitaire.

Ainsi, l'annonceur peut acheter un profil en particulier, et non pas un emplacement spécifique, sur lesquels il a beaucoup moins de visibilité.

La création de nouveaux services

La publicité permet de transformer nos empreintes numériques en valeur monétaire classique. Il s'agit toutefois d'un cas particulier des modèles bifaces, et il existe des sources de valorisation de nos empreintes au cœur même des échanges de l'économie numériques. Nous pensons ici notamment à tous les outils collaboratifs qui doivent leur apparition à l'existence même de nos empreintes.

> Payer sa boite e-mail avec ses empreintes numériques, n'est en effet pas la même chose que de collaborer à la création de GPS intelligent en fournissant ses propres informations de déplacement.

Enfin, la troisième source de financement de cette activité de collecte d'empreintes que nous avons pu identifier est le *marketing* ou la stratégie. Là encore, il s'agit de modèles bifaces où l'acteur en contact direct avec les consommateurs fournit des données sur ceux-ci aux concepteurs de nouveaux produits.

Vendre des données ?

La publicité en temps réel et son modèle de marché non régulé incitent aujourd'hui à obtenir un maximum d'informations sur les utilisateurs.

Prenons un exemple simple pour illustrer ce principe : l'entreprise d'automobiles X lance une campagne de communication sur le portail d'information en ligne géré par l'entreprise Y et est prête à payer 1$ à chaque fois qu'un utilisateur clique sur sa bannière publicitaire (ce qui le redirige par exemple vers le site présentant le nouveau produit de X). On suppose pour simplifier qu'en moyenne, une personne sur dix clique sur cette publicité : l'entreprise Y vendra donc son espace publicitaire à 10 centimes l'unité. Ainsi dix impressions couteront 1$ et en moyenne un des dix utilisateurs qui les visualiseront cliquera dessus.

Supposons maintenant qu'un trader arrive sur ce marché. Ce trader propose une affaire en or au gestionnaire du plus grand blog automobile sur Internet : il lui propose d'acheter les identifiants de sa base d'utilisateurs pour 1$ les cent identifiants. Ce trader est malin car il s'est renseigné : les utilisateurs de ce site sont des passionnés et cliquent deux fois plus que la moyenne des gens sur les publicités automobiles ! Le trader achète ensuite l'espace publicitaire de l'entreprise Y, en temps réel, mais seulement pour les identifiants qu'il connaît déjà, ceux des utilisateurs du blog. Pour le trader, deux personnes sur dix cliquent donc sur la publicité : comme il se fait payer 1$ pour chaque clic par l'entreprise X, dix impressions lui rapporteront en moyenne 2$. Or dix impressions ne lui coûtent que 1$, et le trader gagnera de l'argent quasiment sans aucun risque !

Cette technique, que l'on appelle arbitrage, est très classique pour les marchés financiers où elle a la vertu de stabiliser des cours de bourses interdépendants. Dans notre exemple ci-dessus, ces marchés interdépendants sont celui de la publicité en ligne et celui de la vente de données par le blog automobile. L'arrivée du trader ne change rien pour la plupart des acteurs : l'entreprise Y continue de facturer son espace à 10 centimes, l'entreprise X continue de payer 1$ le clic, et les utilisateurs du site sont même mieux lotis puisque, la publicité étant mieux ciblée, ils sont moins importunés par une publicité automobile pour laquelle ils n'ont aucun intérêt. Pour l'entreprise gérant le blog automobile, la situation s'améliore aussi puisqu'elle peut désormais monétiser sa base d'utilisateurs. Seuls les utilisateurs du blog se trouvent floués car ils ne sont pas avertis de la transaction qui s'opère derrière leur dos : certes cela permet de financer le blog automobile qu'ils consultent régulièrement, mais certains seraient peut-être même prêts à payer le site pour que celui-ci garde cette information secrète (rappelons tout de même que cette information a été vendue dans notre exemple 1 centime au trader, somme qui d'après nos informations est même largement surestimée par rapport à la réalité) !

Ce système a donc de nombreuses vertus, mais l'aspect le plus dérangeant est l'opacité des techniques utilisées. Certaines entreprises du marché de la publicité en temps réel se livrent aujourd'hui à des pratiques similaires, transposées au cas légèrement plus complexe du ciblage comportemental, mais aucune ne communique dessus : il ne faut surtout pas effrayer le grand public qui pourrait demander à garder ses informations secrètes, voir même à lui aussi être rémunéré !

On ne fait pas d'omelette sans casser des œufs

Big Brother is watching you

"War is peace.
Freedom is slavery.
Ignorance is strength."

"If you want a picture of the future, imagine a boot stamping on a human face—for ever."

George Orwell, 1984

Les empreintes, ou « traces », numériques sont ainsi une formidable source de création de valeur, et donc de croissance pour notre nouvelle économie numérique. Pourtant de nombreuses voix dénoncent aujourd'hui l'impuissance des utilisateurs à protéger leurs propres données. Comment expliquer une telle défiance face aux entreprises qui écrivent notre avenir ? Qui pourrait donc bien se plaindre que nos données, cette nouvelle ressource produite de manière illimitée par chacun d'entre nous, serve désormais à financer le développement de nombreux services innovants ? Ces lanceurs d'alertes ne sont-ils que des obscurantistes refusant toute innovation numérique ?

Vers la disparition de la vie privée ?

L'utilisation des données personnelles présente naturellement un certain nombre de risques qu'il convient d'éviter, allant des problématiques de vie privée à des considérations plus larges sur la protection des consommateurs. Le premier, le plus frappant et le plus documenté de ces risques apparaît dans l'œuvre emblématique de George Orwell, *1984*. L'auteur y décrit une société de contrôle dans laquelle le gouvernement épie ses citoyens jusque dans les moindres détails de leurs vies et se sert de ces informations pour les opprimer. Un monde qui fait passer l'ex-RDA pour un véritable camp de vacances libertaire.

Un tel risque paraît-il aujourd'hui complètement absurde ? Après tout, dans les sociétés démocratiques, la longue expérience des régimes totalitaires a conduit à la mise en place de nombreux garde-fous chargés d'empêcher la mise en place d'un tel système. Alors, est-ce une lubie de quelques Cassandre prêtes à tout pour attirer la lumière des médias ?

La réalité semble bien plus complexe, même dans les sociétés qui accordent une importance primordiale aux libertés individuelles. Nous avons notamment pu nous en rendre compte à la lumière des révélations faites sur le programme PRISM, mis en place par les autorités américaines (et les services de sécurité des USA) afin d'obtenir des différents opérateurs télécom – Internet et téléphonie – une quantité de données personnelles considérable. Ainsi, l'opérateur Verizon est obligé d'envoyer aux services de sécurité américains les relevés téléphoniques de ses millions d'abonnés. L'existence de portes dérobées dans les serveurs des géants américains de l'Internet tels Microsoft, Yahoo, Google, Facebook, AOL, Skype, YouTube, et Apple est presque avérée.

Or, bien que la longue tradition de protection des libertés individuelles des Etats-Unis ait été mise à mal ces derniers temps par des extensions du FISA[38] et du Patriot Act[39], celle-ci n'en reste pas moins forte. Si la surveillance a pu atteindre un tel niveau dans une démocratie très attachée à la protection de la vie privée, il est permis de s'interroger sur l'utilisation des empreintes numériques que pourraient faire des gouvernements moins regardants sur la question. Pour Julien Assange, éditeur en chef de WikiLeaks, il est clair que l'avancée des technologies de l'information annonce la fin de la vie privée pour la majeure partie de la population mondiale[40].

La vie privée, un problème de vieux cons ?

Mais même dans l'hypothèse où les gouvernements seraient proprement contrôlés et ne sombreraient pas dans le totalitarisme, la vie privée dans son acception actuelle pourrait être amenée à disparaître. Aujourd'hui déjà, Facebook ou Google disposent de quantités considérables d'informations qui permettent de dresser des portraits-robots très complets de leurs utilisateurs, contenant parfois des informations dont eux-mêmes n'ont parfois pas encore conscience, comme une grossesse[41] ! On comprend bien la sensibilité de ce genre d'information et les conséquences (psychologiques, au moins) que peuvent avoir la révélation involontaire de ces éléments de vie privée.

---

[38] Foreign Intelligence Surveillance Act, loi adoptée en 1978 (et amendée à plusieurs reprises depuis les attentats du 11 Septembre 2011) qui définit les procédures pour la surveillance et la collecte d'informations sur les puissance étrangères.

[39] Loi adoptée au lendemain des attentats du 11 septembre qui restreint les contraintes des agences de sécurité pour la collecte d'informations et étend leur juridiction.

[40] Opinion exprimée dans une tribune du *New York Times* intitulée : « *The banality of don't be evil* ».

[41] http://bugbrother.blog.lemonde.fr/2012/09/30/facebook-sait-si-vous-etes-gay-google-que-vous-etes-enceinte-et-ta-soeur/

Cet état de fait ne semble pas déranger outre mesure une bonne partie des utilisateurs des réseaux sociaux, en particulier les plus jeunes. Le CEO de Facebook, Mark Zuckerberg, déclarait d'ailleurs récemment que sa mission était d'apporter de plus en plus de connexions entre les gens et que la vie privée était un concept dépassé, voire rétrograde, qui n'intéresserait plus que quelques réactionnaires, réfractaires au changement. Serait-ce à dire, pour paraphraser Jean-Marc Manach[42], que la vie privée est « un problème de vieux cons » ? Qu'elle est amenée à disparaître pour faire place au progrès des réseaux sociaux ?

Fort heureusement, la réalité n'est pas aussi simple. De nombreuses études tendent à démontrer que les utilisateurs de services Internet sont de plus en plus inquiets du respect de leur vie privée, et qu'ils n'hésitent pas dans certains cas extrêmes à faire pression sur les services à qui ils ont confié leurs données. Ceci étant dit, Alessandro Acquisti[43] a récemment démontré que plus un individu a le sentiment de contrôler ses données, moins il est prudent avec elles[44].

En conclusion, et sans tomber dans le catastrophisme, il nous semble que, même si ces risques de contrôle de la vie privée sont aujourd'hui bien réels, ils sont suffisamment observés et analysés pour que les institutions chargées de les contrôler fassent leur travail, ou que d'autres acteurs spécialisés se chargent de les dénoncer mieux que nous ne pourrions le faire dans le cadre de ce mémoire.

Les positions dominantes inhérentes à l'économie numérique et l'asymétrie d'information entre le consommateur individuel et la compagnie récoltant des millions de profils est patente et peut entraîner de sévères disfonctionnements de marchés.

Pour cette raison, nous souhaitons analyser dans la suite une catégorie de risques différente, elle aussi bien réelle mais systématiquement passée sous silence par des médias préférant frapper l'imagination de leurs lecteurs en parlant du risque totalitaire. Il s'agit des risques liés à la protection des consommateurs.

---

[42] Jean-Marc Manach est un journaliste français spécialisé dans les nouvelles technologies de l'information et de la communication ainsi que leur impact sur la vie privée. Il est l'auteur d'un livre intitulé : *La vie privée, un problème de vieux cons ?*

[43] Alessandro Acquisti est chercheur en économie comportementale à l'université de Carnegie Mellon, à Pittsburgh, où il enseigne l'ingénierie de la vie privée.

