L’opinion autorisée

Requalification communautaire de l’espace social et techniques d’échantillonnage sur le web

Baptiste Kotras

LATTS/Université Paris-Est

bkotras [at] gmail [point] com

Résumé

Cet article vise à comprendre comment les spécialistes de l’analyse d’opinion sur le web remédient à l’indétermination sociodémographique généralisée des traces conversationnelles qui constituent leur matériau. En l’absence de toute variable classique, il leur faut en effet trouver des formes collectives alternatives pour incarner et imputer les tendances d’opinion qu’ils analysent. A travers le cas de la société Linkfluence (Paris, France), nous montrons comment est élaboré un paradigme alternatif, fondé sur le concept de « communautés en ligne », par lequel les sites sont catégorisés en fonction de leur profil hypertexte et de leur thématique. Dès lors, la sélection des communautés thématiques liées à un sujet donné donne à voir un nouveau mode d’échantillonnage délibérément asymétrique, où seuls ont la parole les sites mobilisés et influents.

Mots-clés : opinion, web, communauté, échantillonnage

Introduction

Depuis les années 2000 a émergé un secteur économique en forte croissance, le *social media analysis*, autour de la disponibilité de données conversationnelles massives et publiques émises sur le web « social », c'est-à-dire essentiellement les blogs, forums et sites de réseaux sociaux. Aussi bien en Europe qu'en Amérique du Nord, des entreprises spécialisées dans la collecte, le traitement et l'analyse de ces données proposent des prestations visant à mesurer l'opinion sur le web, à des fins diverses : gestion de la réputation et de l'image d'une marque, pilotage des relations client, évaluation des campagnes en ligne, études de tendances et de marché ; tous ces besoins mobilisent une technologie plus ou moins élaborée, permettant le traitement de données éparses. Equipés d'outils logiciels adaptés, ces acteurs se proposent donc d'agréger et d'analyser les milliers de messages publiés chaque jour, pour en inférer des tendances qu'ils livrent aux départements de la communication, du marketing, voire des relations consommateur de leurs clients – de « grands comptes », le plus souvent. Outre leur faible coût, ces méthodes présentent selon leurs promoteurs deux avantages essentiels : d'une part, veiller l'opinion en temps réel, ou presque ; mais également, et c'est là leur promesse la plus ambitieuse, recueillir des expressions non sollicitées, donc « spontanées », évitant de ce fait l'écueil de l'artificialité, maintes fois reprochée aux sondages d'opinion par la sociologie critique depuis l'article fondateur de P. Bourdieu (1973).

De façon corollaire à leur « spontanéité », ces données présentent néanmoins l'inconvénient d'être presque totalement détachées de leur contexte d'énonciation : les formats de la prise de parole en ligne laissant encore une large place à l'anonymat, il est impossible de connaître les caractéristiques sociodémographiques des individus produisant les opinions collectées. Cet état de fait pose deux types de problème à ces prestataires. Premièrement, les prestations de type « étude » – qui nous intéresseront ici – supposent d'analyser en profondeur des tendances, des jugements émis sur des produits ou des enjeux spécifiques, ce qui nécessite une lecture humaine et un codage détaillé des *verbatim*, chose impossible à faire sur l'intégralité des volumes recueillis. Pour travailler sur des volumes traitables, se pose donc la question de l'échantillonnage des opinions ; or, l'absence des variables habituelles ne permet pas la construction d'échantillons par quotas. Ensuite, l'indétermination sociodémographique généralisée empêche toute description stratifiée de l'opinion, où celle-ci est associée à des sous-composantes de la population générale, ce qui rend *a priori* malaisé le dialogue avec les clients de cette expertise, habitués à décrire leurs consommateurs en termes de cibles sociodémographiques.

Renoncer à cet appareillage pour exploiter les données conversationnelles a donc un coût pour ces acteurs, celui de devoir reconstruire l'espace social pour échantillonner, puis incarner les opinions sur le web. Dans sa *Fabrique de l'opinion*, L. Blondiaux montre en effet comment les premiers sondeurs avaient tiré parti des acquis alors récents de la statistique inférentielle pour construire, sur la base des variables codées par les bureaucraties fédérales, des échantillons dits « représentatifs », permettant de tester la distribution des opinions sur quelques milliers de personnes *« judicieusement sélectionnées »*[1]. En faisant admettre la partie pour le tout, cette *« Amérique en miniature »* (Blondiaux, 1998, p.173) dotait les sondeurs d'un puissant ressort de justification : dans l'isoloir comme dans l'échantillon, incarnation de l'ensemble du corps social, un homme égalait une voix. Cette analogie démocratique, qui revêt le sondage d'une dimension politique et morale, a selon l'auteur été un facteur essentiel de son adoption ultérieure dans les mondes sociaux de la presse et de l'université[2]. De nos jours, la segmentation

[1] Par opposition avec les dizaines de milliers de répondants aux « votes de paille » organisés jusqu'alors par le *Literary Digest*. Sur les votes de paille, voir Herbst (1995).

[2] Sur la promotion par P.F. Lazarsfeld du sondage comme instrument des sciences sociales, voir également Blondiaux (1990) et Pollak (1979).

sociodémographique de la population est d'autant plus puissante socialement qu'elle est inscrite au cœur de nombreux dispositifs quotidiennement mobilisés, tels que la mesure d'audience (Méadel, 2004), les études de marché ou l'analyse électorale. Parmi les technologies qui équipent notre perception de la société, sa centralité donne consistance aux entités mobilisées pour incarner les tendances d'opinion : les « périurbains », les « classes populaires », les « cadres » et ainsi de suite, prennent vie à travers les énoncés qui les décrivent (Thévenot et Desrosières, 2002, Porter, 1995). Alors que le lien fondateur entre l'opinion et son incarnation sociodémographique semble définitivement rompu, les acteurs du *social media analysis* doivent reconstruire un sujet aux énoncés qui décrivent l'opinion, c'est-à-dire des groupes, collectifs et formes sociales auxquels imputer les tendances détectées. L'enjeu est pour eux de reconstruire un modèle d'échantillonnage qui réponde simultanément à la double question : *« qui écoute-t-on ? »*, et *« qui parle ? »* ; autrement dit : sur quels critères inclut-on ou non un site donné dans le corpus étudié ? Quels groupes portent les opinions que l'on souhaite analyser ?

