Des données au réseau - Big data

laurent.beauguitte@cnrs.fr

Juillet 2017

De la trace numérique au réseau

Extraits de T. Venturini, D. Cardon et J.-P. Cointet, Présentation, *Réseaux*, 6/2014, 188, p. 9-21.

Maîtriser les données numériques demande alors de développer des méthodes raisonnées d'échantillonnage et d'interprétation de ces données. Les travaux conduits à partir d'extraction d'informations issues des blogs, de Google ou de Facebook permettent d'apprendre beaucoup sur les usages de ces plates-formes, sans doute moins sur les phénomènes généraux que, par généralisation, elles voudraient représenter. En revanche, une interrogation sur les articulations entre les enregistrements numériques et d'autres techniques d'objectivation des pratiques sociales est souvent très féconde, à l'instar des travaux étudiant la correspondance entre sociabilité en ligne et sociabilité « réelle » (par exemple : Burke et Kraut, 2014). De nombreux travaux usant de « méthodes digitales » couplent le flux de données extraites des API du web avec des dispositifs d'interrogation ad hoc (questionnaire, échantillonnage, tirage aléatoire, etc.) permettant de contrôler la représentativité des flux capturés. Dans sa thèse sur la formation des opinions politiques sur Twitter, Julien Boyadjian (2014), par exemple, a procédé à un contrôle par questionnaire des propriétés sociales des utilisateurs sélectionnés afin de contrôler son échantillon.

[...]

Interroger les conditions de production des données numériques permet de résister à la tentation de les naturaliser. Comme toute source secondaire, la construction des données est toujours le résultat d'une longue chaîne d'actions dont certains maillons échappent au contrôle direct de l'expérimentateur. Aussi importe-t-il que l'expérimentateur et ses pairs puissent remonter tous les passages de la chaîne et vérifier leur solidité (Latour, 1993). Ce n'est pas toujours complètement le cas avec les données extraites des plates-formes numériques dont l'accès est parfois livré aux desiderata des propriétaires des

sites ou soumis aux décisions d'algorithmes dont le principe de fonctionnement est obscur et secret (Pasquale, 2015). L'enquête en milieu digital suppose toujours une minutieuse investigation des conditions de production et de restitution des données numériques. Même lorsqu'il s'agit de traces mises à disposition par d'autres chercheurs, leur réutilisation reste problématique (Carlson et Anderson, 2007). En effet, dans un corpus numérique, la distinction entre bruit et information ne peut jamais être faite a priori. Elle dépend strictement des objectifs de recherche et du type d'analyse.

Extraits de É. Dagiral et O. Martin, 2017, Liens sociaux numériques. Pour une sociologie plus soucieuse des techniques, *Sociologie*, 8(1), p. 3-22.

p. 4 : [L]es techniques d'information et de communication (TIC) sont entrelacées dans les faits sociaux étudiés et [...] il serait vain de vouloir distinguer la dimension non-TIC et la dimension TIC d'un fait social. Choisir cette expression « liens sociaux numériques » est une manière de dire que les liens sociaux incorporent, ou peuvent incorporer, des aspects prenant forme dans les dispositifs numériques. C'est une manière de dire que la sociologie ne peut pas ignorer la place et le rôle des techniques dans les dynamiques du social.

p. 15 [...] si la capture et l'analyse de données issues d'internet (blogs, réseaux, volume de trafics...) nécessitent de nouveaux outils et de nouvelles compétences techniques pour les sociologues, les fondamentaux de la démarche sociologique restent les mêmes : le « métier de sociologue » n'est pas révolutionné par ces nouveaux objets de recherche.

p. 16 Au moins quatre constats peuvent être formulés à la lumière de ces cinq articles.

1. Premièrement, dans la mesure où certaines relations sociales et certaines pratiques sociales prennent appui ou prennent forme dans des dispositifs techniques, les sociologues ne peuvent évidemment pas ignorer ces dispositifs et les considérer comme négligeables, ni même secondaires.

2. Si le « monde en ligne » et le « monde hors-ligne » ne sont pas étrangers l'un à l'autre et que l'un fait souvent écho à l'autre, [p 17] il ne saurait pour autant être possible de les confondre - raison supplémentaire pour n'ignorer aucun des deux et chercher à les embrasser dans une démarche unique. [...]