[44] http://www.nytimes.com/2013/03/31/technology/web-privacy-and-how-consumers-let-down-their-guard.html

La personnalisation anonyme

Big Brother, l'arbre qui cache la forêt

Dans le scandale du programme PRISM, il est intéressant de remarquer que ce qui a choqué l'opinion publique, c'est avant tout que la NSA ait pu avoir un accès facilité aux informations hébergées par les géants américains d'Internet comme Microsoft, Yahoo!, Google ou Facebook. Le fait que des groupes privés disposent aussi facilement de ces informations, alors que nous savons qu'ils cherchent à les monétiser, ne choque visiblement plus personne. Rappelons tout de même que la NSA est une agence gouvernementale agissant pour la sécurité des Etats-Unis et de ses alliés – dont nous faisons encore parti.

Comment peut-on expliquer ce paradoxe ? Max Weber définit l'Etat moderne par le monopole de la violence physique légitime. Il apparaît alors naturel aux citoyens de chercher à protéger leurs vies privées des programmes de surveillance des Etats. Au contraire, quel est le pouvoir de nuisance d'un réseau social comme Facebook ? Aucune police politique « facebookienne » ne cherchera à se débarrasser de ses opposants, au pire, Facebook dévoilera au monde entier une photographie embarrassante…

Ce raisonnement nous semble incorrect. En effet, c'est oublier l'utilisation de services en ligne de plus en plus nombreux : commerce en ligne, banque, assurance, stockage sur le *cloud* etc. Mais c'est surtout minimiser – ou bien ne pas connaître – les utilisations que peuvent faire les entreprises privées de nos empreintes numériques. La menace de Big Brother est l'arbre qui cache la forêt, et l'objectif de ce chapitre est de montrer que la « protection des données personnelles » est plus importante que la seule protection du droit fondamental à la vie privée.

La différence majeure est que, comme expliqué précédemment, les services Internet peuvent être personnalisés en fonction d'empreintes numériques sortant du champ des données personnelles, phénomène que nous qualifions de « personnalisation anonyme ». En effet, pour pouvoir mettre en œuvre les différentes techniques que nous présenterons dans ce chapitre, il n'est souvent pas nécessaire de savoir qui utilise le service. Ce qui est important, c'est de savoir quel profil de consommateur l'utilise.

Le *behavioural pricing*

La discrimination par les prix est un mécanisme que l'on retrouve dans de très nombreux marchés, dans lequel un agent module les prix de son offre en fonction des caractéristiques de la demande afin de maximiser son profit. Dans le cas particulier du *behavioural pricing*, le vendeur module ses prix en fonction du profil qu'il a pu déduire des empreintes numériques de son client.

Dans le monde « réel », on connaît déjà de nombreux exemples de *behavioural pricing* basé sur la connaissance qu'a le vendeur de son client, par exemple lorsque le fleuriste augmente les prix la veille de la Saint Valentin, ou lorsque le vendeur de légumes sur le marché fait une réduction à ses clients les plus fidèles. Sur Internet, nos empreintes numériques sont disponibles en quantité trop importante pour que les e-commerçants ne soient pas tentés de les utiliser. Les opportunités sont nombreuses, pour les vendeurs mais également pour les acheteurs (leurs empreintes peuvent en effet suggérer qu'il en va de l'intérêt du vendeur de baisser ses prix).

Les tentatives de mise en pratique ont été nombreuses par le passé :

- En septembre 2000, Amazon.com crée la polémique lorsque des clients découvrent qu'ils ont acheté le même DVD à des prix différents. Après avoir vidé ses cookies, un utilisateur affirme avoir vu le prix d'un DVD passer de 26,24$ à 22,74$. Amazon mit un terme à cette affaire en affirmant que les variations observées étaient le résultat d'un test aléatoire, et remboursa les utilisateurs qui avaient payé plus cher.
- Une *start-up* allemande, Kreditech.com a une méthode originale pour calculer les scores de crédit des demandeurs de prêt. Kreditech s'appuie sur « *toutes les données en ligne qui peuvent être trouvées sur un individu* » (soit plus de 8000 points) afin d'attribuer des micro-prêts. Pour les 15% de demandes de prêt qui sont acceptées, l'argent (150€ en moyenne) est en 6 minutes sur le compte en banque de demandeur.
- Toujours dans le domaine du crédit, les californiens de Neo Finance évaluent l'assise professionnelle du candidat au crédit via le réseau social professionnel LinkedIn. Où travaille-t-il ? Depuis combien de temps ? Si son réseau professionnel est vaste et composé de nombreux *senior managers*, il devrait pouvoir accéder à un prêt avantageux : Neo affirme faire économiser en moyenne 50% en intérêts aux jeunes professionnels.
- Une étude[45] montre que le site de e-commerce Staples.com affiche des prix différents à ses utilisateurs en fonction de leur localisation géographique.

---

[45] Mikians, Jakub, et al. "Detecting price and search discrimination on the Internet." *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. ACM, 2012.

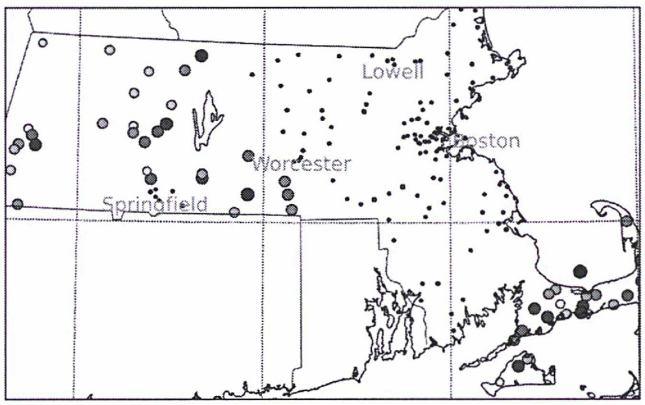


Figure 8 - Figure 10 : Différence de prix pour des utilisateurs du Massachussetts sur Staples.com
La taille du point montre le surplus moyen de prix,
de 0% (petits points) à 3,9% (gros points).

Nous détaillons dans ce qui suit deux applications concrètes du *behavioural pricing* en ligne, afin d'identifier et d'apporter une réponse aux questions que leurs utilisations peuvent soulever.

L'IP Tracking

« Donne moi ton IP, je te dirai combien tu paies »

Vous rentrez dans un magasin de chaussures. Vous essayez une paire, elle vous plaît mais vous ressortez pour en essayer d'autres dans les magasins d'à côté. Lorsque vous revenez le lendemain, elles sont plus chères. Et si vous revenez le surlendemain après avoir demandé son avis à un ami, le prix a encore augmenté.

L'*IP tracking* permet de moduler le prix proposé en s'adaptant à l'intérêt supposé de l'acheteur :

- Lorsque l'utilisateur fait une première recherche de prix, le site de vente lui propose un prix p.
- En même temps que le site propose un prix, il enregistre l'adresse IP de l'utilisateur et l'associe à la requête, gardant en mémoire l'intérêt de l'utilisateur.
- Si l'utilisateur choisit de conclure la vente, il paiera le prix p. S'il choisit au contraire de quitter le site (par exemple pour chercher d'autres informations concernant ce bien, pour comparer les prix avec d'autres vendeurs), et qu'il y retourne un peu plus tard en effectuant la même recherche, le site lui propose un prix $p+s$, en rajoutant un supplément s au prix précédemment proposé.

L'augmentation de prix est justifiée par l'hypothèse que, si l'utilisateur réitère sa recherche, alors il est fortement intéressé par cet achat. De plus, le vendeur cherche à provoquer la vente en simulant une augmentation de la demande, l'utilisateur voulant éviter une nouvelle augmentation de prix, il aura tendance à préférer conclure immédiatement la vente plutôt que la remettre à plus tard.

Il ne faut pas confondre l'*IP tracking* avec la technique dite du *yield management*[46].

Dans le cas du *yield management*, tous les acheteurs voient le même prix à un instant donné, prix qui est modifié par le vendeur en fonction d'une évolution de la demande (qui peut être supposée).

De plus, alors que le *yield management* est une technique commerciale développée et acceptée, l'*IP tracking* apparaît illicite pour deux raisons :

- D'une part, le site vendeur doit enregistrer dans ses bases l'adresse IP du prospect, afin de pouvoir l'identifier lorsque celui-ci reviendra une seconde fois. Ceci est interdit en France par la CNIL, qui considère l'IP comme une donnée à caractère personnel.
- D'autre part, cette activité est une pratique commerciale déloyale contraire au droit de la consommation. L'article L120-1 du Code de la consommation dispose que « *Une pratique commerciale est déloyale (...) lorsqu'elle altère, ou est susceptible d'altérer de manière substantielle, le comportement économique du consommateur normalement informé et raisonnablement attentif et avisé, à l'égard d'un bien ou d'un service.* »

De nombreux témoignages d'utilisateurs assurant avoir été victimes de l'utilisation de l'IP tracking ont été largement relayés par la blogosphère et les sites d'information, notamment concernant l'achat de billets d'avion, au point que la députée européenne Françoise Castex saisisse la Commission Européenne sur ce sujet. En France, la CNIL et la DGCCRF se sont emparées du sujet, mais pour l'heure, aucune enquête n'a permis de mettre en évidence une telle pratique. Il faut tout de même noter que, même si des utilisateurs se plaignent de variations de prix incompréhensibles qu'ils attribuent à de l'*IP tracking*, la totalité de ces variations pourraient être uniquement due aux complexes algorithmes régissant le *yield management*, tenus secrets par les compagnies qui en ont fait un élément stratégique de leurs *business models*.

---

[46] Technique bien connue des voyageurs, qui se sont habitués à être dans un avion ou dans un train où personne n'a payé le même prix pour un service pourtant identique.

En effet, le *yield management* n'a pas pour effet systématique de faire augmenter les prix dans le temps. Nous avons par exemple suivi l'évolution temporelle des prix d'un billet Eurostar. Si nous n'avons pas pu observer d'effets d'un *IP tracking* éventuel (ce qui est confirmé par la direction d'Eurostar), on peut remarquer que si les prix augmentent globalement, il y a ponctuellement des baisses de prix.

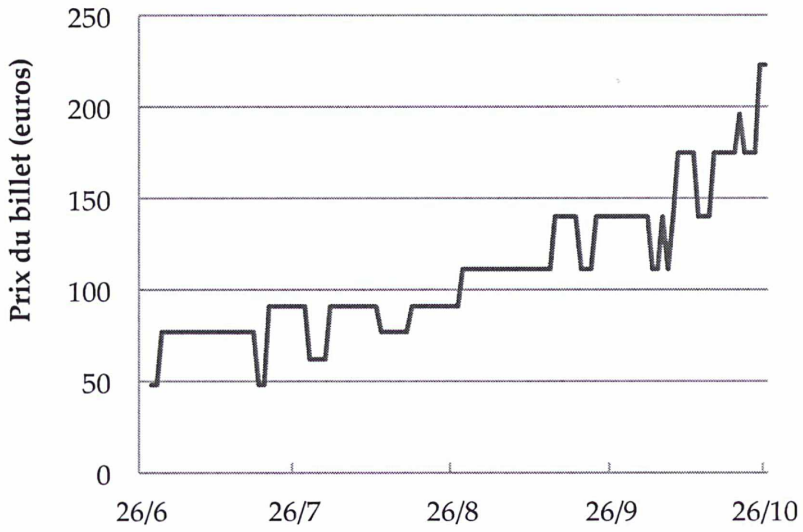


Figure 9 : Evolution du prix du billet d'Eurostar
Billet Standard pour un Londres-Paris le vendredi 26 octobre 2012 départ à 18h31

Alors, l'*IP tracking*, info ou intox ?

