Boullier, D., Crépel, M., et Jacomy. M. (2016), "Zoomer n'est pas explorer: spatialiser les graphes, catégoriser et (dé) construire les réseaux", Réseaux, n° 195, 2016/1.

Zoomer n'est pas explorer : spatialiser les graphes, catégoriser et (dé)construire les réseaux

Dominique Boullier – médialab Sciences Po
dominique.boullier@sciencespo.fr
Maxime Crépel – médialab Sciences Po
maxime.crepel@sciencespo.fr
Mathieu Jacomy – médialab Sciences Po
mathieu.jacomy@sciencespo.fr

Lorsque nous avons vu apparaitre sur notre écran la répartition spatiale des sites du web du livre francophone, il ne nous a fallu que quelques minutes pour commencer à raconter une histoire, sans doute même à NOUS raconter une histoire. L'image du graphe (figure 1) des liens entre les sites nous montrait « à l'évidence » une distance importante entre tous les blogs littéraires et les sites institutionnels des métiers du livre, depuis les éditeurs jusqu'aux bibliothèques. Cette observation tombait à point nommé pour appuyer notre discours général sur le livre numérique : le livre imprimé est déjà numérique par les conversations et commentaires qu'il génère et ce monde social est ignoré par les éditeurs qui prétendent pourtant prendre pied sur le marché du livre numérique contre les plates-formes américaines. Comme tous les sociologues, et même comme Durkheim l'admettait, notre discours à prétention scientifique profitait de cet effet visuel en l'élevant au rang de preuve. Quand bien même notre discours fut peu entendu, cette visualisation produisait son petit effet comme le font les chiffres qui sont supposés « parler d'eux-mêmes » et qui permettent de « prouver et gouverner » à la fois (Desrosières, 2014).

Que cherchions-nous à montrer ? La recherche réalisée en 2010 dans le cadre du projet SOLEN (financée par le FUI) portait sur les réseaux de circulation des livres papier et les formes de conversation-livre (entretiens auprès des acteurs du métier du livre, des réseaux de lecteurs hors web ou sur le web, des réseaux de vente, d'échange ou de don, observations de terrain, séances d'entretiens collectifs, cartographie de sites web, modélisation et production de modèles économiques). L'hypothèse à laquelle les partenaires professionnels du projet semblaient adhérer reposait sur les vertus de la circulation comme création de valeur sur internet. Le livre numérique ne parviendrait selon nous à prendre toute sa place qu'à la condition de pouvoir circuler, en dépassant les blocages des contrôles de type DRM et en inventant des modèles économiques où la revente d'occasion, le prêt et le don deviennent le ressort de plates-formes communes à tous les acteurs du livre. Nos études ethnographiques de terrain permettaient de documenter un véritable continent des activités sociales autour du livre (Le Béchec, Boullier et Crépel, 2016) dans lesquelles les échanges, les dons, les reventes de biens matériels exclusifs étaient très actives comparées aux blocages multiples rencontrés par les lecteurs de livres numériques, sur les plates-formes légales. Mieux même, le livre papier est à l'origine d'une activité considérable d'échanges, de conversations, dans des sites, des blogs, des forums, qui sont parfois devenus très spécialisés, ou très proches de certains auteurs, et qui font même émerger de nouveaux auteurs, par exemple à travers des fans-fictions. L'étude de tous ces sites à l'aide de méthodes automatisées mais supervisées de topologie du web, permettait de rendre visible la prolifération de ces sites. Mais elle faisait apparaitre en même temps des clusters, des grappes de sites à partir de leurs liens hypertextes et la distance entre sites des professionnels du livre et sites des amateurs pourtant très actifs

sur le web. Or, certains de ces clusters n'étaient pas aisés à distinguer, notamment celui entre sites de la bande dessinée et sites de la littérature jeunesse. Il ne nous paraissait pas aisé d'expliquer cette connexion, malgré notre connaissance du domaine. Ce cas atypique dans notre corpus agit en fait comme stimulant pour renforcer la robustesse de nos hypothèses car il risquait de remettre en cause toutes nos explications si « évidentes » sur la distance entre les sites selon leur « type ». C'est pourquoi nous nous sommes lancés dans une exploration des méthodes pour extraire les propriétés de ces sites et de ces liens qui permettraient de rendre compte de cet agrégat. Il nous fallut pour cela mobiliser de nombreux algorithmes différents et notre rapport aux données de la topologie changea alors radicalement. Plutôt que de les prendre comme des acquis « spontanés » ou « indiscutables » que nous nous contentions de labelliser, de catégoriser à partir de notre connaissance du domaine, nous avons dû entrer dans une véritable démarche d'exploration qui supposait de tester plusieurs choix de spatialisation et de clusterisation. C'est cette histoire que nous souhaitons rapporter ici car elle nous parait significative de postures de recherche différentes vis-à-vis de la puissance apparente des algorithmes, en plaidant pour une production de conventions plus exigeantes sur l'usage de ces méthodes.

Reprenons donc le fil de notre récit. Une fois produite la première carte générale du web du livre en France, nous étions donc convaincus mais aussi plutôt convaincants, sans pour autant penser un instant que nous pourrions promouvoir une escroquerie scientifique. Nous avons d'ailleurs constaté que beaucoup de nos collègues pratiquaient de la même façon avec leurs propres résultats basés sur une extraction massive de données sur le web. Bref, la force de conviction des visualisations spatiales fonctionne encore, aussi puissantes que celle de leurs ancêtres, les cartes. Si la critique experte des images de réseaux porte souvent sur leur construction ou sur leurs limites de validité, le reproche le plus fréquent porte avant tout sur leur incapacité à convaincre « au premier coup d'œil », tant la visualisation est complexe, assimilée souvent à un « plat de spaghetti ». Or, dans notre expérience, c'est leur pouvoir de conviction qui paraissait surprenant. Nos images semblaient résumer tout le travail de collecte des liens du web du livre numérique en un propos simple, un récit critique facile à faire circuler : les éditeurs sont déconnectés de ce qui fait la vie numérique du livre, la blogosphère littéraire. Et « il suffisait » de « voir » les distances entre les régions des différents sites (que nous avions labellisés et mis en couleur nous-mêmes) pour s'en convaincre. Pour les besoins d'une publication en noir et blanc, nous avons dû réinventer un code visuel qui n'a pas les qualités d'une sémiologie à la Bertin (1973) mais qui a au moins le mérite de rendre le graphe plus lisible selon nous, tant les nœuds deviennent bien différenciés alors que les couleurs produisaient une équivalence trop grande entre les nœuds. Il s'agit là peut-être d'éléments des conventions (Eymard-Duvernay, 2004) à construire que nous appelons de nos vœux pour l'exploration des graphes.

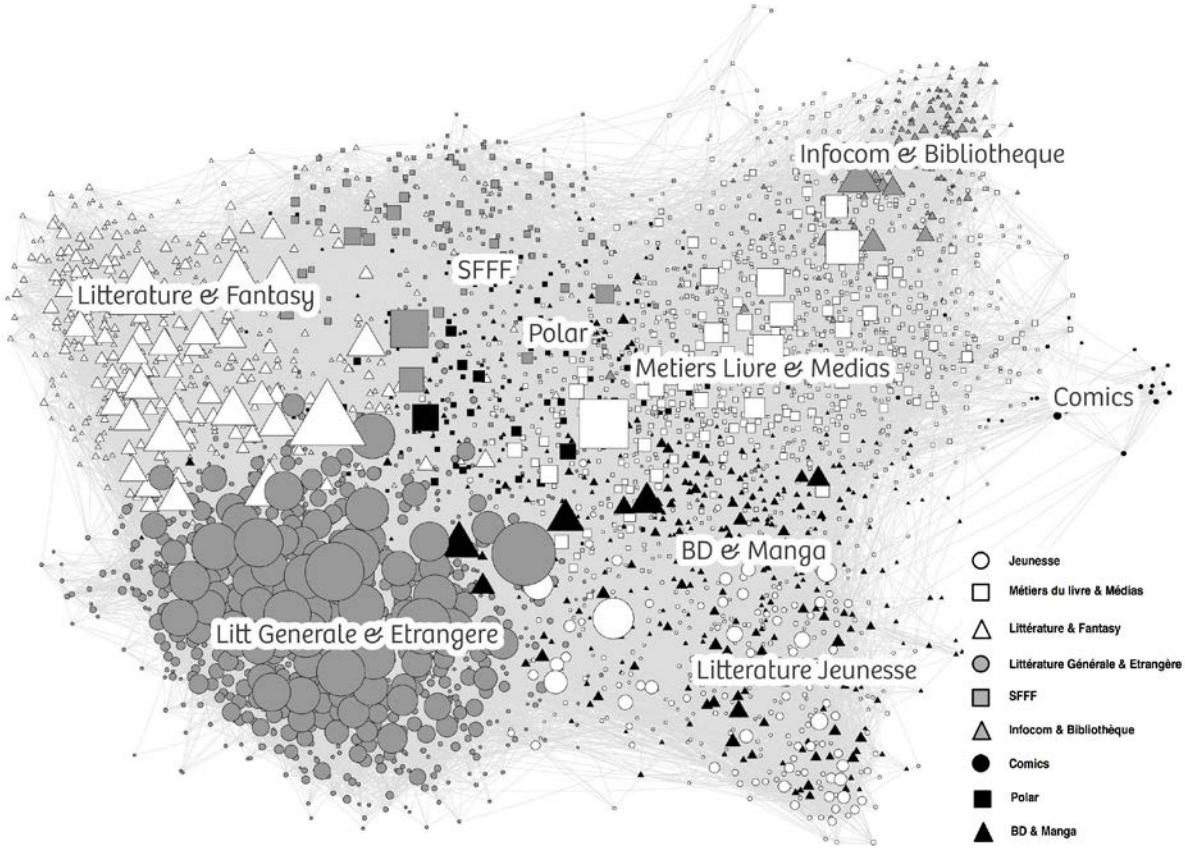


Figure 1- Web littéraire en France en 2010 par clusters thématiques - Algorithme de spatialisation : Force Atlas - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

Web littéraire en France en 2010

Constitution d'un corpus Web et nettoyage du graphe

Le cas proposé pour illustrer notre propos est tiré d'une étude réalisée au médialab de Sciences Po en 2010 qui porte sur les réseaux de circulation des livres et de la critique littéraire. Nous avons réalisé une cartographie du Web du Livre en France ayant pour objectif d'acquérir une connaissance de la structure du web littéraire marchand et non marchand et d'observer les configurations locales d'associations de sites par domaines littéraires. Les données ont été extraites à partir d'un crawl (extraction automatique de données du web) réalisé par la société Linkfluence en Décembre 2010, sur la base d'un corpus de 783 sites web (sites web repérés lors de l'étude ethnographique préalable). Le crawl a automatiquement extrait, à partir de ce corpus, tous les sites à une distance de n+1 et possédant au moins deux liens hypertextes avec le corpus de départ, soit 7 187 sites reliés par 120 217 liens hypertextes.