3. Un troisième apport concerne la manière dont les articles de ce dossier permettent d'éclairer et de questionner la notion de « lien social ». [...]

4. [p. 18] 4. Enfin, notre dernière remarque issue de ces cinq textes concerne les moyens méthodologiques et empiriques nécessaires à la compréhension des liens et à leur analyse. Ces moyens sont pluriels, sans pour autant nécessiter une refonte complète des savoir-faire sociologiques. [p. 19] Par exemple, Margot Delon illustre très bien plusieurs principes méthodologiques, presque toujours indispensables à l'analyse des sources web : la nécessité de replacer le blog dans le contexte social et historique de son éla-

boration et de son évolution et la prise en compte des propos des concepteurs comme des utilisateurs et contributeurs du site.

Extraits de C. Prieur, A. Stoica et Z. Smoreda, 2009, Extraction de réseaux égo centrés dans un (très grand) réseau social, *Bulletin de Méthodologie Sociologique*, 101, p. 5-27

p. 6 La méthode que nous proposons repousse l'étape de simplification du réseau pour tenter de transposer sur de très grands réseaux des approches par réseaux égocentrés. L'élément clef consiste à calculer, pour chacun des nœuds, un ensemble de propriétés de son "réseau égocentré" (défini à partir de ses voisins directs et leurs voisins) pour guider le choix des nœuds sur lesquels on peut souhaiter réaliser une étude plus poussée, constituant ainsi des échantillons à géométrie variable.

p. 7 Les enregistrements concernent deux types de communications : les appels vocaux et les SMS. Pour les personnes appartenant à l'opérateur O, nous disposons aussi de plusieurs données supplémentaires : l'année de naissance, le sexe, la région et le type de contrat (fixe ou mobile). [...] nous avons défini un lien (non orienté) entre deux personnes par l'existence d'au moins un appel vocal entre les deux, dans les deux sens, pendant le mois que dure l'observation. Cette procédure élimine les appels dans un seul sens, des événements singuliers dans la plupart des cas suggérant que les deux individus en communication ne se connaissent pas personnellement. Le réseau que l'on obtient a approximativement 2 millions de nœuds et 3 millions de liens et possède des caractéristiques communes aux très grands réseaux. Pour n'en citer que quelques unes, une forte proportion des nœuds (87%) et des liens (97%) forment une grande composante connexe (c'est-à-dire qu'il existe un chemin qui en relie tous les nœuds).

p. 9 Plutôt que d'étudier de manière globale des propriétés de structure du grand réseau constitué à partir des appels entre deux millions de clients de l'opérateur téléphonique O, nous faisons le choix ici de caractériser les réseaux égocentrés de chacun des nœuds de notre réseau. Nous allons donc définir des indicateurs locaux qu'il sera possible de calculer sur les voisinages de tous les nœuds et dont nous pourrons étudier la distribution sur l'ensemble des individus, pour ensuite nous focaliser sur quelques individus.

p. 10 Pour chaque nœud du réseau, nous avons choisi de définir son réseau égocentré comme l'ensemble de ses voisins, les voisins de ses voisins et les liens entre tous ces nœuds. Autrement dit, on inclut dans le réseau égocentré de chaque nœud (qu'on appelle ego) tous les nœuds qui se trouvent à distance au plus deux d'ego.

p. 26 La méthode proposée ici montre cependant que l'analyse qualitative peut avoir une place non négligeable dans l'étude des très grands réseaux et que dans ce cadre, elle gagnera même en représentativité. En effet, pouvoir comparer des formes de réseaux et des modes de sociabilité dans un échan-

tillon d'individus choisis pour la similarité de leurs caractéristiques égocentrées permet d'explorer la grande diversité combinatoire de ces réseaux. Un des apports les plus évidents de ce dispositif est donc de montrer que le réseau peut difficilement se réduire à une collection d'indicateurs, puisque sous des caractéristiques numériquement identiques peuvent se cacher des formes de réseaux sensiblement différentes.

Exercices

Comment étudier Facebook à l'aide de l'analyse de réseaux ? Précisez la ou les questions de recherche, les données nécessaires et les définitions des liens et sommets du ou des réseaux étudiés.

Comment mobiliser l'analyse de réseaux pour étudier la scène NSBM (National Socialist Black Metal) à l'aide des plateformes discogs.com et metal-archives.com ?