Cette pratique reste pour l'instant à l'état de rumeur, et certains s'étonnent qu'on la regarde avec autant d'attention. En effet, pourquoi le consommateur continuerait-il à utiliser des services d'e-commerce utilisant l'*IP tracking* sur des marchés concurrentiels où toute l'information est disponible grâce aux comparateurs de prix ? Ces outils sont devenus un réflexe chez beaucoup de consommateurs[47], et offrent une information gratuite sur les prix aux utilisateurs. Cette information n'est toutefois pas toujours des plus transparentes, comme le montre une enquête de la DGCCRF[48] :

---

[47] 92% des internautes français utilisent un comparateur de prix sur Internet avant l'achat. Source : *Les internautes et les comparateurs de prix*, IFOP, avril 2011.
[48] Enquête « *Pratiques commerciales sur internet : les comparateurs de prix* » effectuée par la DGCCRF en 2006.

« le panel des sites référencés est souvent présenté à tort comme exhaustif (cas de 5 sites dont un privilégiait manifestement ses partenaires commerciaux) »

D'autre part, il existe de nombreuses parades techniques pour changer d'adresse IP au moment de l'achat final afin de contrer l'*IP tracking*. Par exemple, éteindre et rallumer sa box[49] ou bien utiliser un VPN[50] permettent d'utiliser une adresse IP « toute neuve ». Cependant, comme nous l'avons vu, il existe d'autres identifiants que l'IP permettant d'identifier un appareil de manière unique ou quasi-unique sur le réseau. Et pour ces types d'identifiant, redémarrer simplement sa box Internet ne sera pas suffisant.

[49] Certains fournisseurs d'accès à Internet attribuent une adresse IP dynamique, qui change lors du redémarrage de la box Internet.

[50] Un « *Virtual Private Network* » (réseau privé virtuel) permet de créer un tunnel sécurisé entre deux réseaux via Internet. Une nouvelle interface réseau se crée, qui attribue ainsi une nouvelle adresse IP.

Qui sera assuré par l'assurance de demain ?

« On a la possibilité d'identifier les personnes qui vont sur Internet chercher des informations sur des médicaments. Le côté effrayant, c'est le manque de vie privée. Mais le côté génial, c'est l'opportunité que ça représente. Les informations sont disponibles. Nous devons aller les chercher. »

Alain Buerger, CEO de Coventry[51]

Les informations disponibles sur Internet peuvent également être utiles pour mieux connaître les risques associés à un individu en particulier, une information qui peut intéresser les assureurs. Certains termes de recherche sont par exemple de bons indicateurs pour prédire la propagation des maladies. Google Flu[52] parvient ainsi à corréler le nombre d'internautes recherchant des termes liés à la grippe et le nombre de cas de personnes présentant les symptômes.

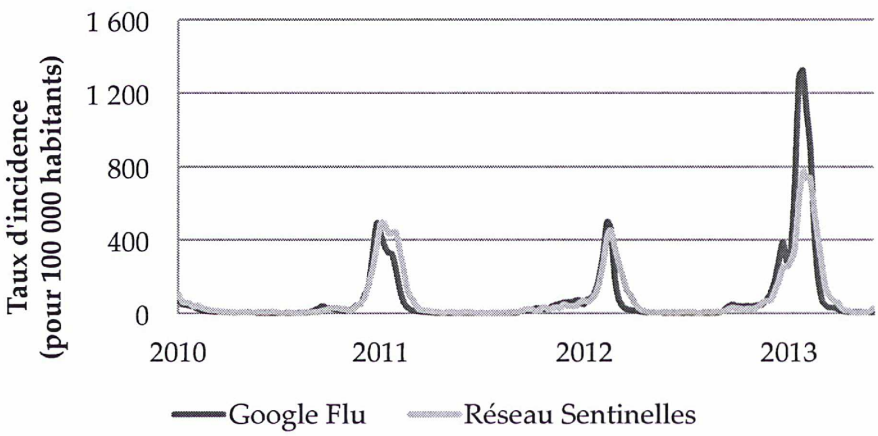


Figure 10 - Corrélation entre la prédiction de Google Flu et celle du réseau Sentinelles

Si l'on peut donc déduire des informations de santé à l'échelle macroscopique grâce aux données de recherche, il en est de même au niveau individuel: savoir qu'un internaute consulte fréquemment la rubrique de doctissimo.fr[53] consacrée au cholestérol n'est pas sans intérêt.

---

[51] Coventry est leader de l'industrie du rachat de contrat d'assurance vie. Source : Pièces à conviction, « *Banquiers : ils avaient promis de changer* ». L. Richard et J.-B. Renaud. Emission du 15/05/2013.

[52] http://www.google.org/flutrends/

[53] Sans tomber dans la paranoïa, il est intéressant de remarquer que Google, Tweeter et Facebook sont en mesure de suivre votre parcours sur doctissimo.fr

Une meilleure information sur un particulier est-elle bénéfique ? A priori, toute information supplémentaire est bonne à prendre …

Prenons l'exemple d'une maladie fatale, et supposons que la population est répartie en deux groupes : les « bons » qui vont contracter la maladie avec une probabilité très faible, et les « mauvais » qui ont une probabilité forte de contracter la maladie. Supposons qu'il y ait une totale asymétrie d'information en faveur de l'assuré, l'assuré seul connaissant le groupe de risque auquel il appartient. L'assureur ne pouvant pas discriminer les individus, il proposera une assurance uniforme. Celle-ci n'est pas tenable. En effet, certains « bons » jugeront l'assurance trop chère pour leurs risques, ce qui aura pour effet d'augmenter le prix, et les « mauvais » restants quitteront à leur tour l'assurance[54]. Seuls les « mauvais » seraient alors assurés, et à un prix prohibitif. Cette « spirale de la mort » est ce que le prix Nobel Arrow, dans un article fondateur de l'économie de la santé[55], appelle « *une sélection adverse défavorable des risques* ». Toute acquisition d'information supplémentaire augmente l'utilité globale. Dans cette logique, les sociétés d'assurance demandent des informations (âge, sexe, situation de famille, antécédents médicaux, antécédents familiaux) pour évaluer les risques d'un potentiel assuré.

A contrario, l' « effet Hirschleifer »[56] démontre l'effet de destruction des opportunités d'assurance engendré par l'acquisition d'information. En effet, si l'historique de recherche d'un potentiel assuré permet à la société d'assurance de déterminer avec une quasi-certitude à quel groupe de risque il appartient, il n'y a plus d'assurance : la prime devenant alors égale au traitement.

La littérature en économie de l'information ne permet donc pas de trancher sur les bénéfices et inconvénients possibles d'une connaissance accrue des individus en matière d'assurance, qui serait rendue possible par l'exploitation des empreintes numériques.

D'autre part, si on peut voir des tarifications très différenciées en ce qui concerne l'assurance non-vie, par exemple en automobile, ce n'est pas le cas en assurance vie et santé. En effet, en assurance

---

[54] Une « solution » qui permet aux assureurs de surmonter cette asymétrie d'information provient du travail de Rothschild et Stiglitz (1963). Pour cela, l'assureur doit amener les assurés à s'auto-sélectionner en proposant un contrat à couverture partielle et prime basse (pour les bons) et un contrat à couverture totale et prime élevée (pour les mauvais).

[55] Arrow, K. J., 1963, « Uncertainty and the welfare economics of medical care », *The American economic review*, 53, pp. 941-973.

[56] Hirshleifer, J., & Riley, J. G. (1992). *The analytics of uncertainty and information*. Cambridge University Press.

non-vie, les assureurs cherchent à influer sur le niveau de risque de l'assuré par les systèmes de bonus-malus, pratique interdite en France pour les assurances vie et santé.

On peut cependant imaginer un système qui résoudrait cette difficulté par la sélection : lorsque l'internaute se connecte à un site d'assurance en ligne, celui-ci ne se verrait proposer qu'un certain nombre de contrats, ceux qui correspondent à sa structure de risque, parmi les nombreux contrats disponibles dans le portefeuille de l'assureur.

Le narcissisme 2.0

« Amusez les Rois par des songes,
Flattez-les, payez-les d'agréables mensonges,
Quelque indignation dont leur cœur soit rempli,
Ils goberont l'appât, vous serez leur ami. »

La Fontaine, Les Obsèques de la Lionne

De nombreux sites web cherchent à s'adapter à l'utilisateur qui les consulte, afin lui fournir le service le plus pertinent possible. Cette « plasticité » caractérise le web 2.0 : le contenu d'une page web est désormais dynamique.

La bulle filtrante du web

Nous nous sommes habitués, un peu sans nous en rendre compte, à voir de plus en plus souvent un Internet personnalisé. Par exemple, lorsqu'un internaute français demande à un moteur de recherche « concerts Rolling Stones», c'est bien les concerts des Rolling Stones ayant lieu en France qui l'intéressent, et non pas ceux dans le Kentucky. Selon cette logique, Amazon met en avant les produits que « *vous apprécierez peut-être également* » : après tout, si un internaute utilise fréquemment Amazon pour acheter des bandes dessinées, pourquoi lui proposer des livres de poésie du XV^{ème} siècle ?

Mais le filtrage des contenus est parfois présent là on ne l'attend moins.

La plupart des gens imaginent par exemple qu'en recherchant « changement climatique » sur Google, ils obtiendront les résultats que Google pense être les plus pertinents sur le sujet. En réalité, depuis le lancement de *Personalized Search* pour les internautes disposant d'un compte Google en

2005[57], service étendu en 2009 à tous les internautes[58], le moteur de recherche personnalise les résultats en fonction des utilisateurs.

Pour illustrer ce phénomène, nous avons créé deux comptes Google *ex nihilo*. Nous avons ensuite « entrainé » le premier utilisateur (A) à correspondre à un profil « intellectuel », le faisant consulter des nombreux sites d'actualités, de blogs, et le faisant chercher des informations sur des sujets économiques, politiques etc. A l'inverse, nous avons entrainé le second utilisateur (B) à correspondre à un « touriste » préparant ses vacances d'été : comparateurs de prix d'avions, sites de réservation d'hôtels, informations sur les meilleurs *spots* de plongée, sur les plus belles plages de la Méditerranée, sur la météo etc. Nous étions alors début juin 2013 : alors que manifestants et policiers se faisaient face en Turquie, nous avons fait exécuter la même recherche Google par A et B : « Turquie ». Sur les 10 premiers résultats du moteur de recherche, tandis que A obtenait 7 liens pointant vers des articles couvrant les manifestations (Le Monde, Le Point, Libération...), B n'en obtenait que 2, recevant à la place de nombreux liens pour un séjour en Turquie (Thomas Cook, Club Med, Nouvelles Frontières...). A et B ont chacun été enfermés dans leurs bulles respectives.

De plus en plus de sites se livrent à cette personnalisation : l'algorithme *EdgeRank* de Facebook ajuste le flux de ses abonnés en fonction des liens sur lesquels il a l'habitude de cliquer, afin qu'un militant UMP ne soit pas pollué par les messages de ses contacts de gauche. Yahoo! News personnalise grâce à *CORE* le flux d'actualités que voit un utilisateur, ce qui lui a permis d'augmenter de 300% le nombre de clics sur les pages d'actualités. Sans que l'utilisateur ne s'en rende compte – tout ceci est complètement transparent pour lui – il commence petit à petit à évoluer dans un monde *on-line* qui lui est unique.

> *« la technologie va être tellement bonne qu'il va être très difficile pour les gens de voir ou de consommer quelque chose qui n'a pas été d'une manière ou d'une autre façonné pour eux »*[59].
> *Eric Schmidt, CEO de Google*

[57] *Search gets personal*. Google, 28 juin 2005
[58] *Personalized Search for everyone*. Google, 4 décembre 2009
[59] Interview de Eric Schmidt dans le *Wall Street Journal*, 14 août 2010.