C'est la façon dont les acteurs élaborent des réponses à ces questions, tout à la fois très concrètes et porteuses de choix éminemment politiques, qui nous intéressera dans cet article. Quelles techniques alternatives de quantification des opinions sont développées sur le web, et de quel type d'objets équipent-elles aussi bien l'analyse que l'action ? Nous reprenons en effet l'hypothèse d'A. Desrosières selon laquelle l'espace public *« n'est pas seulement une idée performative, parfois vague, mais un espace historiquement et techniquement structuré et limité »* (Desrosières, 1992, p.131). Dans cette perspective, il nous faudra comprendre l'épistémologie de ce nouveau mode de quantification de l'opinion, c'est-à-dire les modalités de construction d'une connaissance collectivement validée (Knorr-Cetina, 1999), et ses implications sur le type de connaissance produit. Nous nous baserons pour cela sur la distinction utilement réalisée par A. Desrosières entre les deux types d'opération nécessaires à la quantification et chronologiquement successives : convenir et mesurer. Il apparaît en effet qu'en l'absence de variables sociodémographiques, et avant de « mesurer » quelque tendance que ce soit, nos acteurs doivent auparavant construire des conventions permettant la commensuration ; « *avant de tirer des boules dans une urne, il faut avoir convenu du choix des boules à y inclure, et de la nomenclature de leurs couleurs* » (Desrosières, 2008, p.7) . Autrement dit, ce qui nous intéresse est autant le travail de défrichage et de normalisation de l'espace social en ligne réalisé par les acteurs, que son inscription dans les résultats ainsi produits.

Nous nous centrerons pour cela sur le cas de la société Linkfluence, fondée en 2006 par quatre jeunes diplômés de l'Université de Technologie de Compiègne. Ainsi que nous le verrons, son inscription originelle dans le monde académique a poussé ses membres à produire un certain nombre de textes (communications, articles) décrivant leur vision du web social, nous donnant ainsi un accès unique aux justifications théoriques de l'épistémologie qui fonde encore aujourd'hui leur étude de l'opinion sur le web. En particulier, l'article intitulé « Two visions of the web, from globality to localities » (Jacomy, Fouetillou et Pfaender, 2006) sera au centre de notre démonstration. Nous avons en outre conduit une dizaine d'entretiens semi-directifs avec les employés de la société[3], et réalisé une ethnographie de deux mois dans leurs locaux, constatant ainsi la façon dont ces théorisations ont ensuite été transposées dans un modèle commercial de services.

A partir de leurs productions théoriques, nous montrerons dans un premier temps comment les futurs fondateurs de Linkfluence développent – dans un contexte alors universitaire – une description du web basée sur la requalification du lien hypertexte comme artefact de l'autorité et de l'affinité en ligne. Il s'agira dans un second temps de montrer comment ce référentiel fonde aujourd'hui les prestations de l'entreprise Linkfluence, en autorisant la construction

[3] Fondateurs, directeurs d'étude, chargés d'étude, ingénieurs.

d'échantillons qui incarnent l'opinion dans des totalités stabilisées. Désignées sous le nom de communautés, celles-ci donnent à voir un modèle dynamique et hiérarchisé de l'opinion.

Du graphe à la communauté, histoire d'un assemblage épistémologique

L'objet « communauté », tel que construit par Linkfluence, repose sur l'investissement du lien hypertexte d'un double signifié, qui en fait à la fois la mesure de l'autorité mais également de l'affinité entre les sites. Nous détaillons ce travail de théorisation et ses filiations conceptuelles, puis nous montrons comment cette requalification est fondée par le déploiement d'une épreuve empirique.

Cartographier le web, ou la construction visuelle du graphe

Nous nous basons ici sur les témoignages, recueillis en entretien, et les productions théoriques d'un groupe d'acteurs qui émerge en 2003 autour d'un projet de recherche de l'Université de Technologie de Compiègne. Dans ce groupe, des étudiants en master, parmi lesquels les quatre futurs fondateurs de Linkfluence, collaborent avec des ingénieurs et des enseignants-chercheurs issus de l'équipe COSTECH pour concevoir un outil expérimental nommé TARENTe[4]. Fondé sur les travaux de chercheurs tels que J. Kleinberg (1999), S. Chakrabarti (1999) ou A.-L. Barabasi (2002), ce logiciel combine les technologies du *crawling*[5], de l'analyse sémantique et de la visualisation de données avec l'objectif d'identifier des « agrégats du web » basés sur une corrélation empirique entre proximité sémantique et hypertexte sur le web (Ghitalla et al. 2004). Autrement dit, ils cherchent à montrer que les pages qui se citent abordent les mêmes sujets.

Ce groupe d'étudiants et de chercheurs puise dans deux courants identifiés par J. De Maeyer (2013) en matière de *link studies* : l'analyse des réseaux complexes et l'étude des liens hypertextes comme objet social. Leur référence principale est Jon Kleinberg, universitaire bien connu dans l'histoire du web de par l'inspiration qu'il a constitué, grâce à son algorithme HITS, pour le PageRank de S. Brin et L. Page (Cardon, 2013). Son apport essentiel réside dans la construction d'une mesure standardisée de l'autorité sur le web, fondée sur le ratio entre liens entrants et sortants pour une page donnée. Avec le succès de ses travaux et leur appropriation par les moteurs de recherche, le lien hypertexte est doté d'un signifié univoque, l'autorité. Les futurs fondateurs de Linkfluence s'inscrivent dans cette filiation, en ce qu'ils mobilisent eux aussi la théorie des graphes pour modéliser le web, validant le lien hypertexte comme artefact de l'ordonnancement du web. Cependant, ils critiquent dans un article de 2006 (Jacomy et al. 2006) les méthodes des moteurs de recherche, reprochant à la forme « liste » de diluer le sens des résultats, en ne donnant à voir qu'une succession d'items indifférenciés et abstraits de leur environnement hypertextuel. Cet état de fait, outre qu'il découragerait l'exploration, écraserait les *« hiérarchies locales »* d'un web décrit par Kleinberg lui-même comme fortement décentralisé. Un nouveau mode d'organisation de l'information est donc jugé indispensable[6].

Le groupe de Compiègne se fixe donc pour objectif de restituer avec une plus grande précision les structures hypertextuelles qui caractérisent le web et que rendent invisibles les moteurs de recherche. Pour cela, à l'instar de plusieurs travaux de l'époque (Adamic et Glance, 2005 ;

[4] Tous ne participeront pas à la fondation de Linkfluence, fin 2006. Parmi ceux qui n'intègreront pas la start-up, on trouve notamment Franck Ghitalla, linguiste de l'UTC et coordinateur du projet, et Mathieu Jacomy, concepteur du logiciel libre Gephi.

[5] Exploration et indexation automatisées de pages web via les liens hypertextes.

[6] D'autres projets de même type existent à l'époque. F. Ghitalla fait notamment référence à « Geographies of Cyberspace », mené par le géographe Martin Dodge (University Colllege London).