Difficilement exploitable à cause de sa taille et de la variété des sites web qui le constitue, le graphe obtenu à partir du crawl a subi différents traitements. Ainsi, un travail de filtrage a permis d'éliminer tous les sites jugés non directement liés à la thématique de recherche traitée et composant la « couche haute » du web (moteurs de recherche, sites techniques, réseaux sociaux, etc.), les sites ne relevant pas du web français (suppression des extensions de nom de domaine .ca, .ch, .uk, etc.), les sites inactifs, et les sites les moins connectés sur le graphe

composant la « couche basse ». A partir de ce travail le graphe se compose au final de 3 289 sites reliés entre eux par un ensemble de 52 884 liens.

Spatialisation, clusterisation et Catégorisation

La spatialisation du graphe a été réalisée à partir de l'algorithme Force Atlas (Jacomy, 2014). La proximité entre les nœuds résulte de leurs connexions avec leur environnement. La taille des nœuds est ici liée au degré, soit le nombre de liens entrants et sortants qu'ils possèdent avec les autres nœuds du réseau. Plus un noeud possède de liens entrants et sortants, plus son degré est élevé et sa taille importante sur le graphe. Les couleurs[1] sont fonction d'un algorithme de modularité qui permet de repérer de manière automatisée quelles sont les parties du graphe les plus connectées entre elles, les ensembles qui ont une forte densité interne et une faible densité externe. Ce travail de détection automatique des communautés a nécessité de fusionner à la main quelques clusters de taille très petite en les agrégeant aux composantes plus grandes auxquelles ils étaient le plus fortement liés afin de faciliter l'interprétation du graphe.

Si l'algorithme de clusterisation permet de repérer sur le plan topologique des ensembles plus denses, il revient au chercheur de les interpréter. Parallèlement à ce travail de spatialisation et de détection de clusters, chaque site a été visité par les chercheurs et s'est vu attribuer un ensemble de 5 tags : un tag lié au type de site web (blog, libraire, éditeur, etc.) et de un à quatre autres tags liés au domaine littéraire. Cette tâche laborieuse s'est faite de manière indépendante du travail d'analyse visuelle, de sorte que la spatialisation ou la clusterisation du graphe n'influence pas le travail d'attribution des tags aux sites web, favorisant la production de catégories homogènes pour des sites appartenant aux mêmes clusters. C'est seulement une fois ces tâches de clusterisation et de catégorisation effectuées indépendamment que l'interprétation des communautés à été effectuée à partir des tags (thématiques littéraires) les plus présents au sein de chaque cluster.

Interprétation du graphe

A partir de ce travail de catégorisation et en identifiant les types de sites répartis dans les différents clusters, il a été possible de comprendre les spécificités des différents sous-ensembles détectés de manière automatique par l'algorithme. Ce graphe filtré, spatialisé, clusterisé et catégorisé permet d'observer les différentes communautés qui composent le Web du livre en France. Il est constitué de trois régions distinctes qui diffèrent par le type d'acteurs et les domaines littéraires qui y sont représentés, mais également par leurs caractéristiques structurelles en termes de densité ou de diamètre. Ce travail a permis d'identifier une séparation marquée entre les sites web qui entretiennent les formes de conversation-livre d'une part (les blogosphères littérature générale et étrangère et littérature et fantasy situées à gauche du graphe) et les sites web de création, de production, de diffusion de l'objet-livre d'autre part (Métiers du livre et médias à droite du graphe). Cette séparation démontre la difficulté des constituer des communautés hétérogènes sur le web mêlant acteurs traditionnels et réseaux de lecteurs, à l'exception de petites communautés très thématisées au centre du graphe telles que la communauté Polar, la Science fiction (SFFF), la Bande Dessinée ou encore la littérature jeunesse qui ont su constituer des ensembles mêlant acteurs de la chaîne de production et blogs de lecteurs (Le Béchec, Crépel, Boullier, 2014)

[1] Dans cette version les couleurs ont été remplacées par des symboles (rond, triangle, carré) combinés avec trois teintes (blanc, gris, noir) afin de distinguer pour une version imprimée les différents clusters.

Epiphanie du social enfin révélé par les méthodes numériques qui « ne sont que » des « copies carbone » des « structures » enfin rendues « évidentes » par les liens entre sites. Les méthodes numériques que met en œuvre par exemple R. Rogers (2013) sont très soucieuses de ne pas se révéler in fine captives des formats et des conditions fixées par les plates-formes sur lesquelles elles recueillent les données car elles visent toujours à donner à voir « la société » *à travers* les images de réseaux ainsi construites[2]. Nous avions donc rempli le cahier des charges en filtrant et en spatialisant notre réseau de sites web. Cette exigence nous est apparue suffisante (et toute pratique de recherche effectue ces évaluations sur son propre niveau d'exigence) et nous n'avons pas pris le temps de discuter les éventuels doutes ou anomalies qui apparaissaient.

Depuis ce moment, nous avons eu largement le temps de revenir sur ces premières photos, car il s'agissait bien de photos (et même, vues du ciel), de l'état du web du livre francophone. On ne devrait sans doute pas effectuer ce genre de retour sur un travail passé, car les narrations construites dans l'emballement de la découverte n'y résistent guère, tout comme les évidences recueillies sur un terrain ethnographique après 5 jours d'immersion : tout semble clair, tous les modèles appris semblent s'appliquer parfaitement. Il suffit pourtant de rester quelques jours de plus pour voir s'écrouler ses premières impressions savantes, si savantes qu'elles avaient avant tout comme vertu de protéger le chercheur pour qu'il ne soit pas affecté par la singularité du terrain. Dans notre cas, nous connaissions suffisamment l'univers social du livre pour projeter immédiatement notre vision déjà bien documentée sur le corpus web recueilli et éviter de le laisser questionner ces constructions si précaires. Car ces visualisations dans l'espace d'entités simplifiées en points et en liens, qui tiennent lieu des arcs et des nœuds du graphe, nous font quelque chose, elles peuvent nous affecter, pour renforcer nos certitudes ou pour les déstabiliser, elles nous font faire et nous font croire, elles sont des médiations (Akrich, Callon, Latour, 2006) à part entière. Nous avions cependant gardé à part, comme un repentir, tout en le présentant dans les rapports et papiers produits à cette occasion, un beau cas de démenti de toute la méthode, d'anomalie dans l'effet copie carbone. Il s'agissait des communautés « BD et manga » et « Littérature jeunesse », repérées sur le graphe comme entrelacées (figure 2).

[2] Nous avons montré ailleurs (Boullier, 2015) qu'il existe plusieurs politiques de traitement de ces traces numériques, dont celle qui consiste à reprendre les catégories classiques des sciences sociales. Il est aussi possible de considérer ces traces comme des phénomènes propres aux univers numériques et de développer les méthodes spécifiques pour les penser comme vibrations, dans le cadre de sciences sociales de troisième génération.

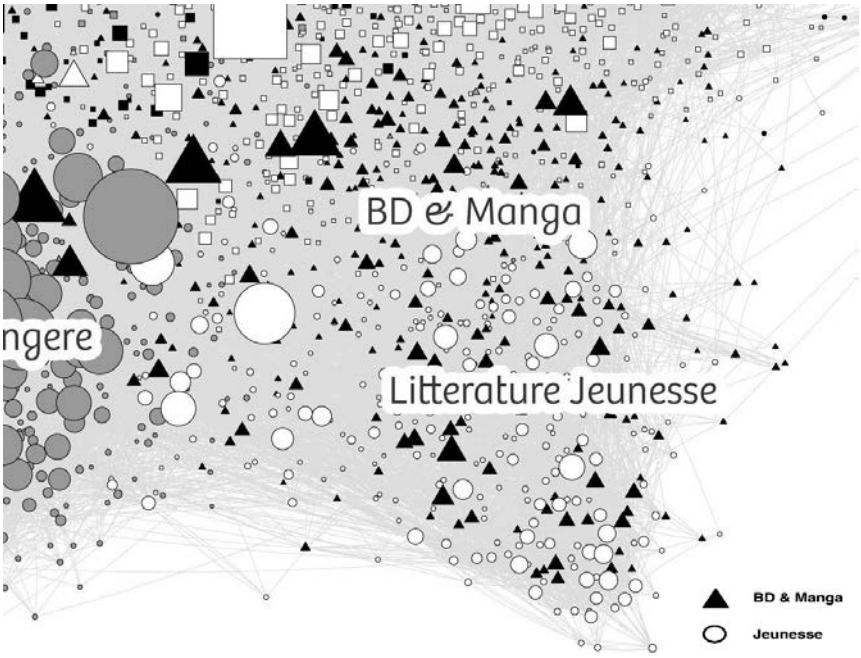


Figure 2 - Zoom sur l'entrelacement des clusters « BD et manga » et « Littérature Jeunesse » sur le graphe complet - Algorithme de spatialisation : Force Atlas - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

Ce cas avait en effet posé la question de l'argument du zoom pourtant si séduisant : le numérique n'agrège pas au point de dissoudre ou de faire disparaitre le singulier, comme le fait le traitement statistique des registres, mais au contraire permet de naviguer entre le tout et les parties, jusqu'à l'élément le plus fin, que sont le blog, le site, voire les citations si l'on construit un graphe avec des données textuelles (Latour et al. 2012). Or, pour démêler nos communautés et comprendre ce qui les composait et les tenait ensemble, il nous fallut procéder tout autrement que ce supposé zoom du sens commun. Le zoom avait à première vue l'avantage d'éviter le piège de la copie carbone et de la photo en permettant non seulement de changer de focale (le zoom de l'appareil photo) mais surtout de le faire dynamiquement, et donc de se transformer en zoom du cinéma, grande avancée pour les sciences sociales. Ce faisant, le zoom cinéma racontait une autre histoire, comme le fait tout mouvement de caméra, qui se déroule dans le temps. Mais cette histoire, nous allons le voir, ne pouvait en fait être racontée et validée dans le même temps qu'à la condition expresse de rompre avec toute idée de zoom photo ou cinéma, ce travelling optique qui ne modifie pas le rendu de la perspective. Car pour effectuer notre investigation de cette anomalie visuellement apparue dans l'image du graphe, il nous fallut sans cesse changer de perspective, précisément. C'est cette nouvelle histoire que nous voulons raconter en prétendant contribuer ainsi à l'élaboration d'une méthode d'investigation topologique, qui doit posséder ses procédures, ses bibliothèques d'algorithmes, ses limites de validité et ses tests de robustesse. Mais pour avancer dans cette voie, il convient de remettre en cause tant la métaphore de la copie carbone que celle du zoom, qui ont empêché jusqu'ici de penser la puissance propre des agencements bâtis sur les visualisations de graphe.

