HAL
open science

L'analyse à grande échelle de nos traces numériques

Matthieu Latapy, David Chavalarias

To cite this version:

Matthieu Latapy, David Chavalarias. L'analyse à grande échelle de nos traces numériques. Mokrane Bouzeghoub; Rémy Mosseri. Les Big Data à Découvert, , 2017, Les Big Data à Découvert, 978-2-271-11464-8. hal-01575457

HAL Id: hal-01575457

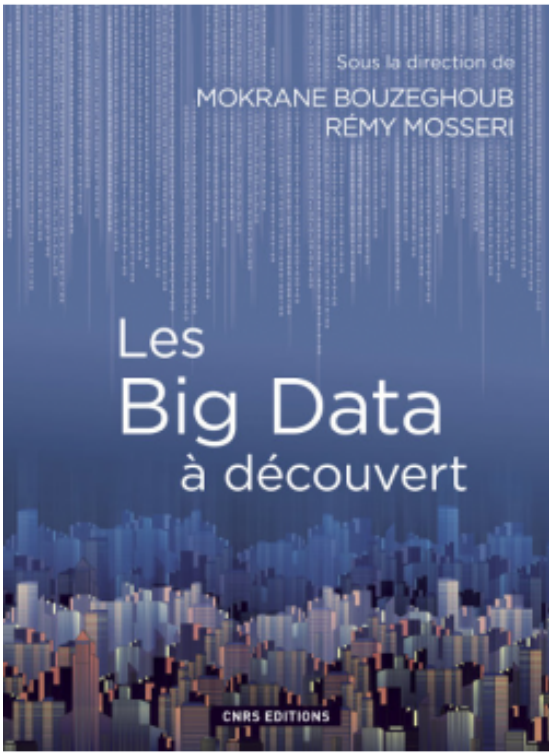
https://hal.science/hal-01575457v1

Submitted on 20 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright



8. L'analyse à grande échelle de nos traces numériques

Matthieu Latapy et David Chavalarias

De l'analyse de traces numériques à la prédiction des comportements, quelles sont les opportunités et limites des approches Big Data ?

En pleine révolution de l'Internet et du numérique, suite à la généralisation des objets mobiles et/ou connectés, une part importante des activités économiques et sociales migre vers des supports numériques. Ceci engendre l'enregistrement de traces très diverses à une échelle sans précédent, contribuant largement au phénomène Big Data : déplacements, relations sociales, rencontres physiques, échanges de messages, flux financiers, navigation sur le web, données de recommandation ou de consommation, données médicales, photographies, etc. Ces données ouvrent la voie à de nombreux progrès. Mais elles sont également exploitées à des fins inappropriées par toutes sortes d'acteurs, qui tirent parti des supports numériques pour en détourner les usages : soit en commettant des infractions ou des crimes rendus possibles par le caractère virtuel de ces environnements, soit en exploitant outre mesure les données privées à des fins qui vont à l'encontre des intérêts des acteurs qui les ont produites. Ceci donne lieu à un *dilemme du numérique* : les moyens à mettre en œuvre pour prévenir les usages illégaux mettent en danger la vie privée et les libertés individuelles.

Les activités criminelles numériques vont de la fraude financière (usurpation de carte de crédit par exemple) à l'exploitation sexuelle d'enfants (échanges de contenus à caractère pédo-pornographique), en passant pas le rançonnage, l'espionnage, l'usurpation d'identité, les attaques sur des services, et de nombreux autres. Parce qu'elles ont lieu dans ces espaces numériques, ces activités laissent des traces. Celles-ci seraient exploitables à des fins de prévention ou d'investigation pour autant que nous enregistrions les activités de l'ensemble de la population. Dès lors, on peut légitimement poser la question de la conservation et de l'exploitation de ces traces numériques, et en creux, la question des atteintes à la vie privée : sommes-nous prêts à réduire notre anonymat sur l'Internet pour faciliter la traque des pédophiles ? sommes-nous prêts à autoriser l'enregistrement de nos achats pour permettre la détection de fraudes ?