Le côté obscur du web personnalisé

A priori, cette personnalisation des services semble être dans l'intérêt des internautes, qui trouveront plus facilement ce qu'ils cherchent sur Internet puisque des algorithmes filtrent les contenus les moins pertinents pour l'utilisateur.

Mais ce web personnalisé possède aussi un côté obscur, identifié dès 2010 par Eli Pariser dans un livre intitulé « *The Filter Bubble: What the Internet Is Hiding from You* ». Un web qui ne nous renvoie uniquement notre propre image, ce qui nous est familier et confortable, nous enferme en effet dans notre propre bulle. Le web que nous voyons n'est alors déterminé qu'en fonction de ce que nous sommes supposés vouloir voir.

Que se passe-t-il si je ne vois qu'un seul type d'information ?

Avant Internet, journalistes et éditeurs décidaient de l'information qu'allait recevoir le public. On ne pouvait que s'en remettre à leur sens de l'éthique et leur professionnalisme pour répondre au « *droit du public à une information de qualité, complète, libre, indépendante et pluraliste* », droit introduisant la Charte d'éthique professionnelle des journalistes.

Que peut-on attendre d'algorithmes qui ne prennent en compte que le taux de clic lorsqu'il s'agit de décider d'afficher ou non une information, dans l'unique but de maximiser des recettes publicitaires ? Où trouver les informations remettant en question ce que nous pensons ? Que deviendront les sujets de fond face à la presse à sensation ; alors qu'il est prouvé que même sur les sites d'actualités les plus respectables, le sensationnel et les faits divers attirent beaucoup plus les internautes que les grands sujets de société ?

Le phénomène de bulle a tendance à radicaliser les opinions des utilisateurs, qui ne sont confrontés qu'à leurs propres croyances. Alors qu'Internet permet de connecter les internautes du monde entier, les réseaux sociaux auront tendance à mettre en avant les « amis » avec qui vous avez le plus d'affinités, et à cacher les autres pour maximiser le taux de clic sur les contenus qu'ils partagent.

Certains diront qu'il en va de la responsabilité des internautes de chercher la diversité des points de vue et de sortir de leur bulle, mais il ne faut pas oublier que toute cette personnalisation est transparente : les algorithmes passent nos empreintes numériques à la moulinette avant de nous afficher une page web qui a tout l'air d'une page « neutre ». Et même lorsque les utilisateurs sont conscients de l'existence de cette bulle, il n'est pas toujours facile de s'en affranchir.

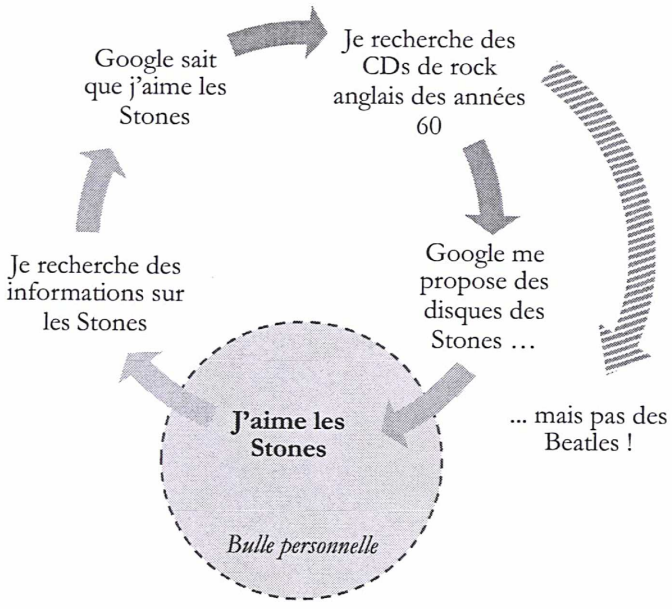


Figure 11 : La guerre entre Rolling Stones et Beatles par Google

Cela dit, force est de constater que les internautes ne cherchent pas forcément à remettre en question ce qui leur est proposé. En effet, 34% du trafic sortant de Google provient des sites sortis premiers lors d'un résultat de recherche[60].

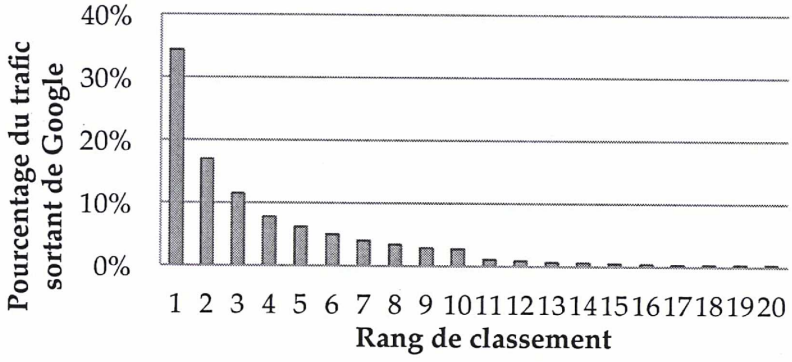


Figure 12 : Pourcentage du trafic sortant de Google en fonction du rang de classement

[60] Source : *The Value of Google Result Positioning*. Chitika. 25 mai 2010

Et si ma bulle était tout simplement fausse ?

Il faut garder à l'esprit que derrière toute cette personnalisation, ce ne sont que des algorithmes qui travaillent. Nous ne nous permettrons bien évidemment pas ici de remettre en question les compétences des ingénieurs de la Silicon Valley, mais il nous paraît tout de même important de conserver un regard critique et de ne pas céder à la confiance absolue dans des algorithmes objectifs et omniscients.

Tout algorithme introduit par exemple des biais systématiques. Les titres à base de jeux de mots ne seront pas pris en comptes par Google News, qui préférera mettre en avant les articles ayant une forte densité en certain mots-clés[61].

Enfin, les empreintes numériques sont le carburant des algorithmes décisionnels, et rien n'assure que ces empreintes soient correctes. Il semblerait que Google Search utilise 57 signaux différents pour décider quels sont les résultats de recherche. Et si ces signaux étaient faux ? C'est en tout cas que ce semble indiquer une étude concernant les données utilisées par les publicitaires[62], étude qui conclut que « *la moitié de mes données sont fausses, mais je ne sais pas de quelle moitié il s'agit* ».

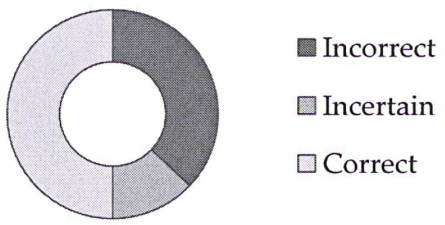


Figure 13 : Exactitude des données revendues à des fins publicitaires

Les algorithmes de personnalisation transforment profondément l'Internet, en structurant *a priori* l'espace dans lequel l'internaute évolue. La personnalisation à outrance crée des boucles de rétroaction qui masquent des pans entiers de la réalité à l'internaute. Les potentialités de création de services sont énormes, mais il nous paraît important d'être en mesure d'incorporer une composante éthique dans ces algorithmes qui prennent une place de plus en plus importante sur Internet. Les services web doivent être suffisamment transparents pour que les utilisateurs puissent avoir conscience de l'existence de cette bulle filtrante et de la possibilité d'en sortir.

---

[61] Source : *Understanding bias in computational news media.* N. Diakopoulous. 10 décembre 2012.
[62] Source : *Data Accuracy Survey Results.* Enliken. 6 mars 2013.

Un risque pour le *business*

Nous avons, jusqu'à présent, évoqué en détails les différents risques encourus par les particuliers et inhérents à la récolte d'empreintes numériques. Mais les entreprises doivent elles aussi faire face à de nouveaux risques, allant des risques d'image aux risques réglementaires.

La CNIL, fossoyeuse des start-ups innovantes ?

« C'est un ennemi de la Nation »

Gilles Babinet, à propos de la CNIL

L'un des plus importants de ces risques est selon nous celui d'une législation protectrice des données personnelles qui ignorerait le potentiel économique de leur utilisation raisonnée. C'est le sens de l'interjection récente de Gilles Babinet[63] qui déclarait récemment que la CNIL devait soit subir une réforme en profondeur, soit disparaître purement et simplement, au motif que sa régulation trop stricte en ferait un véritable « *ennemi de la Nation* ».

Cette appréhension paraît à première vue légitime, et contraindre nos entreprises à protéger les données personnelles pourrait en effet devenir un désavantage compétitif vis-à-vis d'entreprises étrangères n'ayant pas à s'embarrasser de telles préoccupations, et n'ayant pas à investir dans de coûteuses méthodes de protection. Limiter les capacités d'innovation et éventuellement couper les ailes de certaines start-ups est naturellement assez mal reçu à l'heure où l'on s'interroge sur l'absence de grands champions européens de l'Internet. La plupart des pays membres de l'Union Européenne semblent d'ailleurs partager cet avis, puisque le projet de règlement européen sur les données personnelles a été refusé le 6 juin 2013, principalement parce qu'il était jugé trop contraignant pour les industriels du secteur.

Ceci étant dit, cette inquiétude nous semble en réalité très exagérée. Tout d'abord, la CNIL ne se préoccupe pas des éventuelles transgressions de jeunes entreprises innovantes. Les capacités d'intervention de la CNIL sont aujourd'hui limitées, et ce sont avant tout les problèmes à grande

---

[63] Gilles Babinet, ancien président du Conseil National du Numérique, est actuellement le représentant français du programme des champions du numérique auprès de l'Union Européenne

échelle sur lesquels le Conseil concentre ses efforts. Dans la réalité, l'interaction entre la CNIL et les start-ups de l'écosystème numérique est bien plus du registre du guidage et de l'aide que de la sanction. Par ailleurs, les pays où fleurissent les géants de l'Internet (et désormais du traitement des données personnelles sur Internet) ne se privent pas de protéger leurs propres citoyens et de sanctionner les entreprises prises en flagrant délit. L'exemple le plus éclairant à ce sujet est le procès intenté aux Etats-Unis par des particuliers à Facebook et Rapleaf pour traitement de données illégal, alors même que Facebook est un champion national américain et que Rapleaf est une jeune entreprise innovante très prometteuse. Cette attitude protectrice vis-à-vis des données personnelles ne semble pourtant pas fondamentalement impacter les capacités d'innovation américaines. Ce ne sont pas les géants de l'Internet américain qui manquent !

Enfin, il ne faut pas oublier que la confiance est un élément primordial dans l'économie numérique. C'est le cas dans de nombreux autres aspects de l'économie comme les scandales alimentaires récents l'ont montré, mais tout porte à croire que la confiance est un élément encore plus important dans le numérique que dans le reste de l'économie, notamment au vu de la versatilité des utilisateurs sur Internet. D'autant plus que l'on imagine bien qu'un utilisateur serait prêt à partager beaucoup plus de données s'il avait l'assurance qu'il pouvait contrôler ce à quoi elles servaient. Les travaux d'Alessandro Acquisti le prouvent indirectement, ainsi que le développement de nombreux acteurs dans ce qui est en train de devenir une véritable filière de la confiance numérique. Les très fortes réactions des acteurs bien établis dans les différents scandales de données personnelles qui les ont atteints dernièrement ainsi que leurs investissements dans une apparente transparence vont aussi dans ce sens.