LE MOUVEMENT DU ZOOM ET LES DISTANCES SPATIALES OU SOCIALES[3]

La force de la visualisation réside dans son effet synthétique, qui impose un cadre indépassable au commentaire et à l'interprétation. Mais la force de la métaphore du zoom procède de la même façon : les médiations qu'il convient de mettre en œuvre pour réaliser cette interprétation sont avantageusement remplacées par une navigation qui rappelle des expériences communes, comme spectateur ou comme producteur d'images fixes ou animées. Que fait un zoom photo ou cinéma si ce n'est soutenir la perception d'une stabilité des repères, construits selon les principes de la perspective alors même que le changement de focale conduit à un changement de cadrage et que le mouvement de caméra reproduit les effets d'un traveling, dit optique, qui demande mise au point automatique permanente sous peine de perdre la netteté. Les mouvements caméra sont assez limités (zoom, pan, tilt, traveling) mais sont supposés s'appliquer à un espace à agencement topographique. Lorsque les distances sont calculées dans un espace à agencement topologique, avec les méthodes des graphes, qui semblent mêler à la fois principes formels de calcul et principes de visualisation, il semble improbable de retrouver des équivalents des mouvements de caméra. Certains chercheurs ont cependant tenté d'inventer les interfaces qui permettent de naviguer dans des bases de données en transformant les distances calculées en distances spatiales topographiques (Lecolinet, 2002) de façon à provoquer des effets de familiarité pour les usagers non experts. Or, c'est aussi ce qui se passe avec la visualisation des graphes lorsque l'on veut la rendre aisément manipulable par des non-experts. Ils finissent par considérer que les distances calculées entre sites sont « naturellement » représentées dans l'espace à deux dimensions d'un écran comme distances spatiales et mieux encore que ces distances sont des équivalents des distances sociales ainsi rendues visibles. Ces équivalences des espaces de données, des espaces topographiques et des espaces sociaux, jouent sur la polysémie du terme espace. Elles avaient été largement exploitées par Bourdieu lorsque, dans « La Distinction » (Bourdieu, 1979), il présentait ses Analyses Factorielles de Correspondances calculées à l'aide des méthodes de Benzecri sous forme de visualisations le long d'axes (aidé en cela par l'invention récente des eigen-vectors). Il s'autorisait à y plaquer les items de CSP et les axes relabellisés en « volumes de capitaux » (économique et culturel) qui n'avaient en rien été calculés selon les mêmes principes mais seulement postulés par la théorie des différentes formes de capitaux (économique et culturel) propre à Bourdieu.

Il est aisé de critiquer Bourdieu après coup mais la fabrique des images de graphe revient le plus souvent au même et leur force de conviction repose sur le même effet. Certes, il est possible de rappeler que l'orientation de l'image de graphe n'a aucune signification contrairement à celles que l'on peut attribuer aux axes des analyses factorielles de correspondances, et que la visualisation ne permet pas de positionner des hiérarchies. Et pourtant, les clusters découverts par l'algorithme sont immédiatement et quasi intuitivement

[3] Nous entendons toujours dans ce texte « distance spatiale » comme un résultat d'un calcul effectué par un algorithme de spatialisation qui détermine la distance affichée sur l'écran, sans aucune référence à une distance géolocalisée. Mais l'espace d'affichage de l'écran constitue cependant en tant que tel un référentiel topographique quand bien même il permet de visualiser une topologie. C'est pourquoi la confusion est aisée et l'analogie spontanée dans la perception ordinaire. La distance sociale est elle aussi sujette à raccourcis provoqués par les algorithmes de clusterisation. Ce sont eux et eux seuls, qui permettent de produire des agrégats fondés sur les liens hypertextes que l'on a tendance à rapidement assimiler à des communautés dont on mesure les distances sociales. Or, ce n'est qu'après un long travail de catégorisation progressive, que nous présentons ici, associé à des allers-retours avec les connaissances issues de l'observation qualitative des matériaux recueillis, que l'on peut valider la pertinence d'une visualisation pour rendre compte d'entités sociales faisant sens dans le monde vécu des acteurs eux-mêmes. Mais la vigilance méthodologique dont nous voulons faire ici la démonstration devrait rendre plus prudents quant à l'usage des termes distances sociales et spatiales. Sur ces concepts, voir Levy et Lussault (2003).

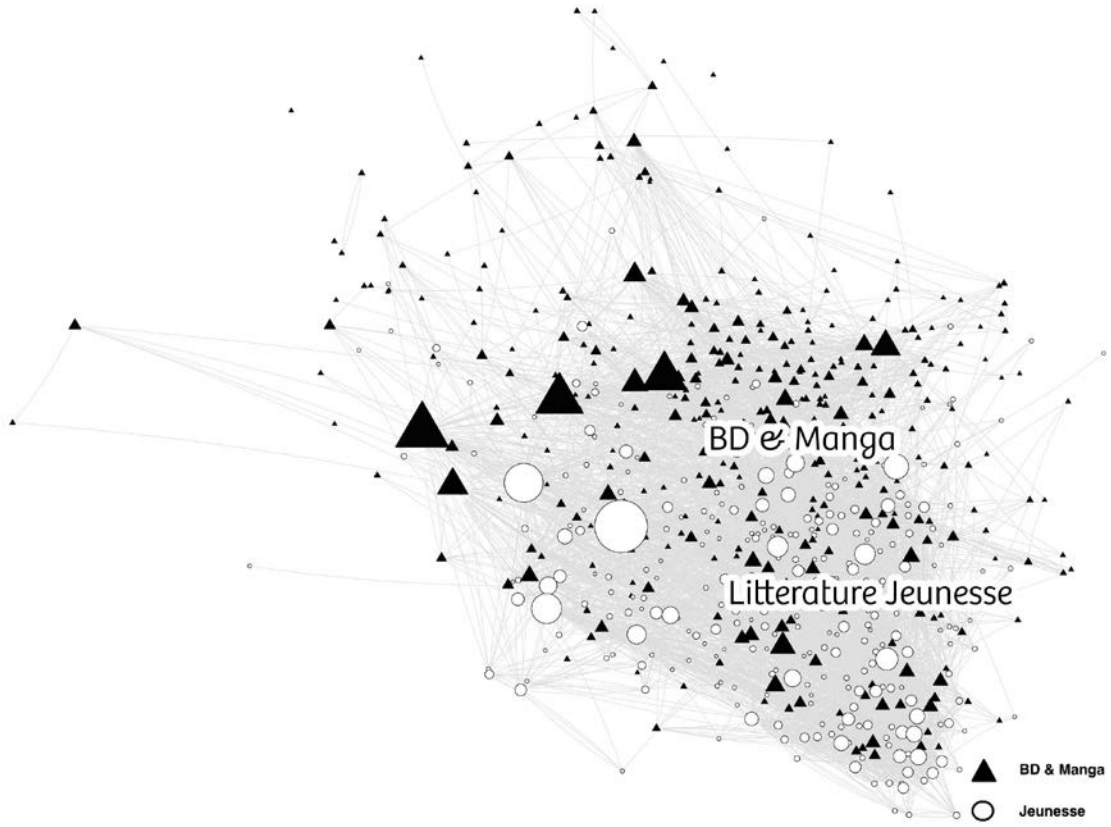
traduits en entités sociales dont les distances sont significatives, ainsi que nous l'avions fait pour la visualisation des « communautés » du web du livre. Le terme lui-même de « communauté » est significatif du glissement immédiat qui se produit entre un effet de cluster produit par le calcul et son équivalent social en termes de communauté. Il n'a le plus souvent aucune valeur par rapport aux traditions sociologiques mais il permet de sauter immédiatement du monde des artefacts numériques (les sites web, les liens) au supposé « vrai » monde social. On pourrait pourtant y appliquer un principe de précaution qui consisterait à parler d'agrégat, mais la puissance de la démonstration sociologique immédiate s'affaiblirait. Nous proposerons plutôt de parler de topoï. Nous explorons en effet un espace à agencement topologique qui produit trois dimensions à la fois :

- des topiques (au sens rhétorique) : les sites qui sont liés entre eux exploitent les mêmes éléments de discours qui se traduisent par des thèmes (topics) partagés, repérables et calculables par la distance entre chaines de caractères (méthodes de word space et de co-occurrence en général par exemple).
- des régions ou des lieux (au sens spatial quasi topographique dans l'espace en deux dimensions d'un écran) : les algorithmes de spatialisation et de clusterisation produisent des proximités et des distances qui sont des conversions des tables issues des rankings issus eux-mêmes des calculs de clusterisation.
- des types (au sens weberien du terme), des catégories ou des profils : en assemblant les résultats des deux premières dimensions, chaque cluster doit regrouper des entités qui présentent les mêmes attributs, censés dès lors constituer le prototype qui va servir à labelliser le cluster en question par une opposition nette avec les autres clusters. Car il ne faut jamais oublier cette étape de catégorisation qui renforcera la lisibilité des distances et facilitera le travail d'interprétation. Les images de graphe ne deviennent lisibles qu'une fois habillées de termes qui sont autant de catégorisations qui font tout le travail scientifique.