Rogers et Marres, 2000), ils défendent un autre mode de spatialisation, dit « cartographique », consistant à représenter visuellement l'ensemble des liens échangés par les sites, sous la forme d'un graphe où les nœuds sont les sites web et les arcs qui les relient sont les liens hypertexte. Ainsi, les chercheurs de l'UTC veulent accéder à un niveau de questionnement « *synoptique* » et non « *contraint par un sens de lecture* » (Jacomy et al. p.3), à même de démontrer l'existence d'agrégats hypertextes et thématiques, nommés par la suite « communautés ».

Le passage au niveau graphique et explicite d'un objet jusque-là purement mathématique en transforme le sens. Ce qui était pour les moteurs de recherche un corpus documentaire à ordonner de manière strictement procédurale et heuristique va devenir un territoire de relations sociales, parcouru de clivages et d'affinités qu'il convient d'objectiver. En effet, le fait de rendre appréhendable sur un plan visuel les différents *clusters* qui composent le web permet à nos acteurs de poser la question du « sens » de ces grappes de sites fortement interconnectés, qui les conduit à investir le lien hypertexte d'un nouveau signifié affinitaire : en montrant que la proximité thématique recoupe la proximité hypertexte, ils vont doter l'artefact du lien hypertexte d'une ontologie double, à la fois mesure de l'autorité et de l'affinité entre deux sites.

Le débat politique comme dispositif de preuve : le cas de l'avortement

Pour forger cette équivalence entre lien hypertexte et affinité thématique, nos acteurs montent un dispositif de preuve qu'ils présentent dans l'article précédemment évoqué, dont le point de départ consiste à indexer les cinquante premiers résultats fournis par Google pour la requête « *abortion* »[7]. Ils codent ces 50 pages web en fonction de la ligne éditoriale, telle que revendiquée par le site lui-même, quant au droit à l'avortement (*prolife*, *prochoice*, neutre)[8]. Un logiciel de *crawling* répertorie ensuite l'ensemble des liens hypertexte existant entre ces pages, permettant la projection du réseau sous forme cartographique. Dans le dispositif visuel et mathématique, deux éléments sont essentiels pour faire apparaître la corrélation que veulent montrer les auteurs : d'une part, le code de couleurs attribuées aux sites en fonction de la position qu'ils revendiquent ; d'autre part, l'algorithme de spatialisation qui place en proximité les sites partageant des liens. De cette façon, le dispositif intègre le questionnement de nos acteurs et rend possible la conclusion à laquelle ils souhaitent parvenir.

[7] *« A l'exclusion des sites non exclusivement dédiés à l'avortement* (www.washingtonpost.com *par exemple*) *»* (p.4). On notera cependant la présence – manifestement erronée – du Wikipédia anglophone ainsi que du site de la BBC.

[8] Les auteurs précisent que deux autres catégories ont « émergé » du codage : des sites « catholiques » non explicitement positionnés sur le droit à l'avortement, et des sites *« proposant des ressources pour passer le moment de l'après-avortement »* (p.3), soit un total de cinq catégories.

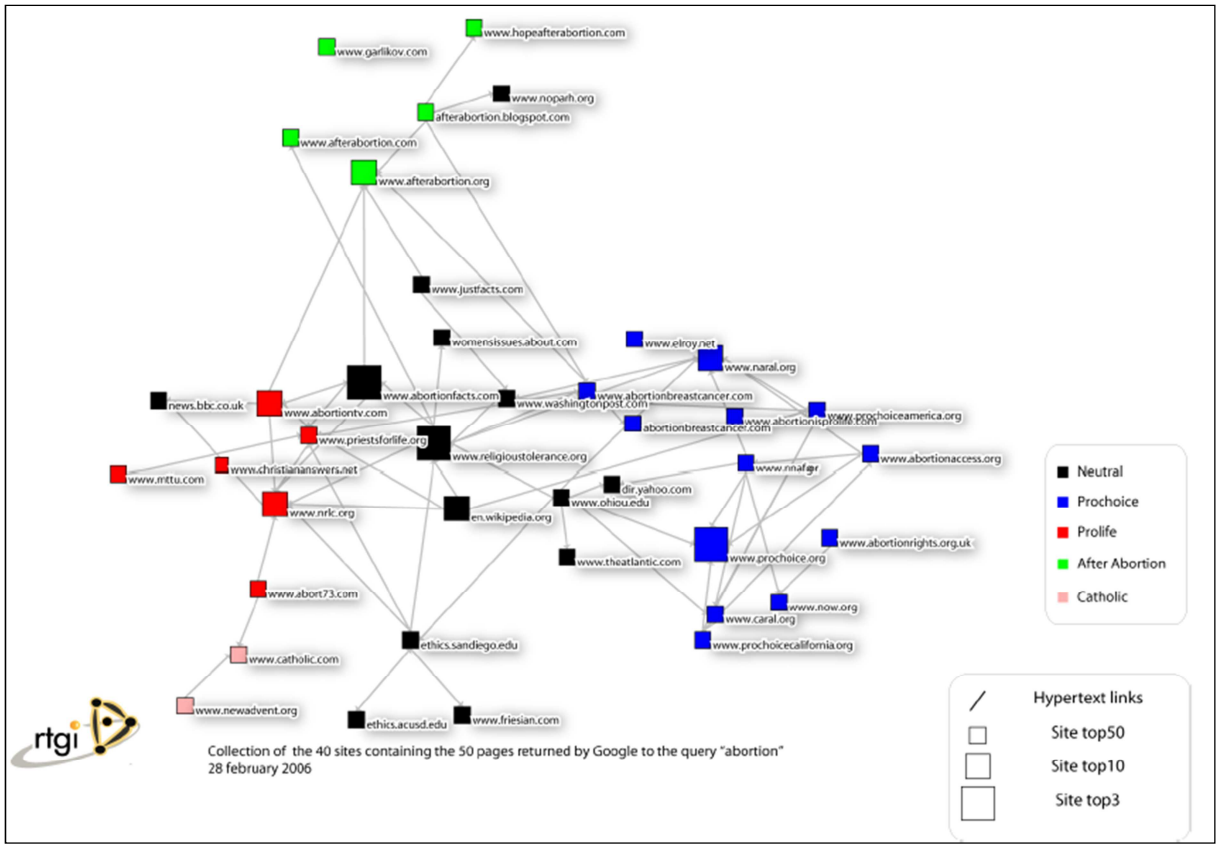

Fig.1 : Cartographie « avortement » - RTGI, 2006 (Jacomy et al. 2006)

Un simple coup d'œil au résultat montre le succès rencontré par l'expérience : le graphe ainsi projeté donne à voir des *clusters* bien différenciés, à la fois en termes de connectivité hypertexte, mais aussi de positionnement éditorial, que matérialise la distribution des couleurs. Les auteurs montrent par cette construction l'homogénéité politique des différentes zones du graphe, qu'ils interprètent alors comme une preuve de la dimension « élective » ou affinitaire du lien hypertexte : les sites partageant une même position se citent de façon préférentielle. Le lien hypertexte n'est donc plus seulement un vecteur d'autorité, mais également d'affinité :

> « Alors que comme nous l'avons vu, la structure de la liste rassemblait toutes les ressources sous une même catégorie, les distances permettent ici aux ressources de se distinguer les unes des autres […]. Ce type de représentation, qui révèle la structure hypertextuelle véritable des sites présents sur la carte, permet des lectures expertes qui se fondent sur la considération du lien hypertexte comme manifestation singulière d'un lien social électif[9] » (Jacomy et al. p.4).