Un marché de la publicité en difficulté

Puisque les entreprises ont elles-mêmes intérêt à porter la confiance numérique, pourquoi défendent-elles des positions apparemment contradictoires, notamment dans leurs efforts de lobby auprès des régulateurs ?

Malheureusement, la pression des utilisateurs n'est pas le seul élément auquel les acteurs d'Internet sont confrontés. Les réalités financières sont d'autres formes de pression qu'ils ne peuvent pas faire passer au second plan. La conjoncture actuelle est difficile pour les entreprises du numérique dont la majorité des revenus provient de la publicité en ligne. Le marché mondial de la publicité représentait près de 800 milliards de dollars en 2012. Sur ces 800 milliards, 100 étaient dépensés sur Internet. D'après les estimations de Google – dont l'objectivité sur la question pourrait cependant être mise en doute – cette part numérique pourrait atteindre 400 milliards en 2017. Est-ce à dire que les revenus des compagnies de l'Internet sont assurés pour les années à venir ?

Dans les faits, l'estimation de Google paraît très optimiste. Au dernier trimestre 2012, la publicité sur Internet stagne aux Etats-Unis, son marché le plus important. D'autres signes semblent mettre à mal ces estimations optimistes, comme la méfiance des marchés vis-à-vis de certaines entreprises basées sur la publicité sur Internet – notamment Facebook, dont le cours de l'action a été très chahuté durant les premiers mois qui ont suivi son introduction en bourse. Ces entreprises subissent donc des pressions de la part de leurs actionnaires, qui souhaitent maximiser leur retour sur investissement. Enfin, les clients – c'est-à-dire les annonceurs publicitaires – se montrent de plus en plus regardants quant à la qualité de la publicité sur Internet. Ils exigent de nouvelles métriques pouvant prouver l'efficacité de cette forme de publicité, ce qui semble bien compliqué. Par exemple, Facebook n'a réussi à prouver un lien entre publicité sur le réseau social et achat que fin 2012 !

Devant ces difficultés, Facebook innove et propose de nouvelles solutions et de nouveaux services. « Promouvoir » permet de modifier le nombre de personnes qui voient les publications de votre compte, crucial pour les entreprises comme le montre le graphe ci-dessous.

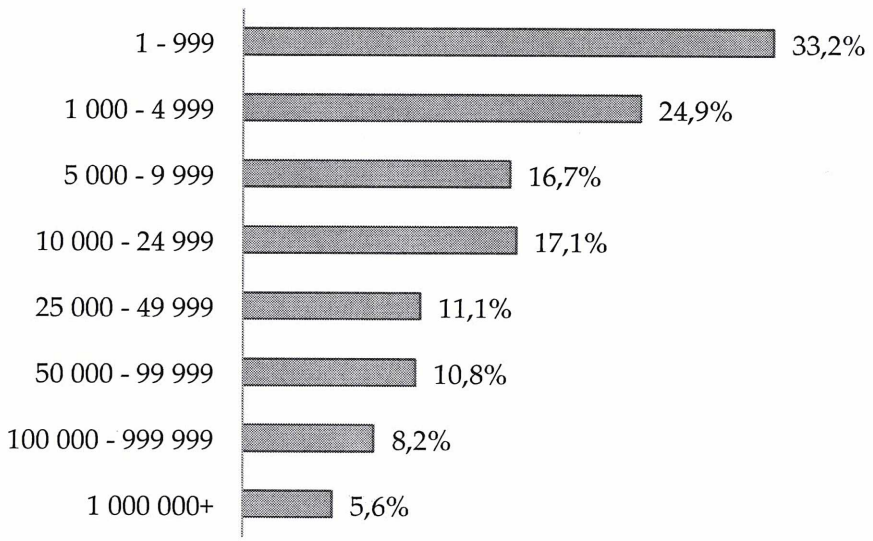


Figure 14 : Pourcentage de fans atteints par un post en fonction du nombre de fans de la page[64]

GraphSearch, est une autre fonctionnalité permettant de faire des recherches croisées sur l'ensemble des données renseignées par les utilisateurs de Facebook (les pages "aimées", les lieux "visités", les entreprises dans lesquelles ils ont travaillé etc.). Ces recherches permettent des recoupements très

[64] Source : *What's The Average Reach Of Your Facebook Post?* SocialBakers, 2012

efficaces : pour donner un exemple parlant, on peut rechercher les employés de MacDonald's qui aiment la page de son concurrent Quick. Enfin, de récentes études[65] ont souligné quelques faits alarmants sur la perception des consommateurs, comme :

- 30% des internautes sont convaincus que la publicité sur Internet ne fonctionne pas ;
- 50% sont convaincus que les bannières publicitaires n'ont aucune incidence sur eux ;
- 66% sont convaincus que la télévision reste un meilleur vecteur pour la publicité.

Cette dernière statistique est particulièrement intéressante. En effet de nombreux analystes affirment que seule la publicité vidéo conserve une efficacité sur Internet. En conséquence, les grands acteurs (Google avec YouTube, Apple avec l'Apple TV etc) investissent fortement le champ de la télévision connectée, possible catalyseur du basculement de la publicité sur Internet et future vache à lait du secteur. Or cette solution demandera l'utilisation d'encore plus de données personnelles pour nourrir des algorithmes performants afin de créer des chaînes de télévision individuelles, véritable graal de la TV connectée (et donc des annonceurs publicitaires).

Sous fortes pressions à la fois internes et externes, les acteurs de l'écosystème sont amenés à devenir de plus en plus « créatifs » dans l'utilisation des empreintes numériques, avec tous les risques que cela comporte.

Emergence de modèles respectueux de la vie privée

Le dernier risque auquel les entreprises de l'écosystème des empreintes numériques sont soumises, est paradoxalement celui de la confiance. Nous l'avons déjà souligné, cette confiance numérique est absolument essentielle pour parfaire l'essor de l'économie sur Internet, alors comment pourrait-elle représenter un risque ?

Tout d'abord, si elle peut représenter un atout pour les entreprises vertueuses, elle peut aussi s'avérer être un handicap important pour les mauvais élèves. Ainsi, le scandale PRISM éclabousse au moins autant les entreprises qui ont collaboré avec les services secrets américains que ces services eux-mêmes. Le *buzz* est rapidement créé et les risques d'image encourus par les compagnies sont très sérieux. D'autre part, si la confiance représente une opportunité, il ne faut pas la manquer, sous

[65] dont notamment une menée fin 2012 par Adobe Systems, *The State of Online Advertising*.

peine de se faire remplacer très rapidement – l'internaute étant notoirement changeant et sans pitié – par des services alternatifs proposant une meilleure qualité de confiance. Pour le moment, ce risque paraît limité, et il existe déjà des alternatives *open source* à des services tels Facebook (Diaspora etc.), mais ceux-ci ne rencontrent pas une adhésion forte. Les alternatives actuelles sont l'œuvre de quelques individus n'ayant pas les ressources nécessaires pour créer de véritables alternatives aux modèles moins respectueux des données personnelles. Mais cette situation pourrait changer très rapidement, dès lors qu'un nouvel acteur économique trouvera son intérêt à investir dans de tels services.

En conclusion, les opportunités offertes par le traitement et l'exploitation des empreintes numériques sont donc excitantes, mais aussi accompagnées de nombreux risques à la fois pour les utilisateurs et pour les entreprises du secteur. D'où l'intérêt actuel très fort des législateurs sur le sujet.

Alors finalement, que doit-on faire ?

Le profilage commercial, ainsi que d'autres activités inhérentes à l'économie numérique que nous avons décrites, ont poussé les acteurs industriels à développer un marché des informations personnelles, et plus généralement des empreintes numériques. Ce marché est une source d'innovation industrielle puissante mais il n'est pas dénué de risques.

L'asymétrie d'information, associée à des problématiques que l'on pourrait qualifier d'*externalité négative*[66] (l'entreprise bénéficie entièrement de la récolte des données de ses clients mais, au contraire de son client, ne subit quasiment aucune perte si les données sont divulguées ce qui conduit à une situation de surexploitation des données) doivent pousser les puissances publiques à se poser la question « Que faire ? ». Il s'agit en effet d'un résultat classique pour les économistes, même les plus libéraux, que ces deux facteurs conduisent naturellement à des défaillances de marché pouvant être suffisamment graves pour motiver des interventions (l'exemple le plus consensuel de ces dernières années est sûrement celui de la pollution et du réchauffement climatique).

Nous souhaitons nous intéresser dans cette partie aux diverses pistes, privées ou publiques, envisageables pour encadrer les pratiques d'exploitation des empreintes numériques.

[66] Une externalité négative caractérise le fait qu'un agent, de par son activité, fait supporter un effet externe négatif à autrui.

Régulation par le marché

Les initiatives privées peuvent être plus flexibles, mieux adaptées et moins coûteuses qu'une intervention publique dont la légitimité est toujours contestable.

Autorégulation : laissons faire la main invisible du marché

La première piste que nous présentons est celle soutenue par la plupart des acteurs industriels que nous avons rencontrés : l'autorégulation.

Ce terme même d'autorégulation peut sembler ambiguë et surtout contradictoire : « *regulation* » (réglementation en français) est en effet un terme anglais utilisé pour nommer un règlement administratif qui clarifie des droits et des responsabilités. L'autorégulation serait donc la transposée de ce terme dans le cas où les règles sont fixées par une partie auto-proclamée des acteurs ? Un concept étonnamment proche de l'efficacité de la main invisible dans la régulation des marchés, mythe dénoncé par Keynes dès 1926[67]...

En réalité, ce concept d'autorégulation est moins caricatural qu'il n'en parait : il désigne un ensemble d'initiatives multilatérales permettant d'œuvrer pour une concurrence plus transparente et plus efficace, ce qui bénéficie, en théorie, à tout le monde. Pour fonctionner de manière crédible, l'autorégulation doit appliquer les règles de séparation des pouvoirs : fixer les règles, contrôler leurs applications et arbitrer les conflits. Il existe des secteurs dans lesquelles ces trois pouvoirs sont assurés par les acteurs eux-mêmes : c'est le cas de l'ordre des médecins par exemple qui dispose d'un pouvoir de sanction en cas de non respect des règles éthiques dictées par la profession. On peut toutefois imaginer un système mixte, avec des règles fixées par les acteurs privés et un pouvoir de contrôle et de sanction laissé à la puissance publique : violer les règles d'autorégulation, c'est alors violer la loi.

La plupart des initiatives actuelles d'autorégulation se contentent aujourd'hui de fixer des normes et des règles de bonnes conduites. C'est le cas par exemple de la *Digital Advertising Alliance* qui publie une liste de principes adressée aux entreprises de publicité comportementale[68]. Ces initiatives misent avant tout sur la transparence et comptent sur le marché, autrement dit le consommateur, pour

---

[67] John Maynard Keynes, *The end of laissez-faire* (1926)
[68] http://www.aboutads.info

sanctionner les abus. La DAA propose ainsi l'achat d'une licence d'utilisation du logo « *Advertising Option Icon* » aux entreprises respectant ses principes.

L'intérêt de l'autorégulation est avant tout de responsabiliser l'écosystème dans son ensemble et d'éviter, lorsque cela est possible, un recours à une législation contraignante nécessairement coûteuse et handicapante pour l'innovation. De plus, les acteurs privés ont dans de nombreux domaines une expertise supérieure aux régulateurs, ainsi qu'une plus grande réactivité que le pouvoir législatif. Il faut toutefois remarquer que, bien que de telles initiatives puissent être parfaitement adaptées à certains cas, elles ne sont pas applicables à tout type de marché. La publicité comportementale est par exemple nécessairement en avance sur ses sujets, car le modèle de ces entreprises est entièrement basé sur la confiance.