Or, toutes ces dimensions ne se trouvent jamais alignées de façon cohérente par la simple puissance des calculs car ce sont des méthodes différentes et des couches différentes du tissage social pourrait-on dire, en reprenant cette fois la métaphore des SIG. Un travail considérable qui fait l'objet de discussions et d'arbitrages préside à cette opération qui peut parfois tenir comme chez Bourdieu du forçage de l'alignement (on retourne les axes dégagés par l'analyse des AFC pour les rendre plus cohérents avec les axes des hiérarchies de capital). Il existe bien une axiologie, une éthique et une morale au cœur de tout travail scientifique pour produire ces effets de vérité finalement si puissants (Latour, Fabbri 1977).
Nous allons montrer comment nous avons effectué un travail long et minutieux de mise à l'épreuve de nos méthodes de calcul pour explorer les propriétés de l'anomalie de l'entrelacement spatiale des clusters BD et Jeunesse. Les clusters qui sont apparus avec les premiers calculs de modularité diffèrent de ceux de l'algorithme de spatialisation utilisé préalablement en ceci que les régions ne sont pas nettes, malgré le forçage des classifications réalisé par tout algorithme de clusterisation. Mais l'incohérence n'est pas seulement visuelle puisque les catégorisations effectuées par nos soins y jouent un rôle important. Pourtant, c'était avant tout la disposition visuo-spatiale non alignée, non cohérente avec les catégorisations (elles-mêmes mises en scène visuellement par des couleurs et des symboles dans cette version) qui alertait nos sens avant même notre vigilance méthodologique ordinaire.

L'ENTRELACEMENT DE LA BD ET DE LA JEUNESSE

Dans ce cas d'entrelacement, il serait possible d'aller voir de plus près, des sous-régions, les types de sites, en « zoomant » au point même d'aller voir chaque site un par un, puisque le numérique permet effectivement de ne pas perdre les parties dès lors qu'on produit des « tout », des agrégats ou des clusters, ce qui est un avantage considérable sur les méthodes statistiques non numériques. Mieux même, il ne s'agit pas de retrouver la fiche descriptive du site, de l'entité de base, mais bien le site actif, sans qu'il ait disparu sous les catégories ou les traitements, comme si l'on pouvait retourner voir sans cesse les personnes interrogées lors d'un sondage pour vérifier qu'elles avaient bien répondu de telle ou telle façon à la question. Nous décidons même, pour plus de clarté, de détacher cette partie entrelacée, cette région, du « tout », c'est-à-dire du reste du graphe, en supprimant les liens qui les unissent aux autres sites tout en conservant leur spatialisation obtenue à l'origine (figure 3). L'entrelacement se voit mieux mais nous n'en comprenons pas plus les propriétés. D'une certaine façon, nous avons recadré, c'est-à-dire enlevé du bruit. Le zoom classique, même recadré, ne donne rien. Il est en effet normal que les parties du graphe restent positionnées en



fonction du graphe complet quand bien même on fait disparaitre visuellement les liens.

Figure 3 - Clusters BD & JEUNESSE détachés du graphe complet spatialisé -Algorithme de spatialisation : Force Atlas - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

Il nous faut alors entrer dans une démarche de recalcul itératif de l'agencement du graphe pour mieux comprendre les propriétés de cet entrelacement. Nous ne cherchons pas à l'expliquer, à en trouver une cause à proprement parler, mais seulement à nous assurer qu'il y a bien entrelacement et à faire apparaitre les propriétés de cet entrelacement. A ce moment, les données sont bien des « obtenues » (Latour, 2012), quand bien même elles l'étaient auparavant : il nous faut retravailler, re-paramétrer, changer d'algorithme, etc. Nous sommes ici plus proches d'une démarche réellement expérimentale. Or, c'est exactement ce que

devraient permettre de faire les visualisations de graphe, de façon à contester les évidences supposées des premiers résultats. Il faudrait même souhaiter qu'aucun résultat immédiat ne soit considéré comme acceptable tant qu'il n'a pas subi des tests de robustesse du type de ceux que nous avons tentés. La démarche de tests a porté sur plusieurs éléments composant au final l'image qui nous parait problématique :

- l'algorithme de spatialisation,
- la catégorisation par l'algorithme de clusterisation et par le chercheur,
- les effets du graphe global (le tout empêche de voir les parties),
- les effets du graphe local et le choix des descripteurs appliqués au graphe

Un biais d'analyse lié au choix de l'algorithme de spatialisation

Tout chercheur qui a déjà pu visualiser des réseaux est étonné de la variété des algorithmes de spatialisation et des résultats qu'ils produisent. Chacun d'entre eux possède ses propres modalités et un grand nombre de paramètres doivent être réglés afin d'obtenir une spatialisation qui soit lisible par le chercheur, sans compter qu'un même algorithme produit des résultats différents d'une fois sur l'autre. Notre premier réflexe face à l'entrelacement des clusters BD et Jeunesse a donc été de nous interroger sur le poids du choix de notre algorithme (Force Atlas) afin de comprendre s'il pouvait être en partie lié au phénomène que nous observions. La manière la plus simple de tester l'algorithme est de procéder à une comparaison des résultats avec d'autres outils de spatialisation. La figure suivante montre les

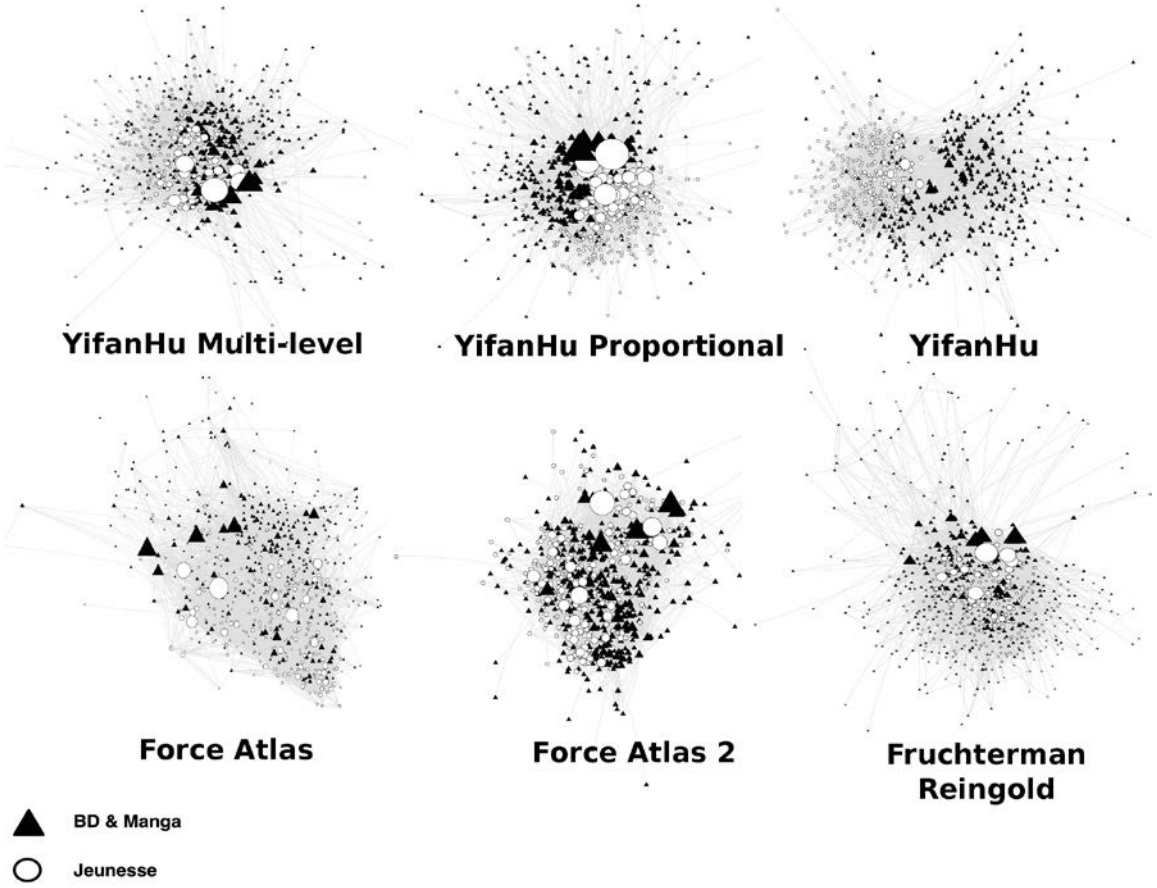


Figure 4 - Clusters BD & JEUNESSE détachés du graphe complet spatialisé avec différents algorithmes précisés pour chaque figure - Taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

spatialisations réduites aux deux communautés BD et Jeunesse, extraites des graphes complets sur lesquels on a appliqué différents algorithmes (figure 4).

