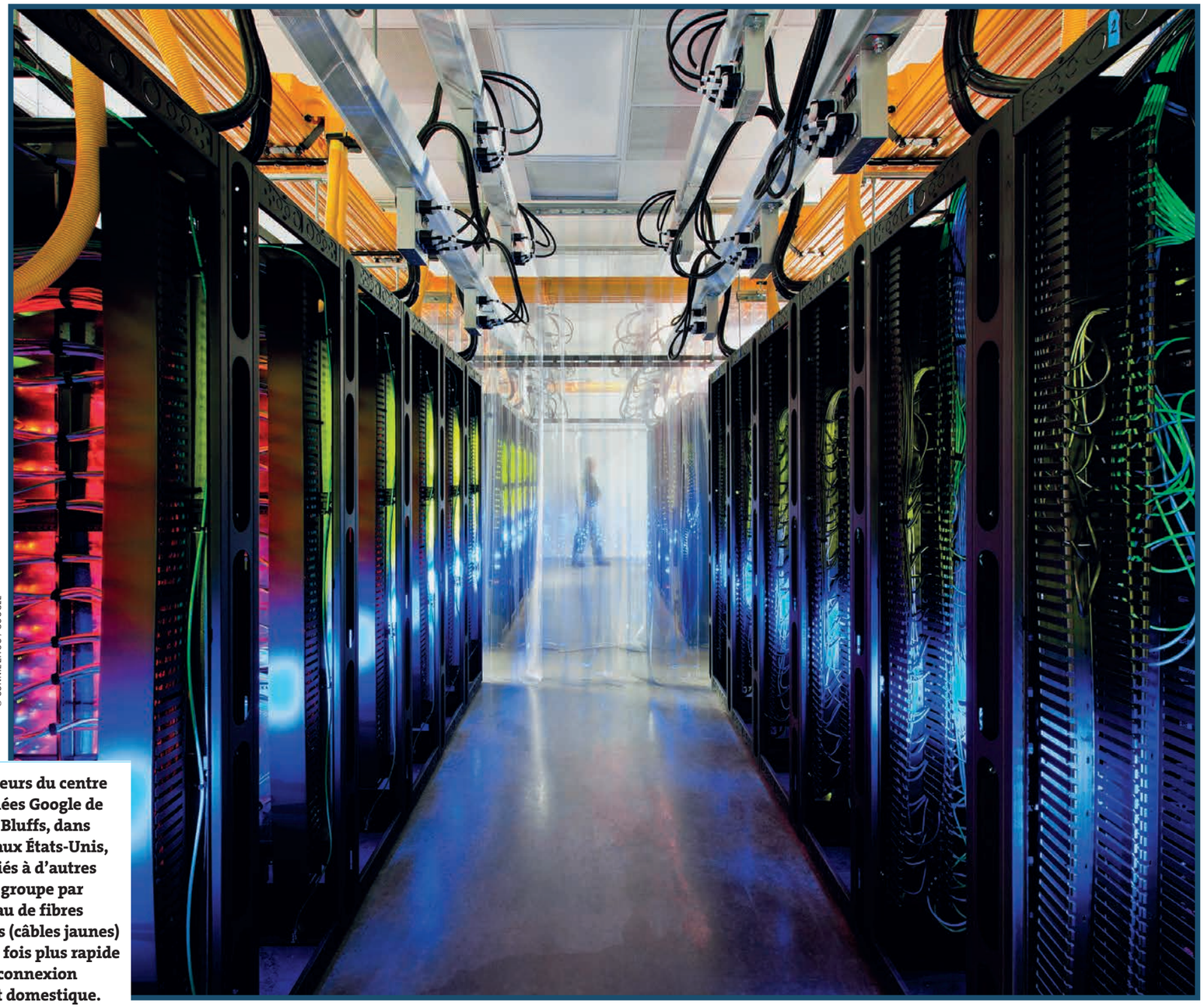
Les promesses du Big Data

Dossier préparé par Luc Allemand



© CONNIE ZHOU / GOOGLE

Ces serveurs du centre de données Google de Council Bluffs, dans l'Iowa, aux États-Unis, sont reliés à d'autres sites du groupe par un réseau de fibres optiques (câbles jaunes) 200 000 fois plus rapide qu'une connexion internet domestique.

En 2013, l'humanité a stocké plus de 2 000 milliards de gigaoctets de données numériques nouvelles. Et les trois quarts de ces données ont été créées par les consommateurs que nous sommes. Mises en réseau et exploitées par des ordinateurs, ces Big Data contiennent la promesse de services nouveaux qui amélioreront nos vies : éviter les embouteillages, adapter l'enseignement à chaque élève, personnaliser nos traitements médicaux, etc. Mais l'ère des Big Data porte aussi le risque d'une surveillance permanente. Connaître les possibilités et les limites des technologies est indispensable pour en réglementer l'usage sans en entraver les développements utiles, et pour que chacun, individuellement, puisse mieux les maîtriser.

1 • VINCENT BLONDEL : « Nous étudions de nouveaux objets scientifiques »

Entretien L'accroissement rapide du volume des données numériques enregistrées promet une compréhension inédite des comportements sociaux. Mais il nécessite de nouvelles méthodes d'analyse.

Vincent Blondel
est professeur à l'université catholique de Louvain, en Belgique, et professeur invité au Massachusetts Institute of Technology, aux États-Unis.

LA RECHERCHE : Comment définissez-vous les Big Data?

VINCENT BLONDEL : Le premier critère est le volume, sous-entendu par le mot « big ». Le domaine des Big Data s'intéresse à des ensembles de données digitales qui, de par leur taille, ne peuvent être traitées avec des méthodes traditionnelles ; en fonction des applications, ce peut-être de l'ordre du gigabit*, du térabit* ou plus encore. Ensuite, ce volume ne cesse de croître à grande vitesse. On estime que le volume de données stockées dans le monde double tous les quatre ans. On a ainsi stocké plus de données depuis 2010 qu'on ne l'avait fait depuis les débuts de l'humanité ! Le troisième critère, c'est la grande diversité des données auxquelles on s'intéresse. Ce peut être la consommation d'électricité dans tous les quartiers de France à tout moment, les 30 milliards de « j'aime » journaliers sur Facebook ou les 5 000 photographies déposées chaque minute sur le site de partage Flickr [1]. Enfin, on s'attache à la « véracité » : les données recueillies sont souvent bruitées et imprécises et doivent être traitées pour en extraire de l'information utile.

*LE BIT est une unité de mesure en informatique désignant la quantité élémentaire d'information ; 1 gigabit = 10^9 bits 1 térabit = 10^{12} bits.

L'essentiel

- LES BIG DATA sont caractérisées par le volume, la vitesse d'accumulation, la variété et la véracité des données numériques.
- IL FAUT METTRE au point des méthodes de traitement dont le nombre d'opérations n'augmente pas trop vite avec la quantité de données.
- CE DOMAINE fournit aussi de nouveau outils pour faire de la science, notamment des sciences sociales.

En quoi ces traitements diffèrent-ils de ceux que l'on réalise déjà en informatique ?

V.B. Ce qui nous a fait entrer dans l'ère des Big Data, c'est l'explosion des capacités de stockage. Un petit disque dur de l'épaisseur d'un livre suffit par exemple à stocker les informations sur les communications téléphoniques belges d'une année. Et cela pour un prix très modeste. Il devient alors possible de s'interroger sur la façon de traiter ces données afin d'en tirer des informations utiles : cette capacité de stockage crée de nouveaux objets à étudier, et il nous faut imaginer comment le faire. Même si nous arrêtions aujourd'hui de recueillir des données, nous aurions besoin de plusieurs années de travail pour comprendre comment analyser tout ce que nous avons déjà enregistré. Mais les données continuent d'arriver, toujours plus vite !

Donc vous recherchez de nouvelles méthodes d'analyse ?