Comme montré par Daston et Galison (1992), la représentation graphique confère ici à la description de l'objet son caractère objectif, puisqu'elle permet de « révéler » les propriétés invisibles du réseau des sites étudiés. Les auteurs se dotent ainsi d' « *inscriptions en deux dimensions, superposables et combinables* » qui leur permettent de « *maîtriser* [leur] *monde* » (Latour, 1993), grâce à un dispositif qui décrit le web en même temps qu'il lui donne un sens. Dans les interprétations construites par les chercheurs de Compiègne, le caractère premier du « lien électif » se mesure à l'aune de sa mise en concurrence avec le sens explicitement produit

[9] C'est ce même concept de « lien social électif » qui constitue le cœur théorique de la thèse de doctorat (inachevée) de l'un des fondateurs de Linkfluence, co-auteur du papier en question : *Le lien hypertexte comme lien social électif, vers une dynamique de constitution du web*, thèse en Sciences de l'Information et de la Communication, sous la direction de Dominique Boullier, Université de Technologie de Compiègne.

par les sites eux-mêmes, ainsi qu'en atteste l'analyse détaillée faite du site www.abortionfacts.com :

> « Ce site se dit être un site neutre, raison pour laquelle il est noir sur la carte. Maintenant, une lecture en précision de la carte nous apprend les choses suivantes : son positionnement est à la frontière entre la zone des sites neutres et la zone de sites pro-life ; il ne possède aucun lien sortant vers des sites neutres ; il possède trois liens sortants vers des sites pro-life. Si ce site n'est pas situé dans la zone des pro-life, c'est uniquement par la présence d'un lien de www.religioustolerance.org pointant vers lui. Ce lien est en quelque sorte sa seule 'attache' au territoire des sites neutres, ses autres attaches étant pro-life » (Jacomy et al. p.4).

Par l'étude de la *« structure hypertexte véritable »*, il devient donc possible d'objectiver les structures communautaires, qui rassemblent les sites web jusques et y compris contre leur autodéfinition, comme c'est le cas avec ce site. Malgré une position revendiquée comme « neutre », celui-ci entretient des liens nombreux avec la sphère anti-avortement, liens que seule la carte permet de révéler et qui accrédite l'idée que sa « véritable » appartenance ne serait pas celle qu'il affiche. Dans cette perspective, le sens enfermé dans l'acte de citation prime sur le sens explicitement produit par les sites quant à leur ligne éditoriale. Plus que par leur identité revendiquée, c'est leurs actions et le sens que celles-ci contiennent qui permet à Linkfluence d'objectiver les structures communautaires.

Cette expérience est cruciale en ce qu'elle fonde empiriquement un modèle de connaissance, qui délaisse les catégorisations *a priori* : d'une part, l'indétermination sociodémographique générale des locuteurs sur le web empêche de les rattacher à des appartenances de type macrosocial ; d'autre part, ainsi que nous le montre l'exemple ci-dessus, même l'alignement revendiqué explicitement par un site n'est pas le paramètre essentiel de l'appartenance communautaire, puisqu'il peut être relativisé par l'attention portée à son profil de connexion. C'est donc bien l'acte de citation plus que la caractérisation *a priori* – même revendiquée par le site – qui permet d'objectiver des communautés. L'espace social en ligne pour Linkfluence se révèle comme une constellation de groupes affinitaires que rien ne fait tenir ensemble si ce n'est la répétition d'actes de citation mutuelle. L'expérience cartographique déployée sur l'avortement permet d'en valider le caractère *« non-aléatoire »* (Rogers et Marres, 2000, p.145), et donc de les interpréter comme la manifestation d'une affinité fondée sur un projet éditorial commun, dans les thématiques comme dans l'orientation avec laquelle elles sont traitées. Le lien hypertexte se retrouve requalifié d'un signifié affinitaire, sans que disparaisse sa fonction de vecteur d'autorité, précédemment établie par Kleinberg. On a vu à ce titre la préoccupation de nos acteurs pour la restitution des « hiérarchies locales » : chaque communauté est définie par un projet commun, mais également par un écosystème de citations qui détermine son centre et ses périphéries, et permet d'identifier les détenteurs de l'autorité locale. Ce point sera déterminant par la suite. Notons enfin que cette expérience déployée sur l'avortement est renouvelée avec une médiatisation accrue à propos du référendum sur le Traité Constitutionnel Européen (2005)[10]. Le « oui » et le « non » formant là encore deux clusters homogènes, la qualification affinitaire du lien hypertexte s'en trouve validée. Fin 2006, la société Linkfluence est créée.

La mise en actes du paradigme communautaire : échantillonner, incarner et cartographier les opinions

Il s'agit maintenant de comprendre comment le modèle communautaire permet à Linkfluence de fournir des résultats qui décrivent l'opinion sur le web. Pour cela, nous avançons de plusieurs années pour nous concentrer sur l'offre commerciale aujourd'hui stabilisée de Linkfluence. Nous

[10] Cette expérience donnera aussi lieu trois ans plus tard à une publication dans la revue *Réseaux*. Voir Fouetillou, G. (2008).

montrons comment sont construits des échantillons délibérément asymétriques du web social, puis comment ce mode d'échantillonnage permet l'incarnation de tendances par des formes collectives ; et enfin, comment l'objet cartographique décrit la circulation des opinions et suggère des pistes pour l'action de communication[11].

Généraliser le modèle, circonscrire le web : l'échantillonnage par affinité et par influence

Les deux expériences fondatrices du modèle conçu par Linkfluence, respectivement déployées sur l'avortement et sur le TCE, ont en commun de prendre pour objet des thématiques politiquement clivées, où les deux positions « pour » et « contre » sont explicitement définies par les sites eux-mêmes non seulement comme l'enjeu de leur pratique éditoriale, mais également de leur sociabilité hyperliée (Badouard, 2013). L'inscription de ce modèle dans le cadre marchand d'une offre de services standardisée passe par l'extension et la généralisation de ces expériences locales, afin de proposer dans un cadre marchand des échantillons pertinents du web social. La double qualification du lien hypertexte est mise à profit pour construire un échantillonnage à la fois par affinité et par autorité.