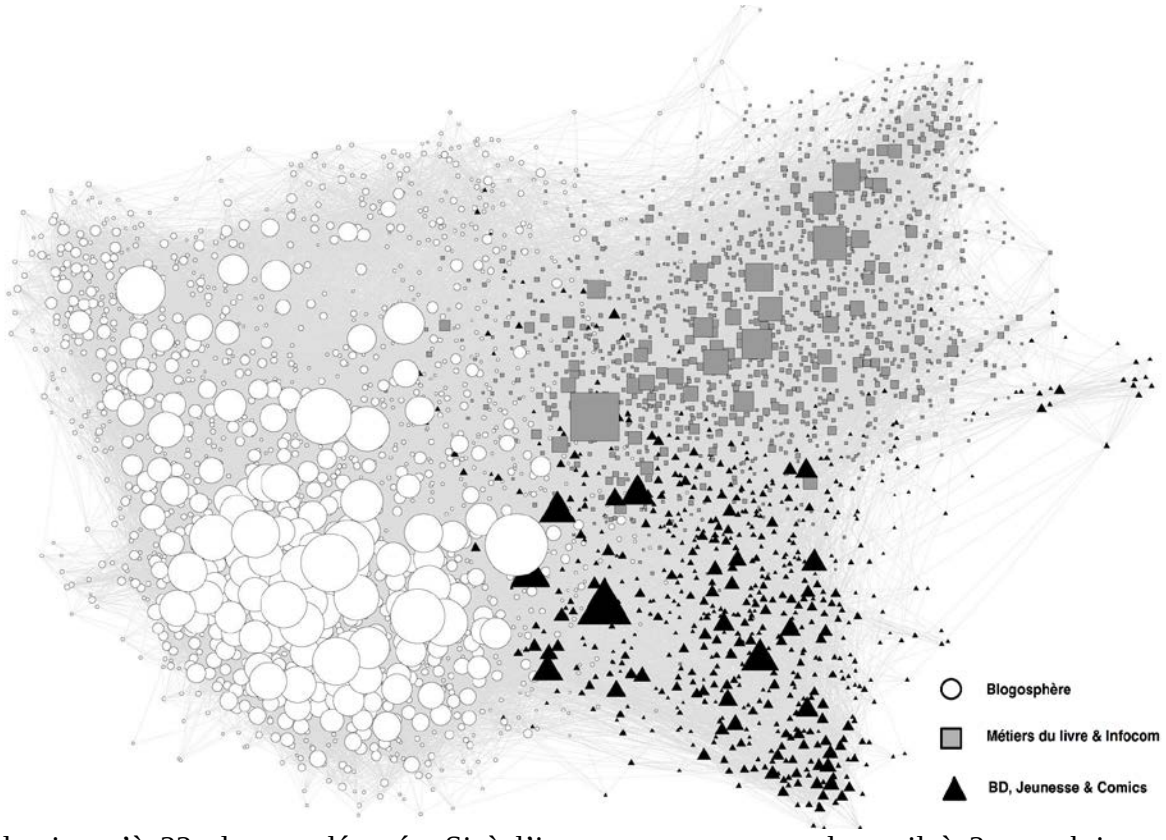
Le premier résultat fait apparaitre que sur six algorithmes présents par défaut dans le logiciel Gephi (Bastian, Heymann, Jacomy, 2009), seul l'un d'entre eux (Yifan Hu) produit une séparation plus marquée des deux clusters lorsque l'on applique ce dernier sur le graphe complet. Plus précisément, il détache les deux clusters de manière plus franche mais on constate encore de nombreux nœuds du cluster BD (triangles noirs) qui viennent se placer dans l'espace à gauche occupé principalement par le cluster Jeunesse (ronds blancs). Dans les cinq autres cas, le phénomène d'entrelacement est très visible et les deux clusters se partagent le même espace. Nous pouvons en déduire que le choix de l'algorithme n'a pas introduit dans ce cas de biais qui aurait produit un phénomène d'entrelacement entre les deux clusters. Tous ces algorithmes ont en commun de montrer les mêmes agrégats visuels, mais de façon différente. Mathématiquement, cette famille d'algorithmes manifeste les clusters optimisant la modularité, c'est pourquoi il y a une correspondance entre la couleur et la position des noeuds dans chaque image. La différence la plus importante entre les algorithmes se trouve dans le rapport des forces d'attraction et de répulsion des noeuds. Le plus ancien algorithme, Fruchterman Reingold, propose un équilibre qui laisse apparaître de nombreux filaments, le résultat n'est pas compact (voir figure 4 en bas à droite). Cette caractéristique rend confuse inutilement l'image et les algorithmes suivants essayent de la minimiser. De ce point de vue, les trois algorithmes de Yifan Hu sont un peu meilleurs, et les deux Force Atlas meilleurs encore. L'équilibre optimal est atteint par le LinLog d'Andreas Noack (non montré ici). Les algorithmes de Yifan Hu ont une philosophie spécifique car ils disposent d'une heuristique pour arrêter le calcul pour déterminer le point d'arrêt optimal, tandis que l'arrêt est manuel dans les trois autres. Les dernières différences concernent l'optimisation de l'algorithme. La version 2 du Force Atlas accélère le calcul avec une approximation, tandis que les deux variantes du Yifan Hu sont deux optimisations différentes.
Il n'est pas anodin de rappeler la diversité des choix d'algorithme de spatialisation. Dans un souci de simplification des outils et de démonstration spectaculaire, on peut faire disparaitre tous ces choix pour mettre en avant « la science faite ». Pourtant les images fournies sont au final très différentes. Les cartes du web ne reposent pas sur une convention unique ni sur des paramètres de référence considérés comme des conventions stables, comme le sont la plupart des cartes topographiques pour des usages ordinaires. Les cartes des géographes, pourtant, sont elles aussi des processus, elles reposent sur des protocoles et les anamorphoses (Lévy, Maitre et Romany, 2016) montrent comment de nouveaux critères de spatialisation peuvent être introduits sur un fond topographique supposé conventionnel et pour cela supposé évident.

Catégories du chercheur et de l'algorithme de modularité

Si le biais de l'algorithme de spatialisation semble écarté, il est important d'interroger les catégories qui servent à interpréter le graphe qui pourraient être à l'origine de l'entrelacement des deux clusters BD et Jeunesse. La catégorisation du graphe résulte, comme nous l'avons expliqué auparavant, de deux opérations conjointes et indépendantes qui ont consisté, d'une part, à appliquer un algorithme de modularité au graphe afin de détecter des clusters dans la structure (Blondel, Guillaume, Lambiotte, 2008), et d'autre part, à étiqueter tous les sites qui composent le graphe de manière qualitative en leur appliquant au moins un tag (type de site web ou type d'acteur) et jusqu'à quatre tags supplémentaires pour décrire de manière synthétique les genres littéraires présents sur les sites. C'est à partir de ces deux types de

classification, algorithmique et qualitative, que l'interprétation des clusters a pu être réalisée. Pourtant ce travail de création des frontières internes au graphe, de découpage en entités distinctes et présentées comme cohérentes relève de choix effectués par le chercheur qui dispose de différents outils et méthodes.

Il existe différents algorithmes de clustering qui possèdent chacun leurs spécificités et produisent sur un même graphe des clusters différents. L'algorithme de modularité, qui est une forme de clustering spécifique, peut lui même varier en fonction du seuil de « résolution » que l'on détermine avant de l'appliquer au graphe. Ainsi le nombre de clusters et leur taille varient suivant la définition de ce seuil. Sur le graphe étudié, la clusterisation en 9 clusters est obtenue avec un seuil de « résolution » (Lambiotte, 2009) réglé par défaut sur 1, mais il faut pourtant préciser que ce nombre de clusters peut également varier d'une passe à l'autre car le modèle fonctionne par comparaison à un modèle aléatoire qui peut entrainer une variation à la marge du nombre de clusters obtenus. Il est parfois donc nécessaire de recomposer les clusters du graphe en associant aux plus stables, des clusters de taille très restreinte (moins de 1% du graphe) de sorte que l'on puisse obtenir une lecture « plus aisée » de ses différentes composantes. Si l'on applique au graphe un seuil de 0.5, le nombre de clusters obtenus peut



aller jusqu'à 22 clusters détectés. Si à l'inverse on augmente le seuil à 2, on obtient un découpage en 3 clusters principaux, la blogosphère (ronds blancs) les métiers du livre et infocom (carrés gris), et une composante commune (triangles noirs) pour les clusters initialement distingués que sont BD et Manga, Jeunesse et Comics. L'entrelacement des clusters BD et Jeunesse n'est plus visible dans ce cas puisqu'ils sont inclus dans un seul et même cluster (figure 5).

Figure 5 - Web littéraire en France en 2010 – Seuil de 2 pour l'algorithme de modularité produisant 3 composantes principales - Algorithme de spatialisation : Force Atlas - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

Cette variété des choix méthodologiques et le travail de recomposition des clusters montrent à quel point le traitement appliqué sur un graphe est important dans le processus d'analyse des résultats. Bien qu'il s'agisse ici d'outils algorithmiques qui semblent automatiques, les choix qui sont effectués par le chercheur dans la configuration de ces outils modifient les résultats obtenus et le récit qui pourra y être associé. Tester la sensibilité et la robustesse de ces outils permet, en procédant pas à pas, d'effectuer les choix les plus cohérents en fonction des objectifs de l'étude et d'être également alerté sur les limites de l'interprétation des résultats obtenus.

Le travail de qualification des nœuds, quant à lui, ne repose que sur une connaissance du domaine étudié et sur un travail qualitatif d'indexation minutieux et lent lorsqu'il s'agit d'un graphe de cette taille. Les catégories du chercheur sont, comme celles des algorithmes, tout aussi discutables. Dans le cas présent, elles sont largement inspirées des classifications en matière de genres littéraires telles qu'on les trouve chez les libraires. Les débats classiques sur les limites des classifications ordinaires et expertes montrent bien à quel point catégoriser le monde amène systématiquement à faire des choix en fonction d'objectifs spécifiques qui n'ont le plus souvent de cohérence que pour une communauté qui partage des centres d'intérêts et un langage communs. Mais cela indique également que ces catégories reposent le plus souvent sur des frontières poreuses que l'on tente de maintenir en prenant des décisions arbitraires face aux cas spécifiques qui entrent difficilement dans des catégories préétablies. De la même manière, les catégories utilisées dans cette étude présentent de nombreuses imperfections ainsi que les principes de généralisation qu'il a fallu adopter afin de faciliter la lecture du graphe. Ainsi, on pourrait largement discuter du terme utilisé pour le cluster « Littérature générale et étrangère » qui, comme son nom l'indique, regroupe des blogs de lecteurs qui offrent des entrées très variées en littérature classique et contemporaine. On pourrait encore discuter de la distinction faite entre le cluster « Littérature et Fantasy » (blogs principalement centré sur l'heroïc fantasy) et le cluster « SFFF » pour Science Fiction, Fantasy et Fantastique qui regroupe des blogs et sites plus axés sur la Science Fiction traditionnelle. Enfin l'existence d'un cluster « comics » séparé de la « BD et manga » est tout aussi discutable. Nous avons limité les perturbations en choisissant un seuil pertinent pour faire « suffisamment » de distinction dans les entités qui composaient notre graphe, sans nous noyer dans une multitude de sous-communautés difficile à interpréter. L'analyse de ces clusters à partir des tags appliqués à chaque site, de manière indépendante de ce traitement algorithmique, nous a amené progressivement à stabiliser nos catégories pour rendre le graphe intelligible et en proposer une analyse qui semble cohérente. Un aller-retour constant entre connaissances expertes et paramétrage des algorithmes doit constituer une phase essentielle du travail pour éviter de fétichiser toute catégorie. Il serait aussi nécessaire de pratiquer des tests de robustesse de ces classifications qui, selon les méthodes utilisées et les volumes de données, peuvent parfois demander un long temps de calcul.