V.B. Exactement. Prenons l'exemple d'un réseau dans lequel des entités sont connectées les unes aux autres. Un problème classique et très général consiste à rechercher des « communautés » : des zones plus densément connectées que d'autres. C'est un problème bien défini mathématiquement, et nous avions depuis longtemps des méthodes pour le résoudre. Mais elles n'étaient pas assez efficaces : il aurait fallu des années pour traiter les énormes réseaux d'aujourd'hui, formés par les utilisateurs de Facebook, qui sont 1 milliard, ou les pages web reliées par des hyperliens, que l'on compte par dizaines de milliards. Désormais, de nouvelles méthodes permettent de résoudre rapidement ces problèmes à l'aide d'un simple ordinateur de bureau. L'efficacité, c'est-à-dire la vitesse de traitement, est aussi l'un des obstacles à surmonter quand il s'agit de détecter des corrélations dans des ensembles très grands ou d'identifier des événements anormaux dans des séries.


© VINCENT BLONDEL

Comment rendre ces méthodes plus efficaces ?

V.B. Il faut que le nombre d'opérations à réaliser, donc le temps nécessaire, n'augmente pas trop vite quand le volume des données s'amplifie [fig. 1]. Pour la détection de communautés, par exemple, ce nombre d'opérations croissait comme le carré de la quantité de données : pour un réseau 10 fois plus gros, il fallait 100 fois plus de temps. Par exemple, imaginons qu'une heure de calcul suffise pour mener une analyse sur les communications téléphoniques d'une seule journée à Bruxelles. Pour traiter les communications de toute la Belgique, cela prendra 100 heures. Et pour les communications de toute l'Europe, il faudra environ 250 000 heures, soit plus de vingt-huit ans. Ce n'est pas possible. Nous devons donc trouver des méthodes dont le temps de calcul croît moins vite avec la taille des données. Linéairement par exemple : le temps augmente seulement proportionnellement à la quantité de données. C'est le minimum si l'on veut lire toutes les données.

Peut-on néanmoins faire mieux ?

V.B. Oui, nous savons aujourd'hui analyser un ensemble de données sans les consulter toutes, en donnant néanmoins des garanties sur la fiabilité de la réponse. Voilà une problématique scientifique récente et typiquement Big Data. Par exemple, il y a quotidiennement des milliards de transactions avec des cartes de crédit. Un algorithme qui n'en analyse que 10 ou 100 millions pourra tout de même indiquer qu'aucune carte n'a eu un parcours correspondant à une usurpation d'identité. La réponse ne sera pas garantie à 100 %, parce que des comportements anormaux pourraient exister dans les données qui n'ont pas été analysées. Mais la probabilité qu'elle soit vraie sera quantifiée rigoureusement.

Cela ressemble à des sondages ?

V.B. En quelque sorte, mais ce type de méthode permet de répondre à des questions plus complexes que celles posées lors de sondages d'opinion. Par exemple, l'évolution au cours du temps des communautés qui structurent un réseau, ou la détermination qu'une entité a eu un parcours différent des autres. La théorie nous permet de déterminer la distribution de probabilités suivant laquelle il faut choisir les données à analyser pour optimiser la précision de la réponse. Elle nous permet aussi de donner des bornes mathématiques pour l'écart entre cette réponse et celle que l'on aurait obtenue si l'on avait examiné toutes les données. Bien entendu, tout cela repose sur des hypothèses en lien avec la structure de l'ensemble de données et dépend du problème particulier que l'on souhaite résoudre.

Deux auteurs ont affirmé récemment que les Big Data sont porteuses d'une révolution scientifique comparable à celle entraînée par l'invention du microscope [2]. Qu'en pensez-vous?

V.B. Les Big Data permettent effectivement de faire de la science de façon totalement >>>

« Nous étudions de nouveaux objets scientifiques »

Entretien avec Vincent Blondel

nouvelle, notamment pour l'étude de phénomènes sociaux. Par exemple, avec Samuel Martin et Corentin Vande Kerckhove, dans mon laboratoire, nous travaillons en psychologie sociale sur les dynamiques d'opinion : comment, dans un groupe, des personnes qui doivent faire un choix s'influencent-elles mutuellement ? Des modèles mathématiques ont été proposés, mais il faut les tester. Autrefois, nous aurions mené les expériences avec quelques dizaines de personnes. Aujourd'hui, grâce au « Turc mécanique » de la société Amazon, nous pouvons très simplement recruter plusieurs milliers de participants qui réaliseront l'expérience de chez eux, en échange d'une somme modique [3]. Nos résultats auront une autre portée !

Un autre exemple est lié au développement des « cours en ligne ouverts et massifs », les MOOC, selon l'acronyme anglais. Des universités proposent des cours en accès gratuit sur Internet. À l'université catholique de Louvain, nous travaillons en partenariat avec la plateforme internationale edX, fondée par l'université Harvard et le Massachusetts Institute of Technology, aux États-Unis, dont l'interface enregistre l'ensemble du parcours de formation de l'étudiant : à quels moments il se connecte, combien de temps il reste sur chaque page, son taux de réussite aux tests qui lui sont régulièrement proposés, éventuellement les questions qu'il pose au sein des forums mis en place, etc. Cela permet de faire des observations et des expérimentations pédagogiques à une échelle inaccessible jusqu'ici. En corrélant le comportement des étudiants avec leurs résultats et leur progression, on pourra comprendre les processus d'apprentissage mieux qu'on ne l'a jamais fait, déterminer si certains sont plus efficaces que d'autres et offrir un parcours personnalisé.

Finalement, ne renoncez-vous pas à établir des lois scientifiques explicatives au profit de simples corrélations, que seul l'ordinateur maîtrise ?

V.B. Si un algorithme peut vous dire avant votre médecin, et sans que l'on comprenne totalement pourquoi, que vous avez une probabilité élevée d'avoir un cancer, je ne vois pas pourquoi on s'en priverait. Ensuite, les Big Data touchent les sciences sociales dans lesquelles les chaînes causales d'explication sont moins claires qu'en physique ou en biologie. Enfin, les analyses de Big Data sont des outils qui ne se substituent pas à la compréhension des scientifiques : elles attirent l'attention sur des corrélations détectées afin que ces derniers recherchent ensuite des explications causales. Bien entendu, pour les entreprises qui s'intéressent seulement aux applications, pour mieux vendre leurs produits par exemple, les modèles explicatifs ne sont pas nécessaires. En science par contre, les Big Data peuvent bien être vues comme un outil, à l'image d'un microscope, pour faire progresser la connaissance

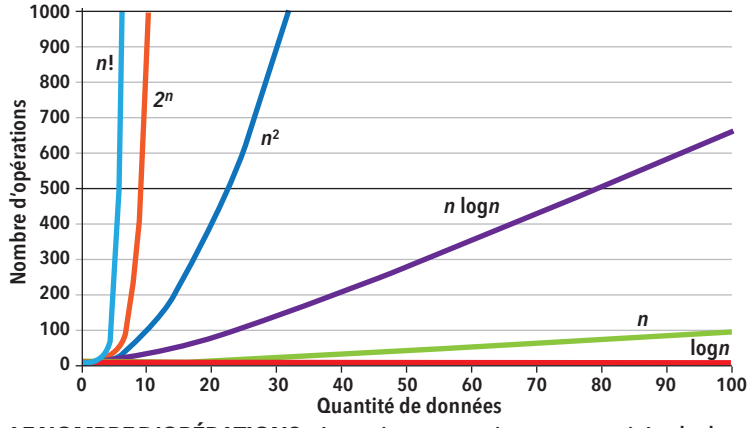
Y a-t-il une question sur laquelle les Big Data rencontrent des difficultés sérieuses aujourd'hui ?

V.B. La protection de la vie privée. Les Big Data promettent des bénéfices énormes pour la société, en faisant progresser la médecine personnalisée, la prédiction de la propagation de virus ou les modèles de croissance économique. Mais comme ces données sont souvent issues des comportements de chacun d'entre nous, il y a des risques d'intrusion, ce qui suscite des craintes. Il est de la responsabilité des scientifiques de contribuer à ces problématiques et d'aider les citoyens et les législateurs qui s'interrogent sur les limites à mettre, comme le font en ce moment ceux de l'Union européenne. Trouver la juste mesure ne doit pas être seulement du ressort de juristes ou de techniciens mais bien de toute la société. Ceux qui élaborent des modèles d'utilisation des données doivent aussi montrer scientifiquement l'intérêt de le faire, les difficultés qui se présentent lorsqu'on veut rendre des données anonymes et quantifier les dangers auxquels on s'expose en partageant des données.