Pour étendre le modèle des communautés affinitaires, construit sur le cas de l'avortement, Linkfluence considère comme équivalents ce qui était dans l'exemple précédent une « cause », au sens militant, et l'expression plus triviale d'un intérêt partagé entre plusieurs sites. Ce faisant, les *crawlers* tendent à montrer le même type de recoupement entre affinité thématique et hypertexte, les sites et blogs partageant un intérêt commun étant amenés à se citer mutuellement, sans pour autant qu'ils ne se définissent subjectivement en tant que collectif. Le directeur de l'innovation explique le choix réalisé à l'époque : *« est-ce qu'il faut se dire communauté pour faire communauté ? Nous, clairement, on était dans une définition relationnelle et émergente qui ne nécessitait pas ce phénomène de reconnaissance »* [12].

Dès lors que tout ensemble de sites web dédiés à un même objet peut être décrit en termes communautaires, Linkfluence propose alors un référentiel général, unique et préexistant dans lequel le web social vient prendre place, naturalisant des « communautés en ligne » définies *ex post*. A partir de la création de l'entreprise Linkfluence commence donc la constitution par le département « écologie » – spécifiquement dédié à cette tâche – de panels thématisés du web social, construits par l'objectivation des proximités hypertextes et leur recoupement par une proximité thématique (ou *« éditoriale »*). Cette segmentation du web intègre ainsi la structure spontanée de l'expression et des sociabilités en ligne dans une offre marchande de plus en plus « tout-terrain » à mesure qu'augmente le nombre de communautés couvertes par les panels. A l'heure actuelle, Linkfluence a recensé et cartographié plus de 120 communautés pour un total d'environ 15 000 sites et blogs, corpus qui selon l'un des fondateurs *« couvre 80% des demandes de* [leurs] *clients »*[13].

Cette entreprise de normalisation de l'espace social en ligne permet la constitution d'échantillons opérationnels pour des études d'opinion. Dans un premier temps, l'enjeu pour les directeurs d'études consiste à trouver un cadre commun avec leurs clients. Ceux-ci, souvent responsables de la communication ou du marketing au sein de leur entreprise, sont par conséquent habitués à définir leurs consommateurs selon des caractéristiques sociales et démographiques, ce que ne peuvent fournir les outils de Linkfluence. Définir un périmètre d'étude réalisable pour le prestataire et acceptable pour le client n'est donc pas une mince affaire. Pour chaque nouveau

[11] Nous précisons que, bien qu'elle constitue pour les acteurs la finalité du processus d'objectivation, nous laisserons volontairement de côté l'analyse elle-même des opinions émises (tonalité, typologie des arguments), pour nous concentrer sur les effets de l'échantillonnage communautaire sur la mise en sens d'énoncés descriptifs de l'opinion.

[12] Entretien, mars 2013.

[13] Entretien, décembre 2012.

client, débute donc ce qui s'apparente à un cas exemplaire de sociologie de la traduction (Akrich et al. 2006) ; ainsi, lors d'une réunion méthodologie du département, un directeur d'études rappelle :

> « Eux [les clients] raisonnent en termes de cibles sociodémo[graphiques], c'est le langage du marketing. Sur le web, personne ne le parle. Il faut traduire ces objectifs dans la langue du web social. Cette langue, considérez que Linkfluence en a la grammaire, et cette grammaire, ce sont les communautés. » (DE, réunion méthodologique, juin 2013)

Ce travail réflexif de traduction suppose la mise en équivalence des objectifs du client et des communautés thématiques proposées par Linkfluence. A cette fin, c'est bien le centre d'intérêt objectivé d'une communauté hypertexte qui est le critère de son inclusion dans l'échantillon d'une étude. Un client désireux de repositionner sa marque de champagne auprès d'un public de *« playful formalists »*, c'est-à-dire aisé, masculin, *« à la fois respectueux des traditions et innovant »*[14], commandera ainsi après deux réunions de tractations une étude centrée sur les communautés de la voile et des régates. Un sociostyle relativement abstrait, éminemment complexe à circonscrire sur le web, est ici traduit par la référence à une pratique sportive historiquement liée au produit du client ; parce qu'ils s'intéressent à la voile, les sites de ces communautés sont jugés pertinents pour une étude sur le champagne. De même, un chargé d'études évoque comment, confronté à un client cherchant à mesurer le succès d'une campagne publicitaire auprès des *« CSP+ de 15 à 60 ans »*, il doit lui expliquer qu'une telle cible n'est pas identifiable dans le référentiel de Linkfluence. En revanche, la campagne en question étant centrée sur des *« valeurs d'épanouissement »* et de *« dépassement de soi »*, il a convaincu son client de mesurer son succès auprès des communautés du sport, idéologiquement les plus à même de la relayer. Au terme de ce processus actif de traduction, c'est bel et bien l'intérêt manifesté par telle communauté pour une thématique donnée qui justifie son inclusion dans le périmètre d'une étude ; la référence sociodémographique, au passage, a disparu.

Les échantillons ne sont pourtant pas uniquement garantis par le critère thématique ; Linkfluence se targue en effet de proposer une sélection du web social « influent », qu'ils déduisent des caractéristiques épistémologiques du modèle communautaire. Ce parti pris, typique de Linkfluence et revendiqué comme tel sur son marché, vise à prendre acte de l'asymétrie structurelle du web, qui fait que *« 99% des contenus sur le web n'ont aucune audience »*[15]. Un des fondateurs explique : *« nous, on n'était pas dans une logique de moteur de recherche, d'exhaustivité, on voulait travailler uniquement sur l'échantillonnage du web qui façonne les opinions. Donc il fallait faire un travail de sélection »*[16]. Dans cette logique, les développeurs de Linkfluence ont bâti un algorithme de quantification de l'influence, basé sur les échanges de liens hypertexte, produisant pour chaque site un score normalisé entre 1 et 100. La pertinence de l'échantillon, outre la thématique des sites retenus, est également garantie par la sélection du sommet de la hiérarchie de chaque communauté.

Comme le résume notre responsable méthodologique, *« il existe des multitudes de communautés, et chacune a sa hiérarchie interne, et chacune a ses grands et ses petits »*[17]. C'est en s'appuyant sur ces deux dimensions de la communauté – affinitaire et hiérarchisée – que les chargés d'études peuvent bâtir les conventions nécessaires à la délimitation d'échantillons, dont la pertinence est alors doublement garantie. Linkfluence bouleverse dès lors la définition même de l'échantillon :

---

[14] Directrice d'études, réunion de lancement, juin 2013.