Concurrents indirects et problématiques concurrentielles

Une autre piste possible d'évolution provient des autres entreprises de l'économie numérique qui ne sont pas directement liées au marché des empreintes numériques. Nous développons ci-dessous l'exemple du choix des paramètres par défaut du navigateur web Firefox, développé par la fondation Mozilla.

En mars 2013, les développeurs du navigateur Firefox ont indiqué que la prochaine version de leur navigateur serait dotée d'une nouvelle fonctionnalité : un blocage automatique des « *third-party cookies* », ces fichiers stockés sur nos ordinateurs par des sites que nous n'avons jamais visités. Cette annonce a suscité de vives réactions, notamment de la part des entreprises de publicité en ligne opérant en « *third-party* ». Ces dernières dénoncent une distorsion problématique de concurrence en faveur des géants américains pouvant se permettre d'opérer en « *first party* ». Il n'y a en effet qu'assez peu de différences techniques pour ces entreprises entre gérer des cookies « *first party* » (déposés directement par le site que vous visitez) ou des cookies « *third party* » (déposés par un tiers lors de votre visite sur le site), si ce n'est que la première solution nécessite un accompagnement des clients (les éditeurs de sites Internet) plus poussé, et donc une force commerciale très importante.

Le débat s'est cristallisé autours du choix de Mozilla d'activer ce blocage par défaut pour tous les utilisateurs. L'immense majorité des utilisateurs ne modifiant jamais les paramètres par défaut, on peut en effet voir ce choix comme l'introduction d'une barrière concurrentielle à l'entrée dans ce marché des empreintes numériques (publicité en ligne, outils analytiques etc.). Or c'est une particularité économique du monde numérique que de disposer de coûts de duplication et de

distribution quasi nuls. Il en résulte un poids prépondérant des coûts fixes dans les structures de coûts des entreprises de l'économie numérique, et donc un pouvoir économique important pour ceux qui dominent à un instant donné. Renforcer cette barrière naturelle ne pourrait alors avoir comme conséquence qu'une dégradation de la concurrence, source d'opacification du marché.

Il faut tout de même noter que Mozilla est depuis revenu sur cette décision. Dans un article publié sur son blog[69], le directeur technique Brendan Eich, explique qu'un blocage automatique est source de problèmes théoriques complexes, comme celui des faux positifs (si le site lemonde.fr décide d'héberger ses vidéos sur lemonde-videos.fr, comment déterminer automatiquement que cette dernière URL est « légitime » ?) ou de faux négatifs (si vous cliquez par mégarde sur une publicité qui vous renvoie sur un site, vous ne voulez à priori pas que ce dernier soit alors autorisé à récoler des informations sur vous). Même si la mise en place d'une telle mesure peut paraitre plus néfaste que bénéfique pour le long terme, toutes ces annonces ont tout de même eu un intérêt majeur: celui d'ouvrir le débat et d'obliger les différents acteurs à exposer leurs points de vue et à s'organiser pour proposer une contre-mesure efficace.

Le pouvoir des utilisateurs

Les utilisateurs sont souvent présentés comme souverains dans leurs choix d'utilisation des services Internet. Ceci permet d'influer sur le comportement des entreprises, le raisonnement sous-jacent étant que la réputation et les ventes d'une entreprise souffriront si celle-ci ne remplit pas les désirs des clients en matière de protection des empreintes numériques.

Des utilisateurs souverains ?

L'hypothèse même de souveraineté de choix des utilisateurs est en réalité contestable : tentez l'expérience et essayez de passer quelques jours sans utiliser un seul des services fourni par Google… Mais outre la dépendance, c'est l'insouciance des utilisateurs qui est le réel problème, ces derniers n'étant bien souvent pas conscients des pratiques ni des dangers potentiels. Bien peu de gens accepteraient de diffuser en libre accès la liste de leurs communications téléphoniques, ni même simplement leur carnet d'adresses email. Pourtant l'immense majorité fournit ces

[69] https://brendaneich.com/2013/05/c-is-for-cookie/

informations à des entreprises privées sans se soucier des protections mises en place. Est-on condamné à attendre le passage d'un cataclysme mondial avant de voir les mentalités changer ?

L'importance de la sensibilisation

Un axe d'évolution très important est ainsi celui de la sensibilisation des consommateurs. Même si l'éducation sur ces sujets ne peut provenir que de la puissance publique, un certain nombre d'ONG œuvrent pour plus de transparence et de communication en la matière. Deux paramètres sont en effet importants pour que l'argument de la souveraineté des utilisateurs soit réellement crédible :

- La sensibilité des consommateurs aux mesures de protections.
- L'ampleur de la communication des entreprises sur leurs efforts de protection.

Ces deux paramètres ayant une influence dynamique l'un sur l'autre (une publicité plus importante entraine rapidement une sensibilité plus importante des consommateurs), cet effort de transparence des ONG entraine un cercle vertueux vers une meilleure prise de conscience.

Nous ne pouvons nous pencher ici sur un examen exhaustif des très nombreux projets d'ONG ou d'associations en la matière et ne citerons donc qu'un exemple, celui du « *Big Brother Award* », décliné dans de nombreux pays dont la France[70]. Cette cérémonie vise à « récompenser » les personnes ou institutions qui représentent le mieux la société décrite par George Orwell dans *1984*.

Les solutions de contournement

En complément des actions de sensibilisation, certains utilisateurs sont suffisamment créatifs pour proposer des solutions de contournement. Les plus connues se nomment AdBlock, qui supprime purement et simplement toutes les publicités de votre navigateur, Ghostery et Collusion, deux systèmes similaires permettant de tracer et bloquer les « *trackers* », ou encore ToR (acronyme pour *The Onion Router*), un système beaucoup plus complexe permettant de naviguer « anonymement » sur Internet.

---

[70] En France, la cérémonie est organisée par l'association « Souriez, vous êtes filmé-e-s ». L'édition 2013 est la 7e édition, après une interruption de 2 ans pour cause de manque de financements.

Les initiatives publiques

Du côté des acteurs publics, la problématique principale est celle de l'étendue géographique pertinente pour une régulation de l'écosystème Internet : les discussions ayant lieu actuellement à l'échelle européenne auront nécessairement plus d'impact que la seule réglementation française. Mais la première question à laquelle nous souhaitons répondre est celle du rôle de l'Etat, et si celui-ci doit intervenir pour encadrer les divers pratiques.

Pour commencer, nous pouvons souligner qu'il y a beaucoup d'exemple où il est unanimement reconnu que l'Etat joue un rôle majeur. Et ces exemples comme la sécurité routière, la lutte contre le tabagisme ou la régulation financière sont assez similaires à notre problématique dans le sens où ils concernent à chaque fois deux types d'acteurs aux points de vue et intérêts conflictuels, dont les actions de l'un se font au détriment de l'autre groupe et où il faut trouver un arbitrage.

Le rôle de l'Etat

Comme nous l'avons souligné dans les pages précédentes, l'incitation à créer des technologies avec une protection accrue de la vie privée est aujourd'hui assez faible, ce qui ébranle fortement la crédibilité des options d'autorégulation ou de laissez-faire. Il ne reste alors plus que la puissance publique. Cette approche répressive permet d'atteindre deux buts principaux: compenser le préjudice subi, mais surtout inciter financièrement les entreprises en rétablissant les inefficacités de marché (faire en sorte que le bénéfice attendu d'une pratique non conforme soit inférieur au risque financier encouru).

Appréhender les limites de la voie réglementaire

« Sur Internet, il ne peut pas y avoir de souveraineté numérique, ou alors il faudrait nous expliquer qui serait en droit d'exercer cette souveraineté : l'Hadopi, le gouvernement français, celui des Etats-Unis ? »

Jean-Marc Manach

Même si l'on suppose que le régulateur est compétent, suffisamment informé et que son but est de réguler sur la base de l'optimum social, la mise en place d'une politique réglementaire induit nécessairement des coûts : pour l'Etat bien sûr, mais aussi pour l'industrie qui doit se conformer aux règles. Ces coûts sont à évaluer comparativement aux gains résultants d'une meilleure protection des données, mais la situation n'est pas nécessairement positive.

On peut schématiquement donner deux options possibles pour la voie réglementaire :

- Définir des règles très précisément, ce qui permet de limiter les coûts induits. Le problème des règles trop strictes est qu'elles ont tendance à être à la fois sur-optimales (elles peuvent interdire des pratiques ayant un bénéfice à la fois pour l'entreprise et pour le consommateur) et sous-optimales car elles ne peuvent prévoir par avance tout les cas de figure possibles.
- Définir des règles plus vagues, ce qui permet de contourner le problème de l'évolution des technologies en la matière, beaucoup plus rapide que celle de la législation (il faut compter au minimum 2 à 3 ans pour qu'un règlement européen soit adopté). Au contraire, des règles vagues augmentent significativement les coûts pour les entreprises et nécessitent de long procès pour arbitrer les conflits.

Une autre limite des politiques réglementaires est celle de la compétence et de l'indépendance du régulateur. Au vu de l'évolution des discussions sur le projet de règlement européen, on peut se demander si le pouvoir des lobbys est réellement plus légitime que le pouvoir du marché. De nombreux exemples, tirés d'autres industries, vont dans ce sens et montrent que dans certains cas, la politique réglementaire a pu conduire à des situations de protection des consommateurs inférieures à celles que le marché aurait générées sans contrainte.

Nous pouvons citer en exemple l'industrie du tabac aux Etats-Unis, qui remporta de nombreux procès grâce à la phrase imprimée sur tous les paquets « *WARNING: THE SURGEON GENERAL HAS DETERMINED THAT CIGARETTE SMOKING IS DANGEROUS TO YOUR HEALTH* ». Initialement, cette obligation fédérale de 1969 était sensée avertir les fumeurs des dangers médicaux de la cigarette. Toutefois, jusqu'en 2008, cette simple inscription suffit à libérer les industriels de la cigarette de toute autre obligation comme la recherche de technologies de filtres moins risquées ou la communication d'informations sur les dangers réels et leur permit de diffuser des publicités vantant les mérites de la cigarette en omettant de mentionner ses dangers[71]…

Enfin, Internet étant par construction un outil décentralisé, la notion de frontière territoriale y a peu de sens. Un contrôle étatique trop puissant ne peut qu'entrainer qu'un mouvement de réaction des consommateurs qui apprendront à utiliser les outils permettant de crypter les communications et d'apparaître anonyme sur le web. Un tel contrôle, en plus d'être inefficace, serait même néfaste puisqu'il compliquerait énormément la tâche de protection contre des fléaux actuellement relativement maîtrisés (discours ou images racistes, haineuses ou encore pédophiles).

---

[71] Pour plus d'informations, voir les articles concernant les procès Cipollone contre Liggett Group Inc (1992) et Altria Group contre S. Good (2008).

Faut-il plus de régulation ?

L'un des arguments principaux des ennemis de la régulation est que l'écosystème est en train de se mettre en place et qu'il ne paraît donc pas pertinent d'intervenir avant de connaître la direction prise par les acteurs de la filière. Ainsi, réguler dès maintenant, c'est prendre le risque d'empêcher des développements potentiels à haute valeur ajoutée. C'est aussi prendre le risque de pénaliser ses industries nationales face à la concurrence de pays plus souples dans leur approche.