La tension relationnelle entre les différentes composantes du graphe

La troisième hypothèse émise pour comprendre cet entrelacement consiste à considérer que ce sont les relations des clusters BD et Jeunesse avec le reste du graphe qui les amènent à se spatialiser dans la même région du graphe complet. Autrement dit, les liens que partagent les clusters BD et Jeunesse avec les autres clusters sont assez semblables pour que les algorithmes de spatialisation les contraignent à se positionner dans le même espace. Afin de tester cette hypothèse, il faut procéder à une suppression progressive des clusters voisins sur le graphe complet et procéder à une nouvelle spatialisation pour observer l'effet que cette perte de relations peut avoir sur l'entrelacement des deux clusters BD et Jeunesse. Il s'agit

donc de décomposer pas à pas le graphe et de le spatialiser à nouveau. Cela revient de fait à produire un effet de zoom car on élimine petit à petit l'environnement mais cela n'est qu'une impression visuelle, car l'opération consiste au contraire à prendre en compte à quel point chaque nœud est tributaire de tous les autres pour son positionnement. Or, le « tout » que nous prenons comme point de départ n'est lui-même qu'un artefact que nous avons construit à partir de notre connaissance du domaine. Aucun « tout » n'est saisi autrement que par la médiation d'un artefact. Mais ici, nous avons les moyens de changer son périmètre à volonté pour voir ce qui demeure dans la spatialisation de départ qui nous posait problème.

Si l'on soustrait au graphe complet la plus importante composante du graphe qu'est le cluster « Métiers du livre et médias » et que l'on applique à nouveau l'algorithme Force Atlas, on constate que les deux clusters BD et Jeunesse tendent à se séparer (figure 6 - première image à gauche). De la même manière en supprimant la seconde composante du graphe représentée par le cluster « Littérature Générale et Etrangère », on observe que la séparation entre BD et jeunesse est plus marquée (figure 6 - image du centre). Le résultat est moins significatif si l'on supprime le cluster « Littérature et Fantasy » mais la suppression de ce cluster semble limiter l'entrelacement des clusters BD et Jeunesse (figure 6 – image de droite).

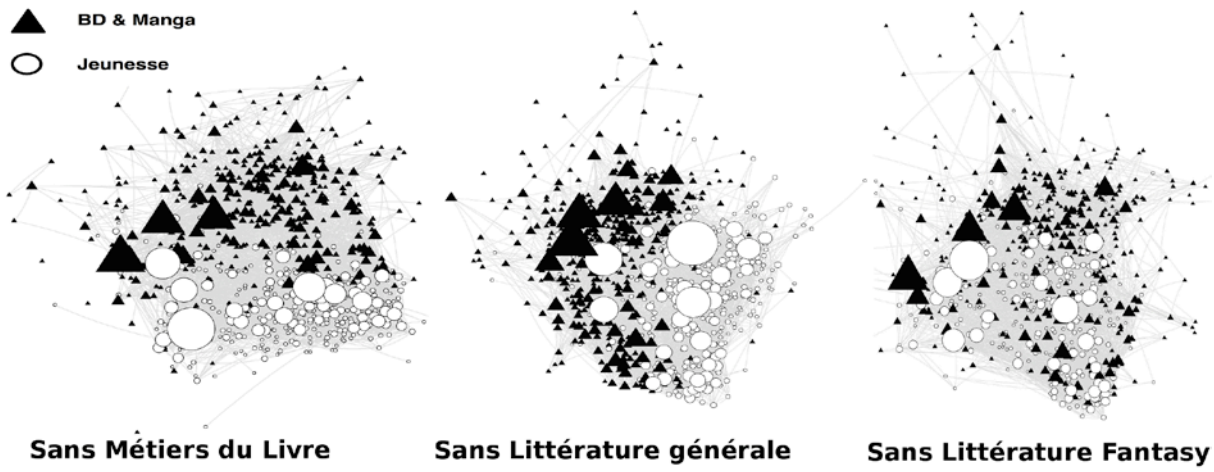


Figure 6 - Clusters BD et Jeunesse découpés du graphe spatialisé par Force Atlas en supprimant successivement les 3 composantes principales - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

Si l'on supprime à présent en même temps les 3 composantes principales que sont « Métiers du livre et médias », « Littérature Générale et Etrangère » et « Littérature et Fantasy » mais que l'on conserve les autres clusters en appliquant un algorithme de spatialisation Force Atlas, on observe une séparation totale des deux clusters BD et Jeunesse (figure 7).

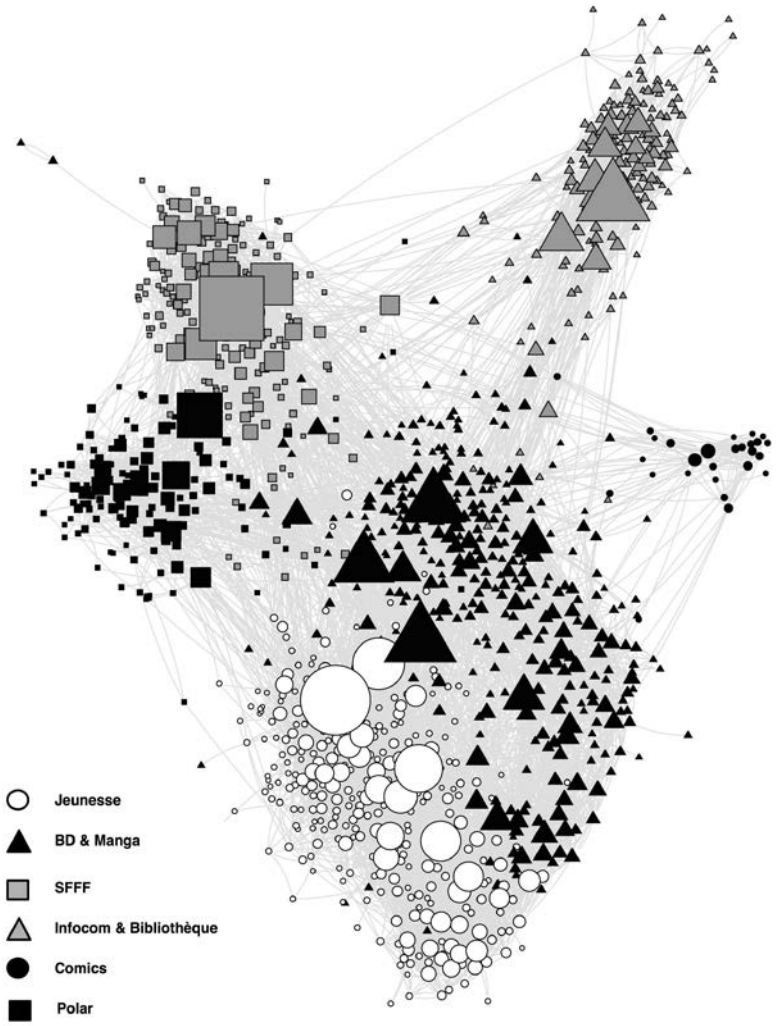


Figure 7 - Web littéraire en France en 2010 spatialisé par Force Atlas en supprimant les 3 composantes principales - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

Il semble que notre hypothèse soit validée en partie car les tensions générées par les trois composantes principales du graphe exercent une contrainte qui tend à faire s'entrelacer les clusters BD et Jeunesse. Une fois libérés de tout ou partie de cette contrainte les clusters détectés par l'algorithme de modularité comme deux communautés distinctes se séparent pour occuper un espace qui leur est propre, même si elles restent juxtaposées.
Ce résultat peut satisfaire le chercheur. En effet, les catégories (« BD » et « Jeunesse ») qu'il avait appliquées au début de son exploitation du graphe continuent de faire sens. Car ce qui posait problème était bien l'inadéquation entre la visualisation et cette catégorisation et non la visualisation en tant que telle. Cet élément de méthode est important car ce sont les imperfections, les marges, les cas particuliers ou incohérents qui constituent le ressort de toute méthode scientifique : lorsque toutes les hypothèses sont confirmées, parfois du premier coup, il existe à coup sûr un risque d'artefact de méthode. Il est donc très important de tenir compte des travaux qui font état d'échec, d'impasses ou de résultats négatifs car ce sont eux qui vont inciter à faire avancer la communauté scientifique. Or, la force d'imposition des images de graphes est souvent mobilisée pour démontrer la validité d'une interprétation particulière qui devrait le plus souvent rester provisoire. Cependant, ce résultat ne nous est pas apparu suffisant à nouveau, car obtenir seulement la confirmation de ses catégorisations ne permet pas de comprendre ce qui provoquait cet entrelacement une fois le graphe général réinjecté dans les liens des deux clusters. Il fallait donc tester encore une dernière hypothèse.

La structure et la composition interne des clusters BD et Jeunesse

Sans invalider l'hypothèse précédente, c'est en s'intéressant à la composition interne et en appliquant d'autres traitements à ces deux clusters que nous comprenons en partie la raison pour laquelle ils ne se distinguent pas sur le plan de la spatialisation mais sont pourtant détectés comme deux clusters différenciés du point de vue de l'algorithme de modularité. Nous partons du résultat précédent, où le graphe des deux clusters a été séparé du reste du graphe global puis respatialisé avec l'algorithme Forceatlas. On distingue bien (figure 8 à gauche) que les deux communautés, ainsi libérées des tensions avec le reste du graphe complet, se séparent distinctement.
Il faut encore vérifier si les catégories mobilisées sont les seules pertinentes et surtout chercher à comprendre ce qui peut produire malgré tout un entrelacement. A la fois remettre en cause et discuter les classifications du chercheur et de l'expert, systématiquement, et se laisser affecter par le démenti apporté par l'entrelacement, sans vouloir le réduire a priori à un artefact. Il faut alors mobiliser d'autres tags descriptifs des sites, qui sont autant de nouveaux principes de classification qui peuvent compléter ou contester les catégories de BD et de Jeunesse. Nous tentons alors d'exploiter la classification par types d'acteurs (Blogs et Blogs ProAm, Editeurs, Libraires, etc.). Cela revient-il à « zoomer » dans une des propriétés de ces ensembles, en déclinant un attribut de leur profil, celui du type d'acteurs ? Pas vraiment, car il s'agit alors de réorganiser tout le graphe en fonction de ces nouvelles catégories et de vérifier si elles nous apprennent autre chose, ce qui peut conduire à invalider la pertinence des premières catégorisations. Là encore, déplacement de point de vue et non de zoom, pourrions-nous dire, car ces supposés attributs sont en fait des candidats comme les autres à prendre le « tout » dans leurs principes. L'absence de hiérarchie et d'emboitement est importante à préserver et nous rapproche d'une méthode monadologique puisque toute catégorie peut embrasser le reste du monde à partir de son point de vue (Tarde, 1893).