Le compromis entre sécurité et préservation de la vie privée est toujours délicat. En témoignent les débats houleux qui ont précédé l'adoption de la loi sur le renseignement en mai 2015, et qui ont vu le conseil constitutionnel plusieurs fois saisi. La législation actuelle

évolue vers l'enregistrement d'un volume toujours plus important de traces, sur lesquelles les forces de l'ordre reposent aujourd'hui de façon cruciale pour mener leurs investigations. Les malfaiteurs l'ont bien compris, et ils sont passés maîtres dans l'art de la dissimulation, de la discrétion, et de l'effacement de traces. Il s'agit pour eux que leurs traces ne soient pas distinguables de celles d'activités légitimes, voire d'éviter qu'elles soient enregistrées.

Dans certains cas, il est extrêmement difficile de cacher les traces d'une activité illégale. Pour les cas d'usurpations d'identité, par exemple le vol d'une carte de crédit, les comportements du malfaiteur sont rarement en cohérence avec les habitudes de la victime. Ces incohérences peuvent éveiller des soupçons sur des transactions, qui seront confirmés ou infirmés facilement suite à une vérification auprès de la victime potentielle. S'il n'est en général pas possible de retrouver l'auteur de la fraude, on peut au moins bloquer la carte et empêcher ainsi son utilisation frauduleuse.

Cette détection est efficace précisément parce que le fraudeur manque d'informations, et parce qu'il existe une personne auprès de qui vérifier les doutes. Cependant, pour autant que la proportion de malfaiteurs reste faible dans la population, ces doutes s'avéreront faux la plupart du temps car chacun a également une certaine probabilité de changer radicalement de comportement (suite à un divorce, au décès d'un proche, à un nouvel emploi, etc).

Ces types d'activités illégales engendrent ce qu'on appelle des *signaux faibles* : dans une énorme masse de données et de traces d'activités légitimes, ils constituent une partie extrêmement minoritaire de l'ensemble. Par exemple, le taux d'échanges de fichiers à caractère pédopornographique a été estimé à environ deux pour mille dans certains systèmes pair-à-pair[i]. De même, les usurpations d'identité lors de transactions financières en ligne restent extrêmement minoritaires. Il s'agit de signaux faibles typiques.

La détection de tels signaux soulève de nombreux défis scientifiques, qui commencent dès l'évaluation du résultat de la détection : si un détecteur de fichiers à caractère pédopornographique affirme systématiquement que le fichier qui lui est soumis n'entre pas dans cette catégorie, il aura raison 98,8 % des fois ; or il sera complètement inopérant ! A l'inverse, il est crucial d'éviter que de très rares déviances invalident un système complet. Il est ainsi souvent affirmé qu'il faudrait limiter l'usage du pair-à-pair ou même l'Internet au motif qu'ils sont sources de fichiers ou informations à proscrire ; pourtant, l'essentiel des échanges qu'ils soutiennent sont légitimes. De même, le commerce électronique est sévèrement freiné par la défiance des utilisateurs par rapport aux usurpations d'identité, pourtant exceptionnelles.

De nombreux efforts sont menés pour développer des méthodes et outils pour la détection de tels signaux faibles. On pourrait être tenté de mener une collecte indifférenciée et massive de données, comme par exemple l'analyse simultanée de comportements de toute une population, ou le croisement de données d'un même utilisateur sur plusieurs types de bases. Ce sont ces traitements complexes qui sont concernés par l'évolution de la législation : pour détecter des individus déviants et prédire les activités criminelles, il faudrait tout collecter, tout traiter, tout exploiter. Edward Snowden a par exemple révélé en 2013 l'application outre-Atlantique de cette doctrine de surveillance de masse suite aux attentats du 11 Septembre 2001.