Propos recueillis par Luc Allemand

[1] www.flickr.fr
[2] V. Mayer-Schöneberger et K. Cukier, *Big Data*, John Murray, 2013.
[3] www.mturk.com

Fig.1 La complexité des algorithmes



LE NOMBRE D'OPÉRATIONS nécessaires pour traiter une quantité *n* de données ne doit pas augmenter trop vite avec *n*. Des algorithmes acceptables pour peu de données (avec une variation en n^2 par exemple) prennent trop de temps dans le domaine des Big Data. Des méthodes d'échantillonnage, où toutes les données ne sont pas lues, permettent une variation plus faible que *n*.

Pour la vie sur Mars, on ne sait pas encore. Pour les cinq vies du papier, c'est sûr.

La force de tous les papiers, c'est de pouvoir être recyclés
au moins cinq fois en papier. Cela dépend de chacun de nous.
www.recyclons-les-papiers.fr

Tous les papiers ont droit à plusieurs vies.
Trions mieux, pour recycler plus !

2 • Un réseau d'autobus redessiné grâce au téléphone mobile

PAR **Francesco Calabrese,**
qui dirige l'équipe Smarter Urban Dynamics du centre de recherche de la société IBM à Dublin, en Irlande.

L'analyse des caractéristiques temporelles et spatiales des appels de téléphones mobiles à Abidjan, en Côte d'Ivoire, inspire des modifications des transports urbains qui réduiraient les temps de trajets des usagers.

Les villes d'Afrique subsaharienne ont connu ces dernières décennies une détérioration de leurs systèmes de transport en commun. Les grandes sociétés de transport se sont révélées peu efficaces : lignes surchargées, temps de trajet long, billet onéreux. Ces insuffisances ont dans de nombreux endroits été en partie comblées par l'apparition de services de transporteurs privés utilisant de petits véhicules.

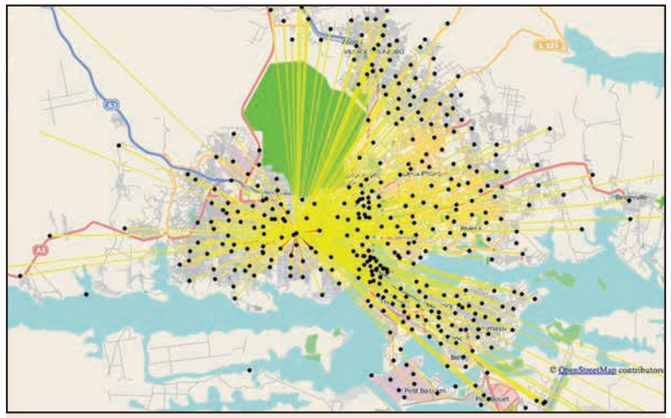
Ainsi, à Abidjan, en Côte d'Ivoire, où vivent 4,5 millions d'habitants, les 539 autobus de la compagnie publique Sotra sont complétés par environ 5 000 minibus et 11 000 taxis collectifs. Les conséquences sur la mobilité sont problématiques. Les minibus et autres formes de transport collectif représentent la moitié du trafic en passagers le long de certains axes. Des axes qui seraient desservis plus efficacement par des véhicules de plus grande capacité, mais trop coûteux pour de petites compagnies privées. L'absence de normes et de contrôles est également préjudiciable à la sécurité – les villes africaines ayant globalement un taux de décès par accident élevé – et à l'environnement [1].

À Abidjan, la compagnie publique Sotra exploite 539 autobus pour le transport des 4,5 millions d'habitants. Ils sont suppléés par quelque 16 000 minibus et taxis collectifs. © KAMBOU SIA/AFP

Quantifier la mobilité. Pour mieux répondre à la demande de mobilité des habitants, en mettant en place un système de transport public efficace et conçu dans une perspective durable, il faut d'abord quantifier celle-ci. Comment y parvenir ? Les méthodes classiques, fondées sur des enquêtes avec des questionnaires auprès des utilisateurs, ont été utilisées de manière très limitée, en raison de leur coût élevé. Et en Côte d'Ivoire, comme souvent dans les pays en développement, les infrastructures numériques (senseurs sur les routes ou GPS dans les véhicules) sont rares.

En revanche, le téléphone mobile, lui, est très répandu. Par exemple, en Côte d'Ivoire, 70 % des habitants en possèdent un. C'est pourquoi les données personnelles des utilisateurs peuvent y jouer un rôle si important. En effet, les communications à l'aide des téléphones mobiles offrent à des villes où l'urbanisation est rapide la possibilité de suivre la mobilité des habitants et d'estimer précisément les besoins en transport.

Si cette nouvelle méthode pour évaluer la mobilité est efficace, c'est parce qu'on a affaire à des données massives : le fort taux de pénétration du téléphone mobile fournit un échantillonnage de plusieurs ordres de grandeur supérieur à celui obtenu par les enquêtes à base de questionnaires. On englobe ainsi plus de catégories d'utilisateurs et celles-ci sont représentées proportionnellement à leur importance statistique. Ainsi, exploiter ces données massives permet de diminuer le biais statistique des enquêtes. De surcroît, les



© OPENSTREETMAP

Les flux moyens de déplacements entre les paires d'antennes-relais de la ville entre 7 et 16 heures sont représentés par les traits jaunes. Ils ont été calculés à partir des appels passés par 500 000 téléphones pour une période de cinq mois.

données sur les communications mobiles peuvent être obtenues en temps réel, autorisant un suivi dynamique du besoin de transport avec, à terme, l'idée de rendre ces services plus réactifs.

Avec mes collègues Michele Berlingerio, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli et Marco Luca Sobio, nous avons voulu vérifier ce que pouvait apporter en pratique ce type d'approche. Nous nous sommes concentrés sur la ville d'Abidjan, pour laquelle l'opérateur de téléphonie Orange avait fourni des données d'appels dans le cadre d'un concours baptisé *Data for Development* (« données pour le développement ») [2]. La base de données globale pour la Côte d'Ivoire contient 2,5 milliards d'enregistrements – appels et SMS – échangés entre 5 millions d'utilisateurs. Chaque enregistrement comprend un identifiant anonymisé, l'heure à laquelle l'appel a été passé ou reçu (ou le >>>

L'essentiel

> L'ABONDANCE des téléphones mobiles en Afrique subsaharienne en fait un outil précieux pour recueillir des données sur la population.

> L'ANALYSE de ces données aide en particulier à mieux comprendre le besoin de transport dans les zones urbaines.

> POUR LA VILLE D'ABIDJAN, les informaticiens ont ainsi proposé l'ajout de lignes de bus qui réduiraient le temps de trajet moyen des usagers.

Un réseau d'autobus redessiné grâce au téléphone mobile

>>> SMS envoyé), et l'identifiant de l'antenne-relais connectée au portable au début de l'appel. Bien que la localisation précise des utilisateurs n'ait pas été fournie dans l'enregistrement, nous avons pu la déduire avec une précision de 500 mètres (en zone urbaine), en supposant que les utilisateurs se trouvaient dans la zone de couverture de l'antenne-relais (la « cellule ») lorsque l'appel était passé.

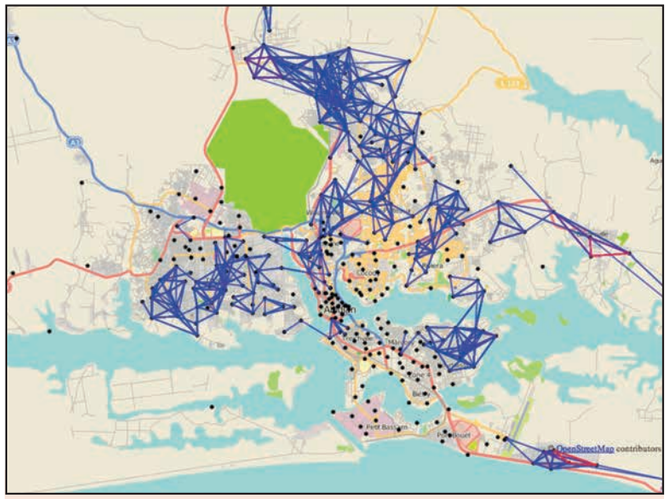
Matrice origine/destination. L'étude a porté sur les enregistrements provenant de 500 000 téléphones pour des appels passés en 2012 sur une période de cinq mois (dix groupes de 50 000 utilisateurs choisis au hasard toutes les deux semaines). À partir de ces enregistrements, nous avons extrapolé les mouvements individuels entre deux appels consécutifs passés par le même téléphone et utilisant deux cellules voisines. La connaissance des mouvements individuels entre les antennes-relais permet de construire la matrice origine/destination. Cette matrice décrit le nombre de personnes qui voyagent de n'importe quel point d'origine à n'importe quelle destination de la ville (carte p. 33). Plus précisément, elle représente le flot de personnes entre chaque paire d'antennes origine/destination durant un intervalle de temps.