La conclusion logique est que la seule régulation possible et souhaitable est l'autorégulation. Mais les résultats de la mise en place d'une autorégulation sont-ils toujours positifs ? On peut rapprocher la situation des empreintes numériques à celle, historique, d'autres écosystèmes en formation. Encore une fois, le parallèle avec l'économie de la cigarette nous semble pertinent.

Alors que la cigarette jouissait encore d'une excellente image dans la société, les cigarettiers ont pu se développer sans aucun contre-pouvoir. Ils ont ainsi pu devenir suffisamment puissants pour disposer d'un lobby solide lorsque les gouvernements ont commencé à vouloir prendre des dispositions pour protéger la santé publique. En revanche, aucun contre-lobby n'existait pour leur porter une approche contradictoire.

On comprend bien pourquoi le parallèle est frappant. Comme la cigarette, les réseaux sociaux et les entreprises qui traitent les données jouissent d'une excellente réputation. Il est aujourd'hui impensable pour les jeunes de ne pas avoir une présence sur Facebook. Ce genre d'entreprises n'a pas été encore trop inquiété par le législateur. Mais pour autant, alors que Bruxelles se décide enfin à voter un nouveau règlement en remplacement de la directive de 1995 (on peut imaginer à quel point celle-ci est devenu inadaptée), on se rend compte que l'industrie des empreintes numériques s'est dotée d'un lobby extrêmement puissant et organisé. Pour mémoire, plus de 4000 amendements ont été déposés à propos du projet de règlement européen, un record absolu en la matière !

C'est là la limite du modèle de l'autorégulation, qui entraîne une forte asymétrie de moyens entre les deux camps différents d'un débat et qui rend la régulation étatique très ardue. C'est donc se fermer cette option que de donner les pleins pouvoirs aux industriels du secteur.

La France et plus largement l'Europe profiteraient donc probablement d'une sensibilisation plus large autour de la problématique des données personnelles afin que des groupes de protection des consommateurs puissent se former et contrebalancer l'influence du lobby industriel dont les motivations ne sont pas nécessairement alignées avec celles du reste de la société.

De l'importance de la formation et de la recherche

Notre exposé précédent montre que toutes les pistes d'encadrement possible ont leurs limites : la politique réglementaire sera toujours en retard sur la technologie, et le laissez-faire du marché ne peut conduire qu'à une surexploitation des données dangereuse pour les consommateurs.

Dans ces conditions, que peut faire concrètement l'Etat en la matière pour protéger ces citoyens ?

Nous avons le sentiment que deux pistes importantes ont été, pour le moment, délaissées et peu explorées. La première est celle de l'éducation et de la formation des citoyens, et la seconde est un soutien public volontariste aux initiatives respectueuses.

Pourquoi former les utilisateurs ?

Nous l'avons vu, deux paramètres importants participent à freiner l'évolution « naturelle » du marché vers une position socialement stable :

- Une trop faible transparence des entreprises sur leurs efforts de protection, sur les données qu'elles récoltent, sur les utilisations qu'elles en font…
- Une trop faible sensibilité des consommateurs aux mesures de protections. La raison principale est que l'immense majorité des consommateurs ne s'intéresse pas aux enjeux du problème. Ceci est dû soit à un manque de conscience des risques, ce qui conduit à sous-protéger ses propres empreintes, soit à un manque de conscience des bénéfices possibles, ce qui conduit à surprotéger ou à une volonté de sur-réglementer.

Nous parlons ici d'évolution « naturelle » car c'est ce que la plupart des théories économiques prévoient : si la concurrence est « suffisante » (un concept restant malheureusement à définir), alors le seul désir des consommateurs permet d'influer sur le comportement des entreprises. Une entreprise qui s'écarterait du niveau de protection souhaité serait alors immédiatement sanctionnée par une baisse du niveau de ses ventes.

Informer les consommateurs aussi bien sur les risques que les bénéfices possibles serait ainsi un moyen élégant et peu coûteux de surmonter cette inefficacité actuelle du marché.

Nous souhaitons insister fortement sur le fait que ces deux actions (information sur les risques et information sur les bénéfices) doivent être menées en parallèle, sans accentuer l'une des deux

composantes. Une surprotection aveugle de tout type de données serait en effet au moins aussi handicapante qu'une sous-protection. Pour illustrer ce propos, prenons l'exemple de la santé : les données de santés traitées par les algorithmes du « *Big Data* » permettent aujourd'hui de changer fondamentalement la dimension des études épidémiologiques. Ceci permettrait d'ouvrir des portes extraordinaires en matière de santé publique. Doit-on s'y refuser au nom de la vie privée, prenant le risque de voir quelques entreprises peu scrupuleuses s'en emparer, ou doit-on au contraire créer un cadre favorisant ces innovations tout en empêchant les dérives ?

Comment former les utilisateurs ?

La formation et la sensibilisation des consommateurs à ces problématiques reste aujourd'hui très polémique en ce qui concerne les méthodes à utiliser : faut-il éduquer aux bons usages (comment protéger correctement les informations que l'on poste sur Facebook) ou bien expliquer les technologies sous-jacentes (comment fonctionne techniquement le bouton Like de Facebook).

En France, les parlementaires et l'éducation nationale réfléchissent à ces problématiques. On trouve ainsi dans la dernière loi d'orientation et de programmation pour la refondation de l'école de nombreuses mentions visant à favoriser les usages numériques à l'école :

> *« La formation à l'utilisation des outils et des ressources numériques est dispensée dans les écoles et les établissements d'enseignement ainsi que dans les unités d'enseignement des établissements et services médico-sociaux et des établissements de santé. Elle comporte une sensibilisation aux droits et aux devoirs liés à l'usage de l'internet et des réseaux, dont la protection de la vie privée et le respect de la propriété intellectuelle. »*
> Article L. 312-9 tiré de la section 2 du chapitre III « La formation à l'utilisation des outils numériques »

L'idée générale est de fournir une éducation par les usages, diffuse et transverse. Pour fonctionner, cette formation aux enjeux par les usages doit commencer très tôt (idéalement avant que ces usages ne commencent dans un cadre personnel) et se poursuivre tout au long du parcours scolaire. La seule manière satisfaisante pour atteindre cet objectif est d'intégrer cette formation aux unités d'enseignement existantes (philosophie, mathématiques, histoire etc.). Le véritable défi reste toutefois de former les professeurs à ces enjeux.

Nous pouvons tout de même sentir, dans la manière dont est écrit cet article L 312-9, une focalisation sur les dangers (« protection de la vie privée ») et les devoirs de citoyens et non sur les bénéfices

potentiels. Il s'agit de notre point de vue d'un strabisme très généralisé : les campagnes d'informations s'acharnent pour la plupart à expliquer qu'il y a des conséquences néfastes à utiliser Internet de façon peu prudente sans pour autant donner de solutions. Ce n'est pas en expliquant que vous pourrez, dans 10 ou 20 ans, regretter les photos que vous postez aujourd'hui sur les réseaux sociaux que l'on parviendra à modifier les usages !

Au contraire, il semble aujourd'hui indispensable de mettre l'accent sur les concepts permettant de comprendre ce qui se passe réellement (le fonctionnement dans les grandes lignes des réseaux comme Internet, celui des algorithmes, les limites des technologies de cryptage etc.). Pour prolonger notre analogie avec les enjeux de santé publique liés au tabac, nous pensons par exemple que l'intégration d'un programme de sciences naturelles (SVT) qui explique le fonctionnement de l'appareil respiratoire, la formation de cancers ou encore les phénomènes d'addiction est bien plus efficace que d'expliquer les conséquences néfastes de la cigarette en obligeant l'impression d'images chocs sur les paquets vendus.

L'école de la sécurité, nouveau domaine de l'excellence française ?

« Finalement l'affaire Prism, si elle est avérée, et même si de nombreux doutes avaient été émis sur ce point, rend relativement pertinent le fait de localiser des data centers et des serveurs sur le territoire national, afin de mieux garantir la protection des données traitées dans des clouds, avec des sociétés françaises implantées en France et soumises à la loi française. »

Fleur Pellerin

Pour toutes les raisons que nous avons exposées dans les premières parties de ce mémoire, nous sommes convaincus que les empreintes numériques représentent un enjeu majeur de la société de demain, notamment en terme de protection de la vie privée et du consommateur.

Pour autant, plutôt que de voir cette problématique comme quelque chose qu'on doit subir et qui nécessite une vigilance quotidienne des gouvernements et des consommateurs, on peut choisir d'en faire une opportunité stratégique. En effet, comme les modèles de protection restent en grande partie à définir et à développer, nous nous trouvons à la naissance d'un secteur qui est voué à devenir très porteur à l'avenir.

Dans cette perspective, la stratégie que nous recommandons et qui nous apparaît la plus intéressante est de se placer dès maintenant, en avance de phase, comme un des acteurs majeurs de cet

écosystème. Dans ce but, il nous faudrait recréer un écosystème des empreintes numériques en France – comme il peut en exister un en Californie.

Pour ce faire, il faudrait tout d'abord fonder une véritable école des empreintes numériques. Il s'agit en effet d'un domaine dans lequel la France, du fait de la qualité reconnue de ses pôles de recherche en statistiques et en mathématiques appliquées, a vocation à exceller. Il s'agirait d'étudier en toute indépendance les enjeux et problématiques liés à l'exploitation des empreintes numériques aussi bien du point de vue des sciences dites fondamentales (mathématiques, informatique etc.) pour comprendre l'étendue des possibilités, que de celui des sciences sociales (sociologie, psychologie…) pour comprendre les dangers complexes potentiels.

Développer dès aujourd'hui une telle initiative aurait un double intérêt :

- A court terme, ce pôle d'expertise permettrait de former une sorte de contre-lobby, ainsi qu'un système favorisant la réflexion des citoyens sur ces enjeux. Nous avons déjà développé l'importance des lobbys sur l'évolution actuelle de la réglementation, et une telle initiative nous apparaît comme particulièrement stratégique.

- A moyen terme, ces enjeux deviendront des atouts économiques exportables et ce pôle d'expertise fournira ainsi des sources de croissance bienvenues pour la France et l'Europe, dans des domaines aussi variés que le service à la personne, les transports, la santé, l'environnement… Ce choix stratégique pourrait même à terme nous faire rattraper le retard certain que nous avons pris dans les deux dernières décennies en terme de services Internet. La proportion d'entreprises américaine (et bientôt chinoise) dans ce secteur est très frappante, et les entreprises françaises y sont trop peu nombreuses.

Il nous semble que nous nous situons à une période clef du développement de l'écosystème des empreintes numériques. Les grands services américains vont souffrir pendant un certain temps des éclaboussures du scandale PRISM. Cependant, cette fenêtre d'opportunité va vite se refermer. On voit déjà que le scandale, s'il reste dans les esprits, s'estompe. Il est donc temps d'agir afin de replacer la France dans la course.

En guise de conclusion, nous vous présentons ici l'un des systèmes alternatifs protecteur des empreintes numériques dont la France pourrait appuyer le développement : la MarinièreBox.

La MarinièreBox

Actuellement, les systèmes des réseaux sociaux et des partages de fichiers sont centralisés, c'est-à-dire que les particuliers envoient leurs fichiers sur des serveurs centraux qui sont ensuite accessibles depuis partout moyennant une connexion internet. Ces serveurs centraux sont gérés par des entreprises privées. C'est ainsi le modèle des services comme Facebook et Dropbox (un service de partage de fichiers).

C'est cette centralisation des données qui fait la force de ces acteurs. Une solution simple serait de les écarter en décentralisant les empreintes numériques, par exemple en les stockant sur les disques durs des boxs internet. Ainsi, les fichiers, photos et autres informations des particuliers pourraient être stockés directement chez eux et non dans des serveurs centralisés.