S'il est probable que ces approches aient permis de démanteler des réseaux criminels ayant commis quelques imprudences numériques, des doutes importants subsistent sur l'efficacité réelle de telles mesures pour la détection de signaux faibles. Ceux-ci seraient plus susceptibles d'être décelés par des opérations ciblées. Ainsi, un rapport de 2013[ii], commandé

par la Maison Blanche à un panel d'experts, a conclu que le programme de surveillance globale de la NSA "n'avait pas été essentiel dans la prévention d'attaques" et que la plupart des éléments qu'il a permis de collecter en enregistrant indistinctement toutes les données de la population "auraient pu facilement être obtenus dans des délais appropriés en utilisant des autorisations conventionnelles."

Sans présager des progrès futurs qui pourraient améliorer l'utilité de l'analyse des méga-données dans un cadre préventif, il est important d'évaluer les programmes de surveillance de masse à l'aune de leurs limites et dérives potentielles. En Nouvelle-Zélande, des militants pro-démocratiques ont ainsi été privés de leurs droits fondamentaux suite à une mauvaise interprétation de leurs courriels par la NSA[iii]. Comme le montrent les révélations Snowden, rien qu'aux États-Unis les dérives sont multiples : espionnage industriel ou politique, opérations de dénigrement[iv]. En France, la collaboration de sociétés françaises avec la Libye quelques mois avant la chute de Kadhafi a permis de déployer des technologies de surveillance de masse[v] dont on sait désormais qu'elles ont permis d'identifier des opposants au régime, par la suite torturés ou éliminés[vi]. La mise en œuvre de ces méthodes dans nos démocraties doit donc tenir compte du fait que ces dernières ne sont pas éternelles, et que les bases de données ainsi constituées peuvent d'un jour à l'autre être exploitées par un régime d'une autre nature.

A un autre niveau, de larges bases de données sur les comportements des populations sont constituées à des fin commerciales par les grands groupes du l'ère numérique, les « GAFA » (Google, Apple, Facebook, Amazon). Elles permettent d'anticiper nos désirs afin de nous vendre plus de produits. Peut-être un jour seront-elles utilisées pour ajuster le montant de notre assurance santé, ou pour déterminer notre niveau d'adhésion aux idées prônées par le parti politique majoritaire, une voix que semble déjà explorer la Chine[vii].

Les outils de surveillance de masse constituent donc une épée de Damoclès pour nos démocraties, épée qui sera en outre toute prête à servir contre les concitoyens de celles qui seraient amenées à changer de nature. La question du Big Data n'est donc pas seulement un défi scientifique ou technologique : le dilemme du numérique qui leur est associé constitue un défi éthique pour nos sociétés modernes, qui ne peut se résumer à une opposition caricaturale entre atteinte à la vie privée et laxisme en termes de sécurité.

i M. Latapy, C. Magnien, R. Fournier. *Quantifying Paedophile Activity in a Large P2P System*, Information Processing and Management, 2013

ii Voir en particulier l'article du Washington Post: "Officials' defenses of NSA phone program may be unravelling" par Greg Miller et Ellen Nakashima, 19 Decembre 2013, http://www.washingtonpost.com/world/national-security/officials-defenses-of-nsa-phone-program-may-be-unraveling/2013/12/19/6927d8a2-68d3-11e3-ae56-22de072140a2_story.html

iii https://theintercept.com/2016/08/14/nsa-gcsb-prism-surveillance-fullman-fiji

iv Voir Glenn Greenwald (2014) *Nulle part où se cacher, Ed.* JC Lattès

v http://www.lefigaro.fr/international/2011/09/01/01003-20110901ARTFIG00412-comment-j-ai-mis-8-millions-de-libyens-sur-ecoute.php

vi http://www.liberation.fr/france/2016/03/15/l-allie-francais-du-paranoiaque-kadhafi_1439846

vii http://www.bbc.com/news/world-asia-china-34592186