La matrice origine/destination donne un premier élément d'analyse de la mobilité d'une ville. Avec cet aperçu de la demande de transport de la population d'Abidjan, nous avons pu évaluer quantitativement l'adéquation entre le système actuel de transport urbain et la demande. À l'aide d'un calculateur d'itinéraire que nous avons mis au point, nous avons associé à chaque élément de la matrice – chaque paire origine/destination – le trajet en bus le plus probable qui serait utilisé pour se déplacer d'un endroit à un autre. Avec ce procédé, on estime la fréquentation attendue de chaque ligne de bus, si tout le monde décidait d'utiliser les transports en commun.

Ce procédé permet en outre d'évaluer le temps de trajet associé à chaque déplacement et le temps d'attente prévu de ces personnes à leur point de correspondance. De la sorte, nous avons mis en évidence des lacunes dans le système de transport public actuel. En effet, nous avons trouvé des paires origine/destination pour lesquelles la durée de voyage par le système de bus était bien supérieure à celle à laquelle on pourrait s'attendre compte tenu de la distance à parcourir.

Ces carences nous ont poussés à étudier comment ajouter de nouvelles lignes de bus en vue d'améliorer l'efficacité du système. Nous nous sommes replongés dans les données d'appels des téléphones afin d'en extraire les motifs de déplacement les plus fréquents, comme des séquences d'antennes-relais utilisées plus souvent que d'autres (carte ci-contre). Bon nombre de ces habitudes de déplacement correspondaient à des lignes actuelles du réseau abidjanais. Mais d'autres semblaient couvrir des zones de la ville où le transport public officiel n'est pas disponible : on n'y trouve que des minibus ou des taxis.

Ajouter des lignes. À partir de toutes ces informations, nous avons élaboré un modèle d'optimisation qui tentait d'évaluer quelles lignes ajouter au système existant pour maximiser le niveau de service. Plus précisément, compte tenu du réseau présent, de la matrice origine/destination, des motifs de déplacement les plus fréquents, de l'estimation des temps de trajets dans le réseau et des ressources budgétaires (en termes de taille des véhicules), nous avons cherché à déterminer un ensemble de nouvelles lignes et de fréquences pour ces lignes de manière à minimiser les temps de trajets à travers la ville. Bien que le problème ne soit pas linéaire*, il peut être approché par un problème de programmation linéaire de grande dimension.



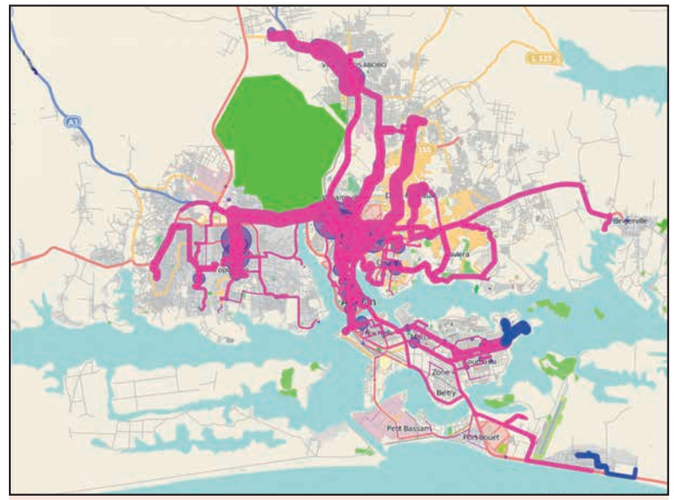
© OPENSTREETMAP

Les séquences d'antennes-relais le plus souvent utilisées (en bleu) reflètent les motifs de déplacement les plus fréquents. Ceux-ci sont analysés afin d'optimiser le réseau d'autobus dans les zones insuffisamment desservies.

* **UN PROBLÈME LINÉAIRE** se résout en un nombre d'opérations proportionnel à la quantité de données traitées.

Après avoir tourné sur nos machines, l'algorithme d'optimisation a recommandé l'ajout de quatre nouvelles lignes dont la mise en œuvre réduirait en moyenne de 10 % les temps de déplacements à travers la ville (carte ci-contre). En rapprochant le système de transport public des zones où les gens souhaitent l'utiliser, ces nouvelles lignes amélioreraient aussi la fréquentation. Avec ces lignes supplémentaires, l'impact sur les 22 lignes existantes serait également positif. Le temps de trajet moyen sur certaines lignes déjà existantes se réduirait, car l'ajout de nouvelles lignes détournerait une partie du flot de citadins voyageant habituellement sur les anciennes lignes [3]. Nous n'avons pas évalué l'impact de cette optimisation sur les minibus et les taxis collectifs, mais il est vraisemblable que leur utilisation diminuerait avec la disponibilité de meilleurs services de transport public dans la ville.

Ce projet illustre bien comment des données massives issues des appels par les téléphones portables sont exploitables pour mieux comprendre la demande de transport. À condition de préserver l'anonymat, ce qui restera un enjeu crucial pour l'avenir, des données analysées en temps réel ou presque permettraient de mesurer la qualité du service fourni aux habitants et d'ajuster dynamiquement les différents réseaux de transport aux besoins : des tarifs variables selon la fréquentation, des lignes qui s'ajustent dynamiquement, ou encore des mesures d'incitation destinées à encourager le changement de moyen de transport pour éviter les congestions de trafic. Des analyses de données qui deviendront plus fréquentes et qui pourront s'appliquer partout à travers le monde. ■



© OPENSTREETMAP

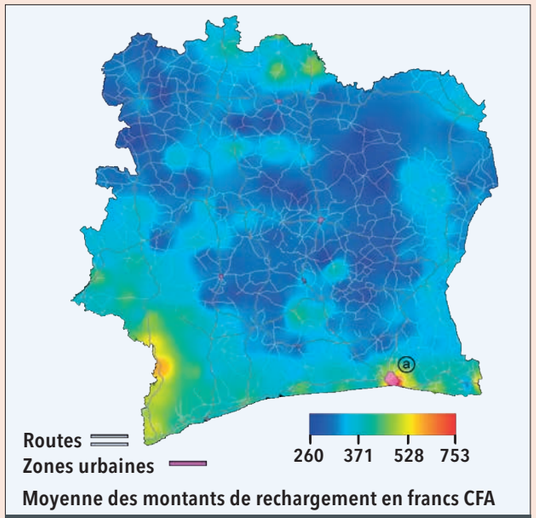
La fréquentation des lignes de bus existantes (en rose) est calculée à partir des flux de déplacements (carte p. 33). En prenant aussi en compte les déplacements les plus fréquents, de nouvelles lignes (en bleu) peuvent être proposées.

[1] K. Gwilliam, *Africa's Transport Infrastructure*, World Bank Publications, 2011.
[2] www.d4d.orange.com
[3] M. Berlingerio *et al.*, *Machine Learning and Knowledge Discovery in Databases, 663*, 2013.

Téléphone et niveau de vie

Les données issues de téléphones mobiles permettent de cartographier la richesse d'un pays dans lequel il n'existe pas de système fiable de collecte régulière de données socio-économiques. Thoralf Gutierrez, de l'université catholique de Louvain, en Belgique, et ses collègues ont ainsi analysé les comportements d'achat de crédit pour les communications mobiles en Côte d'Ivoire : beaucoup d'utilisateurs n'ont pas d'abonnement et rechargent leur compte téléphonique d'une somme qu'ils choisissent à chaque fois. Ces données anonymisées datant de 2012 provenaient d'un opérateur important du pays. Supposant que la taille et la fréquence des achats sont corrélées au niveau de richesse, ils ont utilisé la moyenne des sommes pour établir une carte de richesse du pays. Ils ont aussi analysé le brassage social à partir de la variabilité locale des achats moyens. Ce type de données pourrait servir à faire des prévisions et aider à prendre des décisions socio-économiques. ■ Philippe Pajot

T. Gutierrez *et al.*, ArXiv:1309.4496v1, 2013.



© DR

LE MONTANT MOYEN des achats de crédit téléphonique permet de cartographier le niveau de vie en Côte d'Ivoire. Les zones qui ressortent sont : Abidjan (a) et la frontière avec le Liberia (à gauche), lieu de divers échanges transfrontaliers.