Ensuite, on ajouterait une surcouche logicielle qui permettrait de partager certaines de ces informations avec d'autres particuliers, via des communications cryptées. Il ne serait pas difficile de recréer à la fois des réseaux sociaux et des services de partage de fichiers dans le cadre de cette architecture.

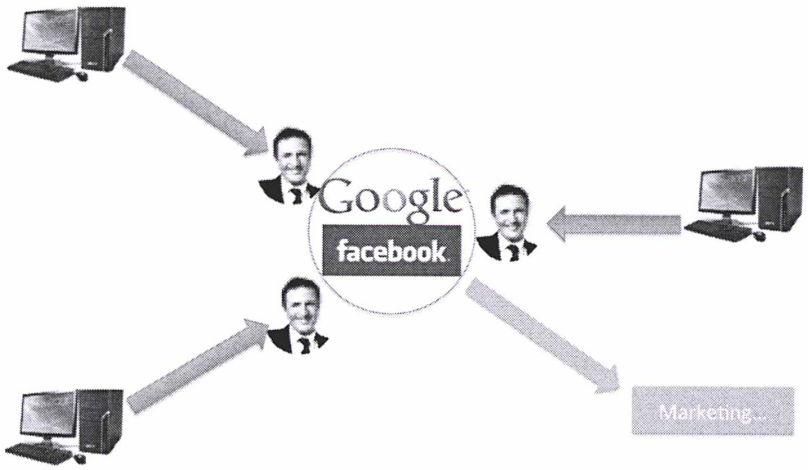
Un tel système donnerait aux individus un contrôle bien plus étroit d'une grande partie de leurs empreintes numériques que s'ils les échangent contre les mêmes services en version centralisée.

Bien évidemment, cela n'entraîne pas un abandon de l'exploitation possible des empreintes numériques. Il ne paraît pas pertinent de renoncer au formidable potentiel de leur traitement.

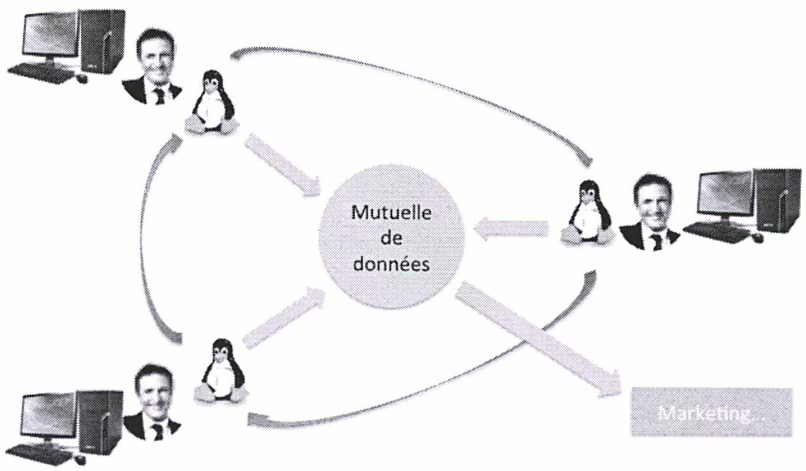
On peut imaginer mettre en place des services mutualistes auxquels on donnerait accès à une certaine partie de nos empreintes numériques et qui pourraient les négocier auprès des entreprises intéressées, pour ensuite reverser les profits effectués aux utilisateurs initiaux.

Ces acteurs auraient un sens économique puisqu'ils auraient accès à de nombreuses données (monnayer des données à un niveau individuel n'a pas de sens, on l'a vu), et pourraient ainsi redistribuer une partie de la valeur aux utilisateurs.

C'est la MarinièreBox. Nous sommes persuadés que ce genre d'initiatives est aisé à mettre en place et pourrait permettre à la France de se replacer dans la course aux services Internet dans laquelle nous avons déjà pris trop de retard.



L'écosystème des réseaux sociaux actuels



L'alternative MarinièreBox

Conclusion

Qu'est-ce qu'une donnée personnelle ? Cette notion peut vous sembler aller de soi : mon numéro de sécurité sociale est une donnée très personnelle, alors que la marque de mon ordinateur ou le navigateur que j'utilise ne le sont pas. Pourtant, nous l'avons vu, une combinaison de données à priori anonymes peut suffire pour vous identifier de manière unique en tant que consommateur sur Internet. Si une combinaison de données non personnelles peut devenir une donnée personnelle, quelle est l'utilité d'une protection juridique encadrant fortement tout traitement de donnée classifiée comme personnelle mais ne s'appliquant pas aux données censées être anonymes ? Dans une époque marquée par une capacité de calcul et un volume de données collectées toujours croissants, l'anonymisation de données apparait comme un mythe détournant dangereusement de l'analyse des risques potentiels liés à l'exploitation de nos données. Nous avons donc élargi notre analyse sur la protection des données à toutes les empreintes numériques, c'est à dire à toute donnée issue de l'utilisation d'un service Internet.

Ce sont aujourd'hui nos empreintes numériques qui servent de monnaie alternative dans l'écosystème de l'économie numérique. Nous permettant d'acheter de nombreux services gratuits, ces données sont une source extraordinaire d'innovation et de croissance. En contrepartie, les entreprises du numérique cherchent à valoriser nos empreintes numériques, ce qu'illustre bien le marché de la publicité en temps réel, certainement un modèle de valorisation parmi les plus aboutis à l'heure actuelle. Mais si ce marché est indéniablement source d'une forte création de valeur, son opacité actuelle a pour conséquence directe une surexploitation non efficace de nos données. Hors l'exploitation de nos données n'est pas dénuée de tout risque. Big Brother et la société de surveillance sont des risques majeurs, mais notre analyse nous mène à conclure que bien d'autres risques sont trop souvent passés sous silence. Au delà des enjeux de libertés individuelles, nous avons analysé les problématiques de protection des consommateurs, de concurrence et de compétitivité liées aux empreintes numériques, ainsi que les diverses pistes, privées ou publiques, envisageables pour encadrer ces pratiques. Toutes ces pistes ont leurs limites : la politique réglementaire sera toujours en retard sur la technologie, et le laissez-faire du marché ne peut conduire, dans les conditions actuelles, qu'à une surexploitation des données dangereuse pour les consommateurs.

Deux pistes sont d'après nous encore insuffisamment explorées. Tout d'abord l'éducation et la formation des citoyens apparaissent comme un moyen élégant de surmonter les inefficacités actuelles du marché. Nous proposons la mise en place d'une véritable école française de recherche sur les empreintes numériques ce qui aurait comme conséquence d'encourager à court terme la réflexion des citoyens sur ces enjeux, réflexion stratégique à la vue de l'influence des lobbies sur le sujet. De plus, cette expertise se transformera à moyen terme en une source de croissance, particulièrement bienvenue à l'heure actuelle, dans des domaines aussi variés que le service à la personne, les transports, la santé ou l'environnement. Enfin, la seconde piste consiste à ne plus considérer ces problématiques comme des états de fait subis que la société doit apprendre à gérer, mais bien au contraire de choisir d'en faire une opportunité stratégique. Ceci passe par un soutien public volontariste aux initiatives respectueuses. Les modèles de protection restent en grande partie à définir et à développer, et nous nous trouvons à la naissance d'un secteur sans aucun doute voué à un grand avenir.

Annexe. Personnes rencontrées

Thibaud Antignac	Chercheur en *Privacy and Computer Science* à l'Inria
David Baranes	*Country Manager* France d'Appnexus
Eric Barbry	Avocat à la Cour d'Appel de Paris Directeur du pôle Droit numérique du cabinet Alain Benoussan
Bertrand Barré	Président du Groupe The Zebra Company
Godefroy Beauvallet	Directeur du Fonds AXA pour la recherche Membre du bureau du Conseil National du Numérique
Pierre-Jean Benghozi	Membre du collège de l'ARCEP Directeur de recherche CNRS et professeur à l'Ecole Polytechnique. Membre du comité de la prospective de la CNIL
Frédéric Bellier	Directeur Général France de RadiumOne
Sabrina Bouguessa	*Legal Counsel* chez Criteo
François Bourdoncle	Co-fondateur et Directeur Technique d'Exalead
Sophie Bresny	Inspectrice principale au Centre de surveillance du commerce électronique (CSCE), Service National des Enquêtes de la DGCCRF
Johanna Carvais	Chargée de projet Labels à la Direction des affaires juridiques, internationales et de l'expertise de la CNIL
Olivier Desbiey	Chargé d'études prospectives à la Direction des études, de l'innovation et de la prospective de la CNIL
Julien Dourgnon	Conseille politique du Ministre du Redressement Productif
Emmanuel Florent	Directeur Général de Digital Virgo

Eric Freyssinet	Chef de la division de lutte contre la cybercriminalité au Pôle judiciaire de la gendarmerie nationale
Paul-Olivier Gibert	Président de Digital & Ethics Président de l'Association Française des Correspondants à la protection des Données à caractère Personnel
Albéric Guigou	Co-fondateur et Directeur Associé de ReputationSquad
Hubert Guillaud	Rédacteur en chef d'Internet Actu
Colin de la Higuera	Président de la Société Informatique de France Directeur-adjoint du Laboratoire d'Informatique de Nantes Atlantique
Xavier Hubert	Conseille juridique du Ministre du Redressement Productif
Willy Lafran	Co-fondateur et Directeur Général de Datarmine
Claire Levallois-Barth	Coordinatrice de la chaire « Valeurs et politiques sur les informations personnelles » de l'Institut Mines-Telecom Secrétaire générale de l'AFCDP
Amirhossein Malekzadeh	Co-fondateur de focusmatic
Jean-Marc Manach	Journaliste spécialisé dans les questions liées à la protection de la vie privée Auteur du blog Bug Brother sur leMonde.fr
Fabrice Mattatia	Responsable d'investissements numériques à la Caisse des Dépôts
Henri de Maublanc	Président du Groupe Clarisse
Xavier Piccino	Directeur adjoint de cabinet adjoint à la DGCCRF, ministère de l'Economie et des Finances
Jean Pinquet	Enseignant-chercheur en économétrie de l'assurance à l'Université de Nanterre – Paris X
Blandine Raoul-Réa	Adjointe-coordinatrice « experts TIC 2nd degré » et usages

	numériques, Bureau des usages numériques et ressources pédagogiques au Ministère de l'Education Nationale
Michel Riguidel	Professeur émérite à Telecom ParisTech
Luisa Rossi	*Regulatory Affairs Manager* chez Orange
Jean-Baptiste Rouet	*Managing Director* France de VivaKi Nerve Center, Publicis Groupe
Françoise Roure	Présidente de la section « Technologies et Société » du Conseil général de l'économie, de l'industrie, de l'énergie et des technologies
Hugo Roy	Chef du projet Terms of Service ; Didn't Read
Jean-Baptiste Rudelle	Co-fondateur et CEO de Criteo
Antoine de Saint-Affrique	Président *Food* d'Unilever
Juan-Miguel Santiago	Adjoint au chef du bureau 6B – Médias et Publicité à la DGCCRF, ministère de l'Economie et des Finances
Nathalie Sonnac	Membre du Conseil National du Numérique Directrice de l'IFP – Département de sciences de l'information et de la communication à l'Université Panthéon Assas
Henri Verdier	Directeur d'Etalab Membre du comité de la prospective de la CNIL
Estelle Werth	*Legal Director, Commercial & Privacy* chez Criteo
Jérémie Zimmermann	Porte-parole et co-fondateur de la Quadrature du Net