[15] Co-fondateur, entretien, décembre 2012. Cette perspective où l'influence est nécessairement l'apanage d'une minorité est classique dans le domaine du marketing digital et notamment viral, ainsi que l'écrit K. Mellet : *« un petit nombre d'individus, qualifiés d'influenceurs ou de leaders d'opinion, est supposé disposer d'une capacité d'influence élevée sur leur entourage »* (2012, p.159).

[16] Directeur de l'innovation, entretien, mars 2013.

[17] Idem.

en rompant avec toute caractérisation démographique, et ne sélectionnant que les sites « influents » et intéressés *a priori* par une thématique, l'entreprise construit un échantillon volontairement asymétrique où ne sont écoutés que ceux qui sont susceptibles d'avoir une opinion sur le sujet, et sont à même de la faire entendre. Peut alors débuter le travail proprement dit de mesure des opinions.

Refermer le modèle. Totalité de référence et incarnation de l'opinion

Nous nous basons ici sur une étude menée par Linkfluence sur une émission d'aménagement et de décoration d'intérieur, conçue par une enseigne de bricolage et régulièrement diffusée à la télévision[18]. Le client de Linkfluence, à savoir l'enseigne de bricolage, souhaite évaluer le succès de son émission[19] – envisagée comme une « marque » – auprès des internautes. A cette fin, le département des études a construit un échantillon de communautés jugées pertinentes au regard de la marque, incluant notamment celles liées à la décoration d'intérieur (« maison », « foyer »), mais également à la télévision – puisqu'il s'agit d'étudier un programme – et au marketing, puisque cette émission correspond à une stratégie de communication de l'entreprise, et est donc susceptible d'être commentée en tant que telle. Une fois circonscrit ce corpus de sites, un chargé d'études a codé (manuellement et automatiquement) les *verbatim*[20] produits par ces derniers selon un certain nombre de critères[21], sur une période d'un an et demi.

Ce travail de sélection et de normalisation d'un échantillon de sites web est crucial dans la quantification des opinions. Ainsi, plutôt que d'écouter l'intégralité des mentions d'une marque sur internet, Linkfluence s'accorde avec son client pour sélectionner un nombre restreint de sites en fonction de leur communauté thématique et de l'influence qu'ils y exercent. Si la plupart de ses concurrents construisent des corpus *ad hoc*, intégrant l'ensemble des pages où apparaissent les mots-clés sélectionnés (par exemple : le nom de la marque cliente), Linkfluence propose de sélectionner une quantité finie de sites de référence pour y évaluer la distribution et la qualité des opinions qui y sont émises en relation à la marque. C'est cette pratique qui lui permet de bâtir une totalité de référence, c'est-à-dire une jauge dans laquelle les opinions vont pouvoir être quantifiées. En ouverture du rapport remis au client, on trouve le constat suivant :

> « Si l'on compare le nombre de sites réactifs à la mention « [nom de l'émission] » (78 sites) au marché des conversations à partir de l'identification de communautés dédiées existantes (*a minima* 1953 sites et blogs *via* notre échantillon représentatif sur notre Live Panel), le positionnement actuel de la marque ne bénéficie pas de ce marché conversationnel important déjà existant. » (Etude « *Engage* », Linkfluence, 2011)

Nous voyons ici comment l'inscription de l'espace social en ligne dans le référentiel des « communautés » permet de ne poser le « capteur » que sur certaines zones du web, jugées pertinentes du point de vue thématique, et d'y mesurer la performance de la marque cliente. Pour cela, il n'est pas besoin de recenser toutes les mentions de la marque sur la période, loin s'en faut[22] ; au contraire, c'est l'identification des sites déjà mobilisés par les sujets intéressant la marque qui permet de constater la faiblesse de la reconnaissance dont l'émission fait l'objet. Ce qui est crucial ici, c'est l'inclusion dans le résultat chiffré de tous les sites qui ne « parlent » pas de la marque mais qui, d'après leur étiquetage communautaire, pourraient le faire, et constituent

[18] Etude « *Engage* », Linkfluence, 2011.

[19] Également appuyée par un site web du même nom.

[20] Billets de blog, *posts* de forums, articles de « sites éditoriaux » (médias le plus souvent).

[21] Entre autres : la mention (ou non) de l'émission, la mention (ou non) de l'enseigne de bricolage, la plateforme d'origine du *verbatim* (forum, blog, site éditorial), le support cité (émission TV ou site web), la tonalité vis-à-vis de la marque (positive, négative ou neutre).

[22] A titre de comparaison, nous avons lancé, sur le simple nom de l'émission et sur la même période de temps, une recherche Google, qui fournit plus de 12 500 résultats.

de ce fait l'étalonnage pertinent de la réputation en ligne de la marque. Ce verdict chiffré, 78 sites pour 1953, parcourt d'un coup toute la chaîne d'équivalence tissée par Linkfluence, depuis la qualification affinitaire du lien hypertexte jusqu'à la segmentation du web et à la construction d'échantillons thématiques.

Le rapport propose ensuite une « ventilation par communautés » des citations de la marque, moment où les opinions recueillies sont incarnées dans des collectifs stabilisés et indiscutés. Le rapport déplore en effet que la moitié seulement des citations de la marque vienne des communautés liées à la maison, cœur de cible de l'émission, et souligne que

> « les deux autres communautés portant sa visibilité sur le web social sont très loin d'appartenir au champ de l'aménagement de la maison. En effet, les sites dédiés aux programmes TV et les sites dédiés aux stratégies marketing et communication (analyse du *brand content* mis en place par [nom de l'enseigne de bricolage]) constituent l'autre principal vivier de citation de la marque ». (Etude « *Engage* », Linkfluence, 2011)

Ce diagnostic nous montre l'importance de la segmentation communautaire du web social qui, en plus d'autoriser la construction d'échantillons par affinités thématiques, est au principe d'une incarnation différenciée des opinions émises au sein même de cet échantillon, ce qui permet d'évaluer la qualité de la présence en ligne d'une marque. Dans cette perspective, l'enjeu est d'être reconnu des communautés directement liées à la thématique de l'émission, à savoir l'aménagement d'intérieur (les communautés « maison »). Symétriquement, les prises de parole issues des communautés « télévision », et plus encore, le métadiscours développé par les communautés « marketing » sur la stratégie de marque du client, sont dévalorisés : si ces mentions ne sont pas néfastes en soi, le fait qu'elles représentant la moitié du total des mentions signale en creux la mauvaise performance de la marque auprès de ce qui s'apparente à un « cœur de cible » communautaire. Linkfluence parvient à inscrire les opinions recueillies dans un référentiel qui, parce qu'il balise les frontières internes et externes de l'étude, permet de mesurer le degré de notoriété d'une marque auprès de ses différentes cibles.