La comparaison ci-dessous des deux graphes, l'un basé sur les genres littéraires (tels que les catégories BD et Jeunesse) et l'autre sur les types d'acteurs (tels que blogs et blogs ProAm[4], éditeurs, libraires) nous parait visuellement très stimulante car nous obtenons une nouvelle partition claire du graphe (figure 8). En effet, en modifiant ces catégories, on constate que deux clusters se distinguent et montrent une séparation entre d'un côté des blogs de lecteurs et de critiques dans le domaine de la BD et de la Jeunesse (les ronds blancs sur le graphe de droite figure 8) et de l'autre les blogs ProAm (les carrés noirs sur le graphe de droite figure 8). Ce qui distingue ces deux types de sites est que Les "Blogs ProAm" sont des sites d'illustrateurs et d'auteurs qui publient leurs créations sur leur blog, alors que les "Blog" correspondent aux sites de lecteurs de BD ou de littérature jeunesse qui publient des billets d'avis ou des notes critiques sur les œuvres qu'ils lisent. Les "Blogs ProAm", qu'ils appartiennent au monde la BD ou à celui de la littérature jeunesse, sont très liés entre eux et jouent ce rôle de *bridges* (Burt, 2005), à l'échelle de toute une catégorie. Cette division offre une nouvelle partition du graphe qui n'est plus fonction du genre littéraire mais qui distingue les simples lecteurs et critiques de ceux qui produisent et publient des contributions (illustrations ou histoires).

[4] ProAm pour professionnel amateur, voir à ce sujet Leadbeater C., Miller P. (2004), *The Pro-Am Revolution: How Enthusiasts are Changing Our Society and Economy*, Demos.

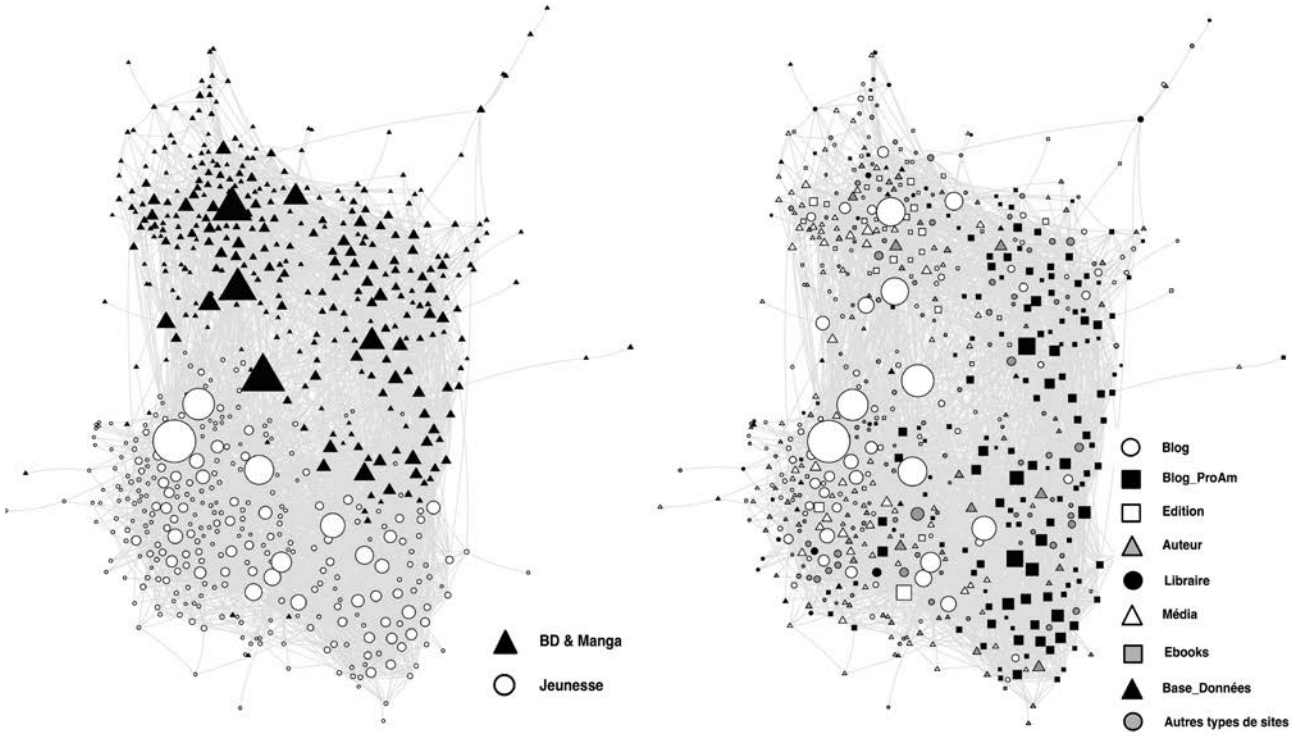


Figure 8 - Clusters BD & JEUNESSE découpés du graphe complet et spatialisés ensemble avec l'algorithme ForceAtlas - A gauche par genres littéraires et à droite par types d'acteurs - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

La mise en évidence de la présence forte de « Blogs ProAm » dans cet entrelacement nous alerte sur leur rôle potentiel. Mais c'est l'analyse qualitative du contenu de ces blogs qui nous permet de constater que la plupart des "Blogs ProAm" produisent des récits et des illustrations dans les deux genres littéraires, en BD et en littérature jeunesse, ce dernier genre littéraire étant composé pour une importante part d'un travail d'illustration. Or ce trait saillant, l'importance de l'illustration dans la littérature jeunesse, ne pouvait pas a priori être considéré comme pertinent pour comprendre les relations entre ces deux clusters. Un expert aurait pu le savoir mais n'aurait pas nécessairement anticipé l'effet de ce trait sur les relations entre communautés. D'un autre côté, la catégorie qui fait l'objet d'un tag général (les blogs ProAm) paraissait significative à une échelle large mais elle n'avait pas a priori d'importance dans la structure du graphe, aucune hypothèse a priori ne soutenait sa présence, si ce n'est un souci de produire une description empirique plus fine. C'est uniquement par cette démarche expérimentale faite de déplacements de points de vue successifs que l'on a pu faire *émerger* le rôle joué par ces nœuds, de façon inductive. Et c'est par la connaissance experte du domaine que l'on peut alors effectuer un retour sur les sites que représentent ces nœuds puis sur les acteurs qui publient ces sites : les professionnels de l'illustration de jeunesse sont aussi des dessinateurs de BD et ils ont un statut de marginal sécant (Crozier, Friedberg, 1977) pour les deux communautés.

Sur le plan de la visualisation, la coexistence de deux divisions « orthogonales » est un effet de complexité typique dans un réseau de ce type, dû à la superposition de multiples facteurs dans l'établissement de liens hypertextes. Les deux divisions sont « orthogonales » visuellement, mais aussi métaphoriquement parce qu'elles s'opposent effectivement dans la

topologie du réseau. La comparaison des deux images nous montre à la fois l'intérêt de projeter des catégories et à la fois le compromis établi par la spatialisation, où l'œil expert peut deviner la double division et le double attachement, vertical ou horizontal. Maintenant que les catégories nous ont montré la voie, nous pouvons orienter notre regard pour « lire » deux groupes verticaux, deux groupes horizontaux, ou encore quatre groupes en carré. Cette superposition de niveaux de lecture est généralement responsable de l'effet « hairball[5] », où le compromis de l'algorithme empêche une séparation nette des clusters, et on voit bien ici comment les catégories permettent de dénouer efficacement la superposition d'attachements en renforçant l'un au détriment des autres.

Une fois découverte la puissance analytique de la catégorie des types d'acteurs, rien n'empêche de tester sur tout le graphe sa pertinence et son effet sur la clusterisation de tout le domaine. La distinction Blogs ProAm et Blogs est-elle pertinente au-delà du cas évoqué ou peut-elle se substituer ou tout au moins compléter l'analyse du « tout » selon une partition nouvelle ? En effet, si l'on applique au graphe des symboles et couleurs en fonction des acteurs et non des genres littéraires, autrement dit si l'on modifie notre grille de lecture du graphe, ce dernier offre un autre visage du web littéraire. On constate alors que la partie gauche du graphe est quasiment exclusivement composée de blogs de lecteurs (ronds blancs) alors que le reste du graphe au centre et à droite est composé d'acteurs beaucoup plus hétérogènes. On peut également repérer en bas à droite un ensemble de blogs « Pro-am » très liés entre eux (carrés noir) dont on a pu détecter qu'ils formaient le ciment des clusters BD et Jeunesse situés dans cette région du graphe.