3 • Les flux de données visualisés en temps réel

Les données arrivent désormais à grande vitesse de toutes parts et à flux continu. De nouvelles méthodes de visualisation nous permettent de les explorer et d'en extraire des informations.

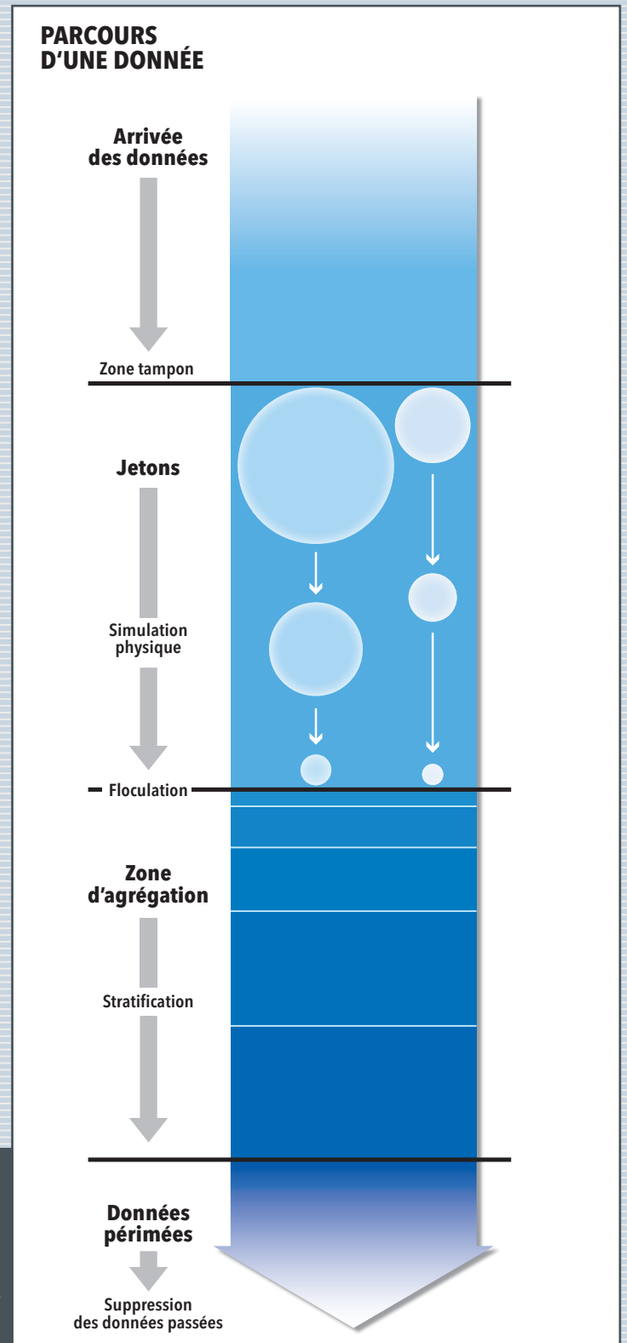
Mises à jours sur les réseaux sociaux, courriers électroniques, ventes ou productions industrielles : des flux de données sont produits en permanence. Dans nombre de cas, ces données peuvent être réparties en catégories. Mais elles apparaissent à des instants imprévisibles, s'accumulent jusqu'à ce qu'elles soient traitées, et doivent être conservées sous forme agrégée pour fournir une information historique et contextuelle. La méthode de « sédimentation visuelle », qui s'appuie sur l'analogie de la sédimentation géologique, a pour but de permettre la compréhension simultanée des différentes étapes [1].

[1] S. www.visualsedimentation.org

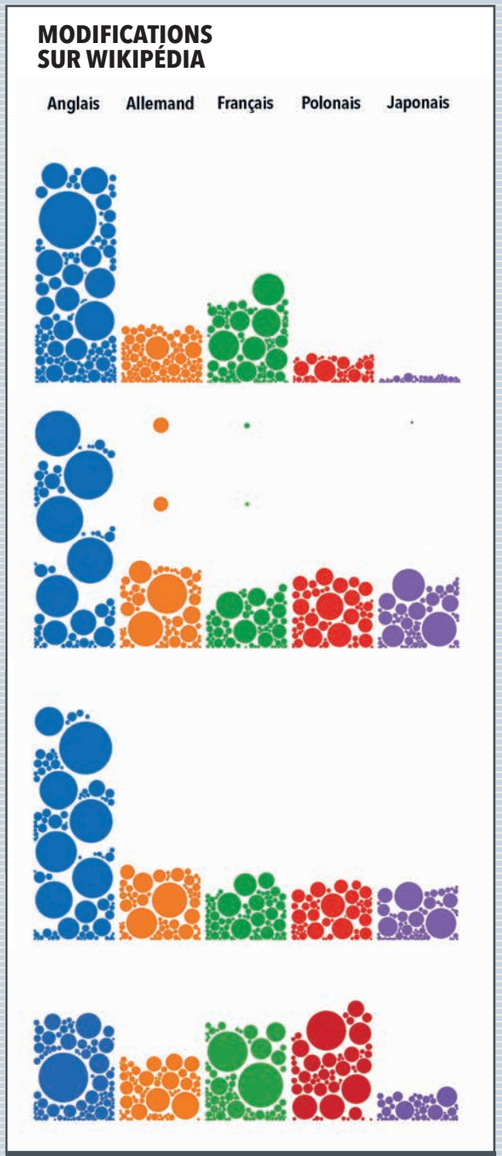
TEXTES ET IMAGES : **SAMUEL HURON**, INSTITUT DE RECHERCHE ET D'INNOVATION DU CENTRE POMPIDOU ET ÉQUIPE AVIZ D'INRIA, À SACLAY, **ROMAIN VUILLEMOT** ET **JEAN-DANIEL FEKETE**, ÉQUIPE AVIZ D'INRIA.

LA MÉTHODE

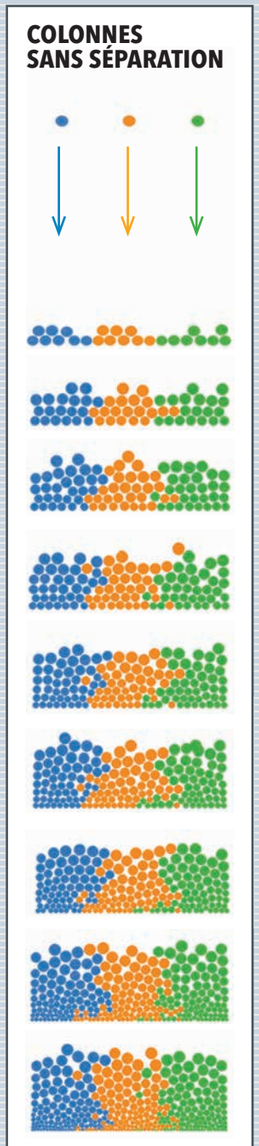
Au fur et à mesure qu'elles arrivent (en haut), les données sont accumulées dans une zone tampon. Chacune apparaît ensuite sous la forme d'un jeton, qui tombe en diminuant de diamètre, sous l'effet d'un modèle de simulation physique. Lorsqu'il arrive sur la zone d'agrégation, il se dépose par floculation, et s'incorpore à une strate. Les strates peuvent avoir différentes couleurs selon leur ancienneté, et même être progressivement supprimées à mesure qu'elles vieillissent.