De la carte au territoire. Un modèle diffusionniste de l'opinion

La double qualification du lien hypertexte, telle que supposée par le modèle communautaire, permet un mode de visualisation synoptique de l'espace social en ligne – la cartographie – qui décrit la circulation des opinions autour de « centres » influents et de « périphéries » influencées. Le référentiel communautaire et les résultats qu'il produit convoquent donc les deux valeurs investies dans le lien hypertexte : une valeur d'affinité, qui permet la clôture des échantillons, et une valeur d'autorité, qui donne à voir un modèle diffusionniste de l'opinion, où chaque site est envisagé comme prescripteur pour ses voisins hypertextes. Ce modèle s'inspire de façon manifeste du paradigme de l'influence personnelle théorisé par Lazarsfeld (2005), recyclé depuis plusieurs années dans le marketing digital (Mellet, 2012).

La dernière partie du rapport remis au client consiste en effet en une *« analyse des territoires socio-affinitaires »*, essentiellement basée sur l'outil cartographique. Dans une première carte, Linkfluence représente l'ensemble des 78 sites qui mentionnent le nom de l'émission, et les liens existant entre eux. Le diagnostic est sans appel : *« un enseignement principal se dégage de l'analyse de la structure de cette cartographie : un territoire composé essentiellement de communautés diverses et dispersées sans lien entre elles »*. Un niveau de sens nouveau s'ajoute à ce moment : on savait déjà qu'une faible proportion de sites mentionnait l'émission ; on sait désormais que ces sites ne forment pas un *« écosystème de citations uni »* (17 communautés se partagent les 78 mentions), ni cohérent, dans la mesure où les différents sites se relient très peu les uns avec les autres.

Le rapport se concentre ensuite spécifiquement sur les communautés de la maison (au nombre de sept, dont « déco *home made* », « autoconstruction », etc.) ayant mentionné l'émission, soit 39

sites qui *« constituent logiquement l'essentiel de la proximité socio-affinitaire avec la marque »*. Cependant, ces sites présentent très peu de liens entre eux (moins d'une dizaine au total), ce qui fait écrire au chargé d'études que *« cette structuration dispersée montre que les sites réactifs* [à la mention du nom de l'émission] *ne font pas partie du même réseau d'internautes : il n'y a pas de reconnaissance mutuelle qui passerait par l'échange de liens entre ces sites »*. Il précise en outre que selon la connaissance développée par Linkfluence au cours des années sur la structure communautaire du web, et notamment des communautés « maison » et « décoration », cette absence de liens est relativement inhabituelle. En fait, le seul endroit où s'esquisse un partage soutenu de liens hypertextes est le site de l'émission, qui rassemble autour de lui un *« îlot de blogs proches du positionnement de marque »*, ce qui est décrit comme un point positif mais insuffisant pour soutenir la notoriété de la marque.

Ainsi, il ne suffit pas de constater la répartition des opinions à un temps *t*. La forme cartographique, construite sur un lien hypertexte vecteur d'autorité et d'affinité, permet également de poser la question de leur diffusion sur un mode dynamique. Un écosystème de citations dense aurait été jugé favorable dans la mesure où il aurait manifesté l'existence d'un groupe de sites mobilisé autour de la marque, ce que ne montre pas la cartographie : l'absence des marqueurs d'affinité que sont les liens hypertextes montre que la marque étudiée ne « fait » pas « communauté » au sens de Linkfluence. Pour remédier à ce diagnostic jugé problématique, Linkfluence propose un certain nombre de « pistes » pour des actions de communication ultérieures, susceptibles d'améliorer quantitativement et qualitativement l'image de la marque[23].

Pour cela est construite une seconde carte, représentant les sites mentionnant l'émission, mais également ceux qui mentionnent le nom du magasin de bricolage (soit 286 sites contre 78 auparavant). L'élargissement à cette mention montre la permanence de la logique affinitaire, selon laquelle un site mentionnant le magasin peut potentiellement être intéressé par l'émission qui y est liée. Cette nouvelle cartographie fait apparaître, à côté du *cluster* déjà étudié, un nouveau *cluster* plus étendu composé de sites essentiellement issus des communautés « maison », « foyer », mais aussi « lifestyle » ou « société ». Ces nouvelles zones sont jugées particulièrement intéressantes pour deux raisons. D'une part, et contrairement à la carte précédente, elles forment des ensembles de citations compacts, densément reliés entre eux, ainsi qu'en atteste le grand nombre de liens réciproques entre sites (distingués par une couleur spéciale). Le rapport conclut de cette intense activité hypertexte que *« toute action de communication digitale auprès de cette cible peut donc s'appuyer sur une viralité potentielle notoire »*. D'autre part, le texte accompagnant la cartographie remarque que ces *« communautés féminines et influentes »* prescrivent déjà fortement le magasin client de Linkfluence, et sont donc potentiellement intéressées par l'émission.

L'épistémologie cartographique de l'opinion permet alors à Linkfluence de proposer des stratégies pour atteindre ce nouvel ensemble de blogs. Sont alors repérés des *« ambassadrices »*, c'est-à-dire des blogs mentionnant déjà l'émission, situés à l'interface entre les deux *clusters* du graphe. En s'adressant à ces auteurs de manière appropriée[24], le client pourrait étendre la notoriété de son émission à des ensembles communautaires plus larges. Filant la métaphore martiale, le rapport décrit un véritable *« territoire »* à *« préempter »*, en occupant dans un premier temps les points de passage hypertexte entre le *« territoire déjà investi »* et les *« zones d'opportunités »* voisines. Ce sont donc bien les signifiés proprement graphiques qui servent ici l'analyse, grâce aux propriétés encapsulées à l'étape précédente (Latour, 1993), qui ne sont plus discutées à ce stade. C'est parce que le lien hypertexte implique la délégation d'une autorité qu'il

[23] Actions sur la nature desquelles le rapport reste délibérément vague, Linkfluence n'étant pas une agence de communication digitale ; ce refus d'intervenir dans ce domaine marque la spécialisation de cette entreprise sur les questions d'opinion et de réputation.

[24] Pour cela, une analyse qualitative de leurs pratiques éditoriales est proposée.

est possible de cartographier les points stratégiques d'un territoire social, les prescripteurs à enrôler pour diffuser largement un message. Le référentiel communautaire de Linkfluence, outre l'imputation des opinions, donne également à voir leur circulation suivant un principe hiérarchique.