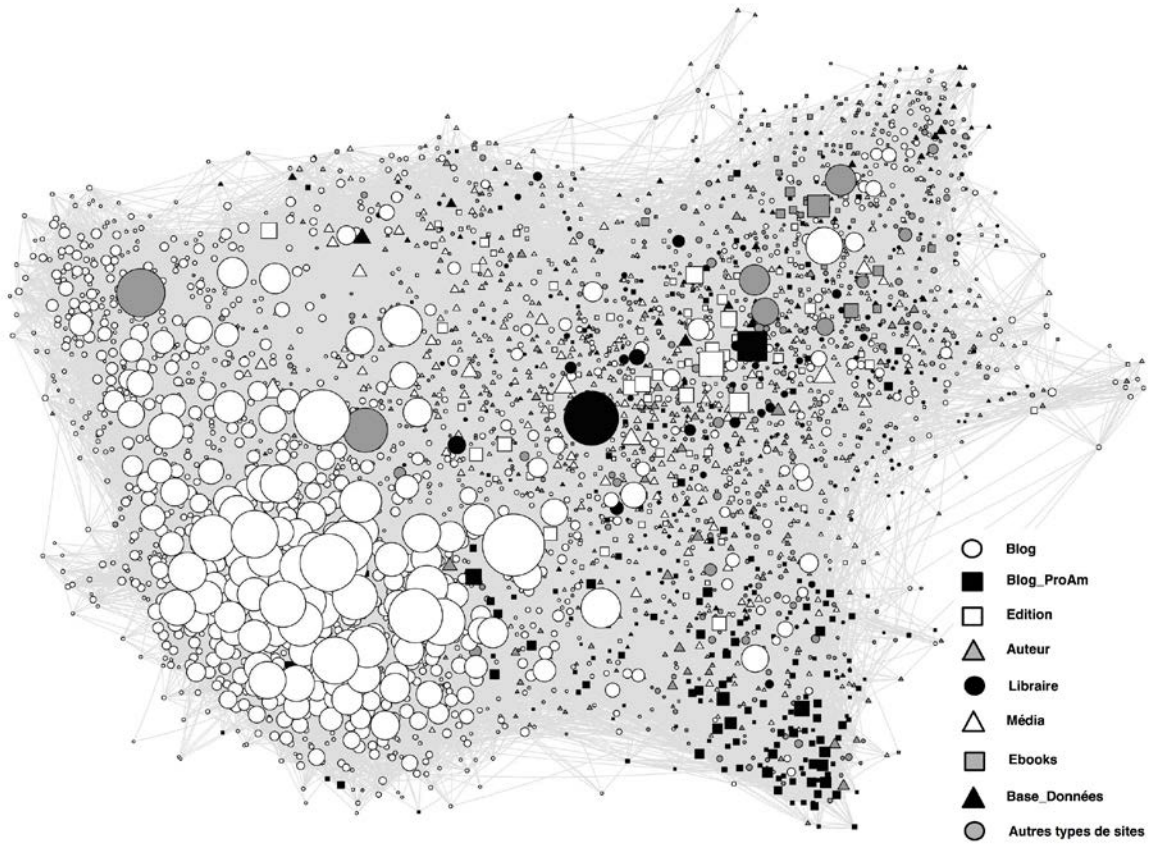


Figure 9 - Web littéraire en France en 2010 par types acteurs - Algorithme de spatialisation : Force Atlas - La taille des nœuds est fonction de leur degré (nombre de liens hypertextes partagés avec les autres noeuds)

[5] Un réseau dense et uniforme pour lequel l'oeil n'est pas capable de distinguer des clusters.

CONCLUSION

On peut comprendre aisément la fascination exercée par la continuité apparente fournie par les dispositifs numériques entre le tout et ses parties, notamment parce qu'elle permet de soutenir la critique du monde en plusieurs niveaux comme B. Latour l'a proposée. C'est pourquoi la métaphore du zoom cinéma a constitué une ressource rhétorique fréquemment utilisée. Elle avait l'avantage de contrecarrer les oppositions macro-micro qui organisent une division du travail sans grand intérêt au sein des sciences sociales. Le compte-rendu que nous venons d'effectuer d'une exploration des données issues d'un crawl du web du livre francophone a été réalisé délibérément sur le mode d'un récit de sociologie des sciences ou encore selon un principe ethnométhodologique : nous avons rendu compte de nos propres impasses, de nos tentatives, de nos choix successifs effectués avec toutes les hésitations et les incertitudes propres à « la science en train de se faire ». Ce faisant, nous ne contestons pas la nécessité de fournir des comptes-rendus moins détaillés qui donnent la vision synthétique des résultats. Nous souhaitons avant tout montrer comment les conventions d'exploitation des méthodes de graphe dans les sciences sociales restent encore à établir pour garantir la qualité nécessaire aux résultats, de la même manière qu'elles font l'objet de discussions dans d'autres disciplines scientifiques. Il n'est pas possible de se contenter des métaphores de la copie carbone et du zoom pour rendre compte du travail effectivement réalisé lors de l'exploration des graphes. Notons bien qu'il ne s'agit que d'un domaine des humanités numériques, d'une technique particulière, qui privilégie de fait une vision statique des réseaux à base de liens hypertextes, qui produit ce que nous avons appelé ici des topiques (dans le cadre d'une topologie). D'autres travaux de suivi des vibrations sous forme de *timelines* et d'arbres génétiques sont aussi nécessaires pour comprendre d'autres dimensions des entités produites par les plates-formes numériques (Boullier, 2015). Le cas des topologies et des visualisations est cependant intéressant pour montrer la force des images qui ressemblent à des cartes, et auxquelles on applique des métaphores visuelles inadaptées mais issues de conventions plus anciennes. Nous avons voulu raconter l'histoire des changements successifs de points de vue qui permet de nous détacher de l'immédiateté supposée de l'effet zoom. Explorer les parties du tout constitué par notre graphe d'origine nous a obligé à abandonner le premier récit basé sur des distances visuelles « évidentes » pour nous focaliser sur les zones à problème, là où les catégories et la visualisation ne produisent pas d'équivalence ni de frontière nette. Se focaliser sur l'entrelacement de clusters de la BD et de la Jeunesse n'a pas consisté à zoomer, à changer d'échelle, mais à recalculer cette zone en faisant varier plusieurs paramètres pour comprendre l'anomalie : faire varier les algorithmes, les connexions avec le tout du graphe, et finalement faire varier les catégories qui présidaient à la première clusterisation. Et enfin, il a fallu retourner vers les sites en question individuellement pour comprendre ce rôle particulier de « bridge » joué par certains nœuds des illustrateurs et dessinateurs de BD et Jeunesse, ce qui constitue une forme de travail d'interprétation que l'on dit qualitatif par observation du matériel publié et usage de notre connaissance experte du domaine. Dans cette démarche, il n'est pas question de « naviguer » dans les données, de « plonger » dans les détails ou de « zoomer », mais en permanence de *déplacer le point de vue* et d'y adapter les calculs. L'algorithme est au service de notre investigation et des solutions techniques différentes donnent toujours des résultats différents. Dans la préhistoire des sciences du web dans laquelle nous nous trouvons, il serait nécessaire de parvenir à produire des protocoles d'exploration et de tests de robustesse qui deviennent des prérequis pour toute exploitation de ces données. Notre compte-rendu espère y avoir contribué.

BIBLIOGRAPHIE

AKRICH M., CALLON M., LATOUR B. (2006), *Sociologie de la traduction. Textes fondateurs*, Paris, Presses des Mines de Paris.

BASTIAN M., HEYMANN S., JACOMY M. (2009). « Gephi: an open source software for exploring and manipulating networks ». *International AAAI Conference on Weblogs and Social Media*.

BERTIN J. (1973), *Sémiologie graphique: les diagrammes, les réseaux, les cartes*, Paris, EHESS.

BOULLIER D. (2015), « Vie et mort des sciences sociales avec le Big Data », *Socio*, n°4, pp. 19-37.

BOULLIER D. (2015), « Les sciences sociales face aux traces du Big Data. Société, Opinion ou Vibrations », *Revue Française de Science Politique*, n°5-6, Volume 65, pp. 805-830.

BOURDIEU, P. (1979), *La distinction. Critique sociale du jugement*, Paris: Editions de Minuit.

BLONDEL V., GUILLAUME J.L., LAMBIOTTE R., LEFEBVRE E. (2008), « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics, Theory and Experiment* 2008 (10), P1000.

BURT, R. S. (2005). Brokerage and Closure: An Introduction to Social Capital. Oxford University Press

CROZIER M., FRIEDBERG E. (1977), *L'acteur et le système, Les contraintes de l'action collective*, Sociologie politique, Seuil.

DESROSIERES A. (2014), *Prouver et gouverner : Une analyse politique des statistiques publiques*, La Découverte, 284 p. (Recueil posthume de textes choisis et rassemblés par Emmanuel Didier)

EYMARD-DUVERAY F., FAVEREAU O., ORLEAN A., SALAIS R. et THEVENOT L. (2004), « L'économie des conventions ou le temps de la réunification dans les sciences sociales », *problèmes économiques*, n° 2838, Janvier 2004, La Documentation française, Paris.

JACOMY, M., VENTURINI, T., HEYMANN, S., & BASTIAN, M. (2014), « ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software ». *PloS One*, 9(6), e98679. doi:10.1371/journal.pone.0098679

LAMBIOTTE R., DELVENNE J.C., BARAHONA M. (2009), Laplacian Dynamics and Multiscale Modular Structure in Networks, arXiv preprint arXiv:0812.1770

LATOUR B., JENSEN B., VENTURINI T., GRAUWIN S., BOULLIER D. (2012), « The Whole is Always Smaller Than Its Parts'. A Digital Test of Gabriel Tarde's monads », *British Journal of Sociology*, Volume 63, Issue 4, pages 590–615.

LATOUR B., FABBRI P. (1977), « La rhétorique de la science », *Actes de la recherche en sciences sociales*, Année 1977, Volume 13, Numéro 1 p. 81 – 95.

LATOUR B. (2012), *Enquête sur les modes d'existence. Une anthropologie des modernes,* Paris, La Découverte.

LEADBEATER C., MILLER P. (2004), *The Pro-Am Revolution: How Enthusiasts are Changing Our Society and Economy*, Demos, 2004.

LE BECHEC M., CREPEL M., BOULLIER D. (2014), « Modes de circulation du livre sur les réseaux numériques », *Études de communication*, n°43, pp.129-144.

LE BECHEC M., CREPEL M., BOULLIER D., (2016), *Le livre-échange*, à paraitre.

LECOLINET E., POOK S. (2002), « Interfaces zoomables et control menus, Techniques focus+contexte pour la navigation interactive dans les bases de données », *Les Cahiers du numérique*, 2002/3 Vol. 3 | pages 191 à 210.

LEVY, J. et LUSSAULT, M., (2003, réédition 2013), *Dictionnaire de la géographie et de l'espace des sociétés*, Paris, Belin, 1034 p.

LEVY, J. MAITRE O., ROMANY T. (2016), « Rebattre les cartes. Topographie et topologie dans la cartographie contemporaine », *Réseaux*, n°195.

ROGERS R. (2013), *Digital Methods*, Cambridge, MA: MIT Press.

TARDE G. (1893), *Monadologie et sociologie*, Paris, Alcan, 55 p.