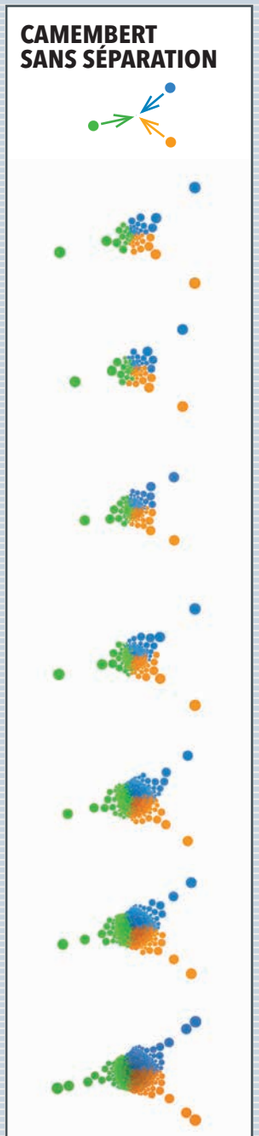
© INFOGRAPHIE BRUNO BOURGEOIS



Les modifications des articles de Wikipédia dans cinq langues (colonnes) sont suivies en temps réel : chaque modification est représentée par un jeton, dont la taille est proportionnelle à sa longueur. Ce jeton vient s'ajouter à la colonne, dont le contenu se tasse au cours du temps (de haut en bas), les images ne sont pas prises à des intervalles réguliers. En passant le pointeur sur un jeton, on obtient des informations sur la modification. En cliquant sur le jeton, on ouvre la page modifiée.



Les jetons peuvent s'étaler sur les côtés de chaque colonne. Cette variante, présentée ici avec des données synthétiques, conserve des informations sur les évolutions relatives des flux dans les interpénétrations successives des colonnes.



Un diagramme en camembert peut être produit sur le même principe : les jetons tombent vers le centre. Ils s'agrègent sur les différents secteurs dont la taille angulaire, variable en l'absence de séparation, traduit l'importance relative.

4 • Une vie privée est-elle encore possible ?

Chacun de nous, en utilisant son GPS ou son téléphone mobile, indique sa position. *A priori,* une donnée sans importance. Pourtant, recoupée avec une ou deux autres, elle permet de nous identifier parmi les utilisateurs.

PAR **Adeline Decuyper,**
doctorante à l'université catholique de Louvain, où **Vincent Blondel** est professeur de mathématiques appliquées.

Vous êtes en voiture, un dimanche soir, et vous vous dirigez vers une grande ville d'Europe. Le système de guidage GPS de votre véhicule vous indique : « Embouteillage dans 12 kilomètres, vitesse moyenne 20 kilomètres à l'heure. Voulez-vous être déviés ? » Bien heureux d'être informé à temps, vous changez d'itinéraire, gagnant ainsi un temps précieux.

Cette situation est bien réelle (elle est arrivée à l'un d'entre nous). La plupart des systèmes GPS commercialisés aujourd'hui ne se contentent pas, en effet, de calculer votre position à partir de signaux envoyés par des satellites. Toutes les trente secondes environ, ils indiquent celle-ci à un service central, qui l'enregistre anonymement. En combinant ces informations pour tous les véhicules équipés, un logiciel calcule l'état du trafic, dont il informe en retour les utilisateurs.

Services personnalisés. Habitués depuis quelques années aux services personnalisés, précis et disponibles en temps réel, nous utilisons tous les jours les résultats de l'analyse instantanée de nos données et de celles des autres, collectées en masse. Le prix à payer ? Partager ses données avec le gestionnaire de l'application et lui permettre de les utiliser. Mais quel degré d'anonymat peut-on conserver quand on enregistre la trajectoire, même approximative, d'un grand nombre de personnes ?

À en croire les opérateurs, vous n'avez pas à vous inquiéter pour votre vie privée : ni votre nom ni le numéro de votre appareil ne sont conservés avec vos données de localisation. Impossible à quiconque de les utiliser pour vous suivre à la trace. Malheureusement, ce n'est pas tout à fait vrai. L'étude menée par l'un d'entre nous (Vincent Blondel), avec Yves-Alexandre de Montjoye, César Hidalgo et Michel Verleysen, du MIT, aux États-Unis, et de l'université catholique de Louvain, en Belgique, montre que très peu d'informations de localisation du type de celles transmises par un GPS suffisent pour distinguer une trajectoire singulière et, partant, pour reconnaître une personne particulière [1].

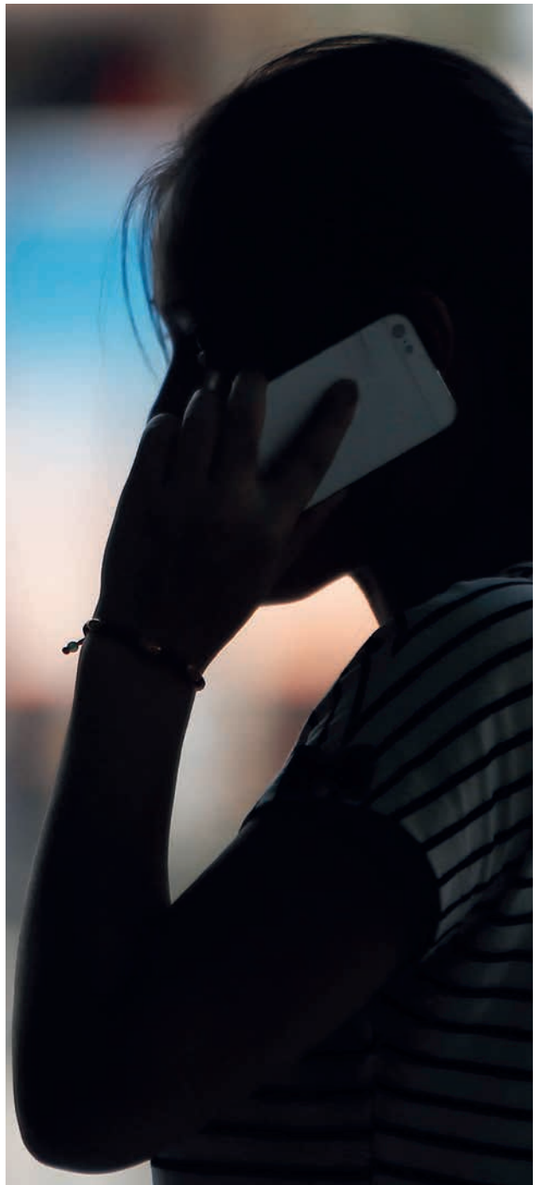
Les bases de données enregistrées par les opérateurs de téléphonie mobile, comme celle utilisée pour cette étude (un million et demi d'abonnés sur quinze mois), contiennent généralement, pour chaque utilisateur, l'endroit et l'heure auxquels chaque appel a été passé. En reprenant les antennes qui ont relayé les appels, on peut retracer un itinéraire approximatif pour chaque personne présente dans les données. Chaque point de cet itinéraire est caractérisé par une date et par une position géographique, en pratique la position de l'antenne-relais qui a retransmis le début de l'appel. Combien de points suffisent pour caractériser complètement une trajectoire particulière ?

La réponse, étonnamment, est quatre ! Si vous connaissez quatre des points de passage d'une personne, dans 95 % des cas, il n'y a qu'un utilisateur dont l'itinéraire passe par ces quatre points précis [fig.1]. Vous pouvez ainsi le retrouver dans la base de données et suivre le restant de son trajet, de ses appels et de son activité téléphonique.

Informations en chaîne. Pour mesurer l'information que donnent quatre points de passage, pour une résolution temporelle à l'heure près et une précision spatiale déterminée par la zone couverte par chaque antenne, on choisit une personne au hasard dans la base de données et parmi ses points de passage, on en sélectionne quatre au hasard. Ensuite, pour ces quatre points de passage précis, on cherche, parmi le million et demi d'utilisateurs, ceux qui ont une trajectoire compatible avec ces quatre points. ›››

L'essentiel

- **LA PLUPART DES DONNÉES de géolocalisation recueillies sont anonymes.**
- **TOUTEFOIS, QUATRE POINTS spatio-temporels suffisent dans la plupart des cas à individualiser une trajectoire, donc une personne, au sein d'une base de données.**
- **IL EST DONC PRÉFÉRABLE de ne partager sa localisation qu'avec des opérateurs de confiance, qui fournissent en échange des services dont nous avons réellement besoin.**

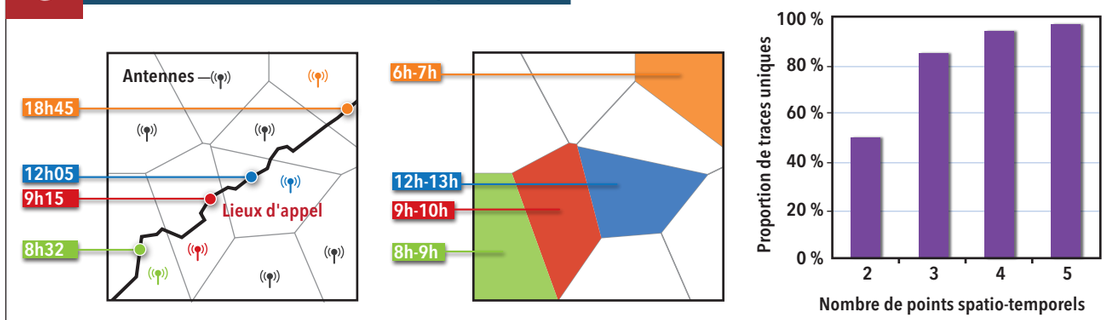


© ALY SONG / REUTERS

Quand on recoupe les bases de données

Que se passe-t-il si l'on associe les informations de plusieurs bases de données ? Le risque de voir sa vie privée exposée devient alors d'autant plus grand. C'est ce qui est arrivé pour NetFlix, entreprise américaine qui propose de regarder des films en streaming sur Internet. En 2006, la société a lancé un concours pour améliorer son système de recommandation et a rendu publique l'activité en ligne d'un demi-million d'utilisateurs, tous identifiés par des numéros de clients anonymes. Deux chercheurs de l'université du Texas ont cependant réussi à remettre un nom sur plusieurs de ces numéros, en comparant les informations de NetFlix et les avis sur les films disponibles sur le site de référence IMDb qui, eux, ne sont pas anonymes. Ils ont ainsi révélé jusqu'aux opinions politiques et orientations sexuelles de plusieurs utilisateurs. L'affaire a été portée devant la justice.

Fig.1 L'individualisation des trajectoires



© INFOGRAPHIE BRUNO BOURGEOIS

LA TRAJECTOIRE des appels téléphoniques d'un abonné (à gauche) est enregistrée dans une base de données sous forme de points spatio-temporels : la zone couverte par l'antenne qui a relayé l'appel est associée à l'heure de celui-ci, à une heure près (au milieu). La moitié de ces trajectoires sont individualisées lorsque l'on en connaît deux points (à droite). Avec quatre points, on en individualise 95 %.

Une vie privée est-elle encore possible ?

© CONNIE ZHOU / GOOGLE

Les serveurs de données de Google (ici près d'Atlanta, aux États-Unis) conservent les historiques de recherches des utilisateurs connectés. Même rendus anonymes, une très faible partie de ces historiques permettrait d'identifier ceux qui les ont produits.

››› On répète alors l'expérience un grand nombre de fois : choix d'une personne, choix des points de passage, comptage du nombre de trajectoires compatibles. Et on observe alors que, parmi les 2 500 essais de ce type réalisés, 95 % des quadruplets ne correspondent qu'à un seul utilisateur. La connaissance des quatre points de passage suffit donc dans ce cas à retrouver une personne précise.

En résumé, cela veut dire que si vous savez qu'une personne a passé des appels depuis chez elle le matin vers 8 heures ; de son lieu de travail, dont vous connaissez la localisation géographique, vers 10 heures ; du supermarché où elle fait ses courses à 19 heures ; et à nouveau de son domicile vers 21 heures, vous pouvez la retrouver dans la base de données. Vous saurez ainsi où elle est allée à d'autres heures et les autres jours.

Une fois le numéro de client de la personne retrouvé dans les données, vous pouvez connaître toute son activité téléphonique. En répétant l'expérience avec les personnes qu'elle a appelées, vous pourriez savoir qui elle appelle régulièrement, quels sont les amis de ses amis, ainsi que retracer toutes leurs activités téléphoniques, et les lieux qu'ils fréquentent régulièrement. Toutes ces informations peuvent être déduites de l'activité téléphonique d'une personne : elles sont enregistrées tous les jours par votre opérateur téléphonique, à chaque appel que vous passez ou que vous recevez.

Identification spatio-temporelle. Quatre points, cela paraît vraiment peu pour être unique. Cette valeur dépend bien sûr de la précision en temps et en espace des données dont on dispose. Les données utilisées dans l'étude contenaient pour chaque activité téléphonique, la date de l'appel, à l'heure près, ainsi que l'antenne d'où partait l'appel : par exemple, l'utilisateur A a appelé B entre 9 et 10 heures le 8 octobre 2013, depuis l'antenne Z. La précision spatiale dépend donc du nombre d'antennes autour de Z : elle est meilleure en zones urbaines qu'en zones rurales.

Protégerait-on mieux l'anonymat des abonnés au téléphone en modifiant la résolution spatio-temporelle des données ? Comment l'unicité des trajectoires varie-t-elle si, par exemple, au lieu de savoir qu'un appel a été passé entre 9 et 10 heures, dans un rayon de 1 kilomètre carré, on sait seulement qu'un appel a eu lieu entre 8 et 13 heures, dans un rayon de 5 kilomètres carrés ? De façon assez surprenante, le nombre d'observations nécessaires pour rendre une personne unique n'augmente que très peu. Si on sait quand a eu lieu l'appel à cinq heures près, et que l'on regroupe ensemble les zones couvertes par cinq antennes à la fois, quatre points d'observation suffisent encore à retrouver une personne unique dans plus de la moitié des cas [fig.2].

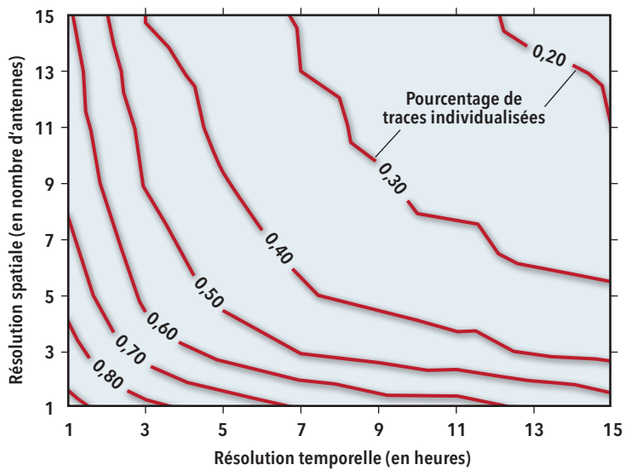
Peut-on généraliser ces résultats à une population plus dense ou à une zone géographique plus grande ? Cette étude ne se fondait que sur l'analyse d'une base de données spécifique, mais si la densité de population augmente, le nombre d'antennes augmente aussi, car les antennes ont une capacité limitée. Donc, même s'il y avait plus de clients dans la base de données, on s'attendrait à ce que l'augmentation de la précision géographique qui l'accompagnerait rende les résultats similaires à ceux observés dans les données analysées ici. Dans le cas où les données disponibles couvriraient une zone géographique plus vaste, les résultats ne changeraient pas beaucoup non plus, car la mobilité des utilisateurs est locale. En effet, si l'on regarde toutes les antennes utilisées par une même personne, on voit dans les données analysées ici que 94 % des utilisateurs ne se déplacent que dans un rayon de 100 kilomètres.

Contrats de confidentialité. Compte tenu des informations très personnelles que l'on retrouve dans les données de mobilité des personnes, le souci de préserver l'anonymat s'est assez vite installé dans la société. Les propriétaires de grandes bases de données hésitent avant tout partage, en prenant garde de ne pas s'exposer à des poursuites en justice ou aux critiques de leurs clients. Pour accéder aux données, les chercheurs doivent d'abord signer un contrat de confidentialité avec le fournisseur des données.

Les données sont donc toujours au préalable anonymisées : par exemple, au lieu d'un numéro de téléphone, on y retrouvera juste un numéro de client pour désigner une personne. C'est aussi la raison pour laquelle nous n'avons pas eu accès aux heures des appels à la seconde près, mais à l'heure près. Les sociétés tentent ainsi de se prémunir contre de potentiels scandales comme celui qui a touché NetFlix (lire « Quand on recoupe les bases de données », p. 39), et de protéger aussi la vie privée de leurs clients. Comme nous venons de le voir, cela risque toutefois de ne pas suffire.

Du côté du grand public, le partage et la diffusion des données personnelles des utilisateurs du Web sont devenus des sujets ›››

Fig.2 La dégradation de l'information



© INFOGRAPHIE BRUNO BOURGEOIS

LE POURCENTAGE de trajectoires individualisées par la connaissance de quatre points de celles-ci décroît assez lentement si l'on ne diminue la précision qu'en temps (en abscisse) ou qu'en espace (en ordonnée). Il est bien plus efficace de la diminuer pour les deux paramètres simultanément.

Yaniv Erlich : « Nous avons cassé l'anonymat de données génétiques »

Yaniv Erlich est généticien à l'Institut Whitehead, à Cambridge, aux États-Unis.

LA RECHERCHE : Votre équipe a identifié des donneurs d'ADN anonymes [1]. Comment avez-vous procédé ?

YANIV ERLICH : Nous sommes partis du génome de dix hommes stockés dans la base de données appelée Projet 1 000 Génomes et accessible gratuitement aux chercheurs dans le monde [2]. Grâce à un algorithme, nous en avons extrait des séquences d'ADN répétitives du chromosome Y que les hommes se transmettent de père en fils. Leur corrélation avec le nom de famille est si forte que des sites web de généalogie proposent de retrouver des gens de sa famille à partir de ces séquences. Grâce à deux de ces sites, nous avons découvert le nom de famille de cinq des dix hommes. Il s'agissait de noms de famille mormons dans l'Utah.
Puis nous avons consulté le site web du Coriell Cell Repository, organisme qui fournit aux chercheurs des prélèvements biologiques effectués sur les donneurs du Projet 1 000 génomes [3]. Ce site indique en particulier l'âge des donneurs au moment du prélèvement. C'est ce qui nous a permis de déterminer l'année de naissance des cinq hommes.

Comment avez-vous établi leur identité précise ?

Y.E. En continuant à surfer sur le Web comme n'importe quel internaute. Nous avons cherché des hommes portant ces noms de familles, ayant ces âges et résidant dans l'Utah au moment des prélèvements. Nous avons consulté des moteurs de recherche, des annuaires, des sites généalogiques, des archives nécrologiques, des sites fournissant des données démographiques publiques… et nous avons démasqué ces cinq hommes en moins de sept heures. Nous avons aussi identifié des membres de leurs familles, soit cinquante personnes au total.

Peut-on remédier à une telle faille ?

Y.E. Bien sûr, nous n'avons pas publié l'identité de ces personnes et nous avons tout de suite informé les autorités. Depuis, le Coriell Cell Repository n'indique plus l'âge des donneurs. Mais avec la multiplication des bases de données génétiques, l'anonymat des donneurs va devenir de plus en plus difficile à garantir. Toutefois, je suis contre la restriction d'accès à ces bases, car cela nuirait au progrès scientifique. En revanche, je pense qu'il faut mieux informer les donneurs, en leur expliquant clairement le risque que leur identité soit un jour découverte. Il faudrait peut-être aussi prévoir des sanctions significatives en cas de mauvais emploi de ces données génétiques.

■ **Propos recueillis par Jean-Philippe Braly**

[1] M. Gymrek *et al.*, *Science*, *339*, 321, 2013.
[2] www.1000genomes.org
[3] ccr.coriell.org

La sécurité sociale anglaise fiche les patients

« Cher docteur, je vous écris pour vous informer que je refuse mon autorisation pour que les informations identifiables me concernant soient transférées hors de votre cabinet pour tout autre but que me prodiguer des soins. » **C'est en ces termes que les promoteurs de la campagne MedConfidential proposent aux citoyens anglais d'écrire à leur médecin traitant [1]. La réforme du système de santé anglais, entrée en vigueur en avril 2013, prévoit en effet, par défaut, la récupération centralisée de toutes les données concernant les patients enregistrées dans les fichiers des médecins. L'objectif annoncé est d'abord une amélioration des soins, qui bénéficierait aux patients. Mais il est aussi prévu que ces données soient mises à la disposition de chercheurs universitaires et de compagnies privées. Ce qui a fait naître de sérieuses craintes sur la préservation de l'anonymat de ces informations si personnelles. ■ Luc Allemand**

[1] http://medconfidential.org

››› sensibles. Une application pour un téléphone intelligent ou une tablette semble utile et répondre à nos besoins, soit. Va-t-elle nous faciliter la vie ? Il suffit de cliquer sur le bouton « Accepter les termes et conditions d'utilisation », et de profiter du service. Un petit clic, c'est vite fait, souvent sans prendre le temps de lire le long texte qui décrit comment et dans quel but les données seront utilisées.

Géolocalisation fréquente. Plus de la moitié des applications gratuites demandent d'enregistrer la position géographique de l'utilisateur [2]. On remarque que beaucoup d'applications qui ne fournissent pas d'information en rapport avec la localisation l'enregistrent quand même : ceux qui l'ont produite se constituent ainsi une base de données des endroits d'où leur application est utilisée, sans toujours un but bien défini.

Que peut-on alors faire pour protéger la vie privée tout en profitant des nouveaux services développés sur la base des données partagées ? Interdire aux sociétés d'enregistrer et d'utiliser les données disponibles ne ferait qu'empêcher d'utiliser les services auxquels nous nous sommes rapidement habitués et qui se fondent sur les données de chacun pour améliorer leur efficacité.

Une solution récemment proposée par une équipe de l'université de Princeton et des laboratoires AT&T aux États-Unis serait de synthétiser des données artificielles, qui présenteraient les mêmes caractéristiques que les données authentiques, afin de pouvoir les rendre publiques sans danger d'intrusion dans la vie privée [3]. L'utilité de cette approche reste cependant limitée, puisque toute analyse faite sur les données de synthèse devrait être vérifiée sur les vraies données pour être validée. De plus, pour construire de telles données de synthèse, il faudrait déjà disposer de données authentiques pour en utiliser les caractéristiques.

Le plus efficace, afin de continuer à utiliser les services que la technologie actuelle permet tout en préservant sa vie privée est finalement sans doute de recommander à l'utilisateur d'applications et de réseaux sociaux de surfer consciencieusement, de bien personnaliser ses paramètres de confidentialité et de n'accepter de partager sa géolocalisation qu'avec des partenaires de confiance. Notre banque en sait souvent beaucoup sur nous : salaire, allocations familiales, cotisation à un syndicat, factures d'hôpital, etc. Nous fournissons ces informations de manière volontaire et en confiance. Il en va de même pour nos informations privées telles que notre géolocalisation. Nous ne devrions les partager qu'avec des partenaires en lesquels nous avons confiance. ■

[1] Y.-A. de Montjoye *et al.*, *Scientific Reports, 3,* 1376, 2013

[2] Appthority Quarterly Publication, App Reputation Report, Whitepaper, February 2013.

[3] S. Isaacman *et al.*, MobiSys '12, Proceedings of the 10th international conference on Mobile systems and services, 2012, p. 239.

Pour en savoir plus

Livres et articles

› V. Mayer-Schönberger et K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think,* John Murray, 2013.

› *Big Data Gets Personal, Technology Review, Business Report,* mai 2013.

› J. Stuart Ward et Adam Barker, ArXiv:1309.5821, 2013.

› Henri Verdier, Les Big Data de A à Z, http://tinyurl.com/paristechreview-etat-2012

› Jean-Pierre Malle, La triple rupture des Big Data, http://tinyurl.com/big-data-revolution-culturelle

Sur le Web

› **www.visualsedimentation.org** L'équipe Aviz D'Inria propose exemples et code source en accès libre pour la sédimentation visuelle.

› **www.d4d.orange.com** La société Orange a organisé entre juin 2012 et mai 2013 le challenge « Data for Development » qui a suscité les contributions de 80 équipes scientifiques.

› **www.futurict.eu** Le projet européen FuturICT vise le développement des technologies de l'information et de la communication au service de la compréhension de la société.

› **http://hd.media.mit.edu** Le Human Dynamics Laboratory du MIT consacre son activité à l'utilisation des Big Data pour la société.

› **http://fing.org/?-Infolab-** La Fondation Internet nouvelle génération mène la campagne Infolab pour sensibiliser entreprises et collectivités à l'utilisation des données.

› **http://dataveyes.com** La société Dataveyes présente et vend des services de visualisation de données.