Conclusion

L'étude du modèle communautaire permet en définitive de répondre à la question posée au début de cet article : qui écoute-t-on sur le web, c'est-à-dire aussi et surtout, à qui donne-t-on la parole ? Nous avons vu comment, par l'introduction de nouvelles techniques d'échantillonnage, Linkfluence construit une opinion doublement *autorisée*, en donnant la parole à des groupes mobilisés *a priori* par un enjeu, et dotés d'une compétence rare à faire entendre et partager leur avis. Dans les échantillons de Linkfluence, tous ne peuvent participer, au contraire ; l'opinion qui s'y exprime est celle d'individus dont on a préalablement certifié la compétence. La figure de l'internaute de base, du consommateur moyen, ou du citoyen lambda, y est absente. Les critiques formulées par Bourdieu puis Champagne à propos des sondages semblent dès lors dépassées, ou intégrées : la méthodologie de Linkfluence implique en effet que toutes les opinions (en ligne) ne se valent pas. L'asymétrie délibérée des échantillons de Linkfluence constitue une innovation majeure en matière de mesure de l'opinion, et peut être reliée à deux types de causes. La première est commune à tous les prestataires sur le marché du *social media analysis*, et tient à la spontanéité du matériau en ligne lui-même, qui empêche de fait de collecter artificiellement les réponses des « sans-opinion » ; quelle que soit la méthodologie déployée, il n'est possible d' « interroger » sur le web que ceux qui veulent bien s'y exprimer, ce qui constitue un argument de poids pour ces méthodologies émergentes.

Cependant, la pratique de l'échantillonnage communautaire relève également d'un parti pris, revendiqué par Linkfluence et d'autres acteurs sur le marché, qui ne constitue finalement que l'une des manières d'appréhender des traces numériques extrêmement abondantes et variées. Si ces échantillons présentent le mérite d'imputer des tendances d'opinion à des groupes stabilisés, le paradigme qui les sous-tend ne peut prétendre à l'heure actuelle au statut de convention dans ce domaine. Dans le contexte propre à une expertise émergente, d'autres normes, d'autres conventions sont avancées, notamment par des concurrents proposant au contraire une veille voulue exhaustive et largement soutenue par les outils logiciels. Réaliser une anthropologie compréhensive des techniques de quantification de l'opinion sur le web suppose donc de pouvoir mettre en présence ces projets alternatifs, leurs justifications et leurs méthodologies. Le parti pris opposé à Linkfluence, celui du traitement extensif et automatisé des données conversationnelles, a d'ores et déjà été exploré par D. Boullier et A. Lohard (2012). D'autres contributions seront nécessaires pour obtenir – là aussi – un point de vue synoptique sur ces nouvelles épistémologies de l'opinion.

RÉFÉRENCES

Adamic, Lada A., and Natalie Glance. 2005. "The Political Blogosphere and the 2004 US Election: Divided They Blog." In *Proceedings of the 3rd International Workshop on Link Discovery*, 36–43. New-York, USA.

Akrich, Madeleine, Michel Callon, and Bruno Latour. 2006. *Sociologie de la Traduction : Textes Fondateurs*. Presses de l'Ecole des Mines.

Badouard, Romain. 2013. “Les mobilisations de clavier: Le lien hypertexte comme ressource des actions collectives en ligne.” *Réseaux* 181 (5): 87–117.

Barabasi, Albert-Lazlo, and R. Albert. 2002. “Statistical Mechanisms of Complex Networks.” *Review of Modern Physics* 74 (145).

Blondiaux, Loïc. 1990. “Paul F. Lazarsfeld (1901-1976) et Jean Stoetzel (1910-1987) et Les Sondages D’opinion : Genèse D’un Discours Scientifique.” *Mots*, no. 23 (Juin): 5–23.

Blondiaux, Loïc. 1998. *La Fabrique de L’opinion. Une Histoire Sociale Des Sondages*. Paris: Seuil.

Boullier, Dominique, and Audrey Lohard. 2012. *Opinion mining et Sentiment analysis*. Open Edition. Collection Sciences-Po | Médialab 1. http://press.openedition.org/198.

Bourdieu. 1973. “L’opinion Publique N’existe Pas.” *Les Temps Modernes*, no. 318 (January): 1292–1309.

Cardon, Dominique. 2013. “Dans L’esprit Du PageRank. Une Enquête Sur L’algorithme de Google.” *Réseaux* 1 (177): 63–95.

Chakrabarti, Soumen, B. Dom, and D. van den Berg. 1999. “Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery.” In *Proceedings of WWW 1999*. Toronto.

Daston, Lorraine, and Peter Galison. 1992. “The Image of Objectivity.” *Representations*, no. 40: 81–128.

De Maeyer, J. 2013. “Towards a Hyperlinked Society: A Critical Review of Link Studies.” *New Media & Society* 15 (5): 737–51.

Desrosières, Alain. 2008. *L’argument Statistique (I) - Pour Une Sociologie Historique de La Quantification*. Paris: Presses de l’Ecole des Mines.

Fouetillou, Guilhem. 2008. “Le Web et Le Traité Constitutionnel Européen. Ecologie D’une Localité Thématique Compétitive.” *Réseaux*, no. 147: 229–57.

Ghitalla, France, Camille Maussang, Fabien Pfaender, and Eustache Diemert. 2004. “TARENTe: An Experimental Tool for Extracting and Exploring Web Aggregates.” In *Proceedings of ICCTA ’04*.

Herbst, Susan. 1995. *Numbered Voices. How Opinion Polling Has Shaped American Politics*. Chicago: The University of Chicago Press.

Jacomy, Mathieu, Guilhem Fouetillou, and Fabien Pfaender. 2006. “Two Visions of the Web : From Globality to Localities.” In *Proceedings of ICCTA ’06*. Damas.

Kleinberg, Jon. 1999. “Authoritative Sources in a Hyperlinked Environnment.” *Journal of the ACM* 46 (5): 604–32.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Latour, Bruno. 1993. “Le Topofil de Boa Vista Ou La Référence Scientifique. Montage Photo-Philosophique.” *Raison Pratique*, no. 4: 187–216.

Méadel, Cécile. 2004. “L’audimat Ou La Conquête Du Monopole.” *Le Temps Des Médias* 3 (2): 151–59.

Mellet, Kevin. 2012. “Contagion, Influence, Communauté. Petite Socio-Économie Des Agences de Social Media Marketing.” In *Du Lien Marchand : Comment Le Marché Fait Société. Essai(s) de Sociologie Économique Relationniste*, 402. Toulouse: Presses Universitaires du Mirail.

Pollak, Michael. 1979. “Paul F. Lazarsfeld, Fondateur D’une Multinationale Scientifique.” *Actes de La Recherche En Sciences Sociales* 25 (1): 45–59.

Porter, Theodore. 1995. *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton University Press. Princeton.

Rogers, Richard, and Nortje Marres. 2000. “Landscaping Climate Change: A Mapping Technique for Understanding Science and Technology Debates on the World Wide Web.” *Public Understanding of Science* 9 (2): 141–63.

Thévenot, Laurent, and Alain Desrosières. 2002. *Les Catégories Socioprofessionnelles*. La Découverte